Agence de Modernisation des Universités et Etablissements

Le projet « Entrepôt de Données »

Jean-François Desnos

Résumé : Le projet a commencé en septembre 1999, dans le cadre du groupe de travail « Pilotage de Etablissements ». Il a débuté par la réalisation d'une maquette[1] et un rapport de pré-étude[2]. En janvier 2000, le programme proposé a été retenu par l'Agence, dont le groupe de travail « Pilotage de Etablissements » a établi le cahier des charges. En novembre 2000, une première version a été livrée à cinq universités pilotes : Amiens, Paris VI, Rennes I, Strasbourg I et Versailles, rejointes en mars 2001 par Besançon, Lille I et Lille II. Une version stabilisée est prévue pour novembre 2001. L'objectif est de constituer une base de départ, chaque université enrichissant ensuite son entrepôt de données selon sa politique et ses besoins propres.

1. Introduction.

A partir des applications actuelles de l'Agence de Modernisation des Universités et Etablissements, il est déjà courant d'exploiter un certain nombre d'indicateurs. Le module « Pilotage » fourni par l'Agence pour Apogée permet par exemple, à l'aide d'un outil d'infocentre[3], de présenter des tableaux personnalisés, réalisés à partir d'extractions de la base de données Apogée. Des développements similaires ont été réalisés par les universités pour Harpège, Nabuco et la Paye, que ce soient des « univers B.O. » comme pour Apogée, des présentations de tableaux sur l'intranet, ou des sorties sous forme de fichiers à exploiter de manière autonome (avec Oracle, Excel, Word,...).

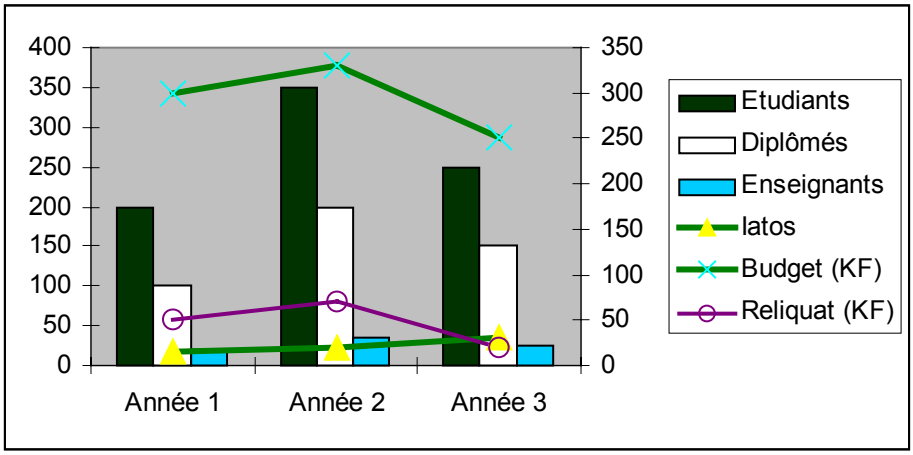
Lorsqu'on veut aller plus loin, croiser, pour une discipline donnée, le nombre de chercheurs, d'enseignants, d'étudiants, de diplômés et les moyens financiers, pour donner un exemple général, les choses deviennent vite plus compliquées.

---

[1] La maquette informatique réalisée en 1999 montre des extractions faites avec l'outil Business Objects sur l'ensemble "Apogée, Nabuco, Harpège". Elle met en évidence la faisabilité et les limites d'une telle démarche.

[2] Rapport à l'AMUE : "Pré-étude Entrepôt de Données", J.-F. Desnos, 14/01/2000.

[3] Le produit Business Objects (B.O.) a été retenu par l'Agence de Modernisation parmi les logiciels du commerce. Certains développements locaux en SQL ont aussi été réalisés.



Exemple de tableau de bord : quelques indicateurs pour une composante donnée. L'année est une dimension, la composante pourrait être une seconde dimension, l'entité géographique (campus) une troisième.

Tout d'abord, les données des bases d'exploitation évoluent à chaque instant. Pour établir des situations arrêtées, pouvoir les reprendre, et établir des séries chronologiques, il faudrait procéder à partir de copies archivées. Ensuite, on constate rapidement que les données sont hétérogènes et de qualité variable. Elles sont hétérogènes parce que les applications ont été conçues et réalisées indépendamment. Elles sont de qualité variable, car souvent incomplètes (reprises d'historiques partielles lors de la mise en service, champs non renseignés car considérés comme peu utiles, erreurs,...), et quelquefois de sens imprécis (définitions non fournies). Enfin, il existe une hétérogénéité structurelle liée aux différentes représentations de l'université : selon que l'on traite de finances (Nabuco) ou d'étudiants (Apogée), on considère des années différentes (année civile ou année universitaire), et des ensembles de composantes ou d'UFR non identiques. Ces divergences s'accentuent si l'on examine les structures plus en détail.

Les bases de données d'exploitation, supportant Apogée, Nabuco, Harpège et Paye, sont donc organisées pour une utilisation transactionnelle - inscrire un étudiant, enregistrer une facture, suivre la carrière de personnels par exemple - et non principalement pour en extraire des tableaux de bord.

La réalisation d'un "Entrepôt de Données" (en abrégé ED) répond à ce dernier besoin. On construit un ED par extraction d'informations sélectionnées dans des bases de données "sources", informations qui sont vérifiées et le cas échéant transformées avant d'être injectées dans une base de données "cible" ou ED. Ce processus de chargement de la base cible, effectué périodiquement, fournit des "couches historiques", qui sont des photographies de la base d'information de l'établissement. Contrairement aux bases sources, l'organisation de la base cible est orientée vers la production de rapports, indicateurs et tableaux de bord.

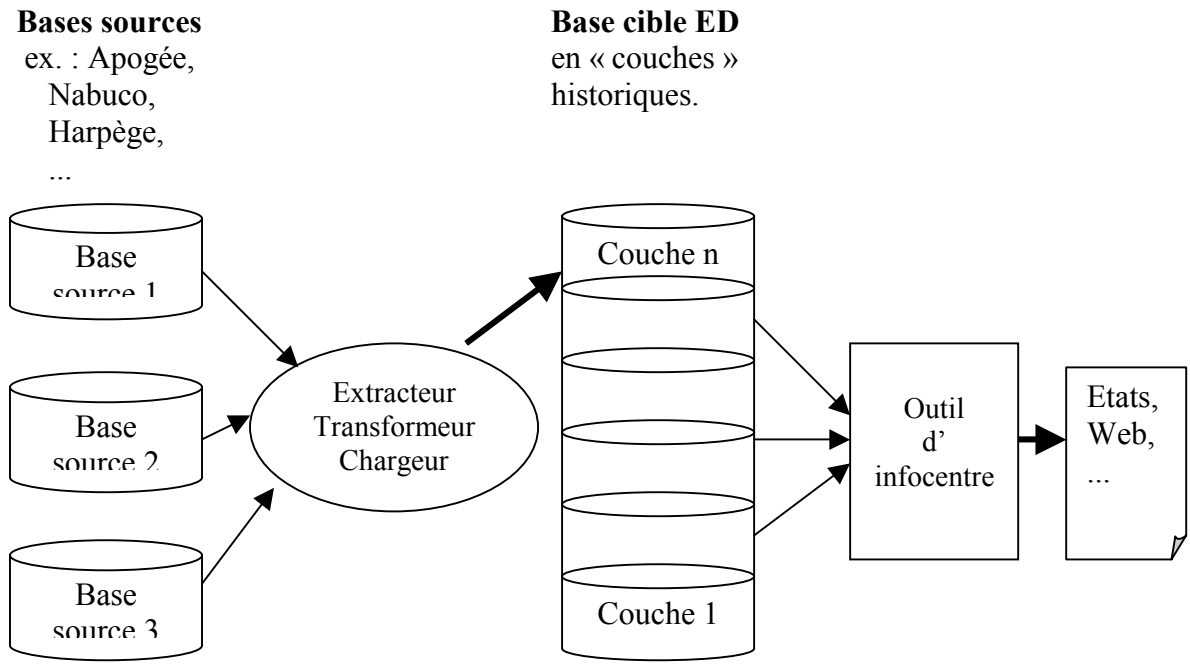
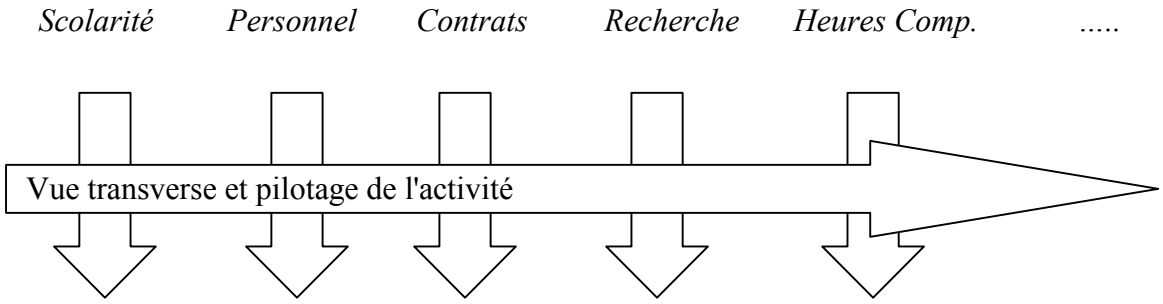


Schéma de principe de chargement d'un ED.

Dans le cas d'une université, les bases sources peuvent être celles des applications de l'Agence déjà citées : Apogée, Nabuco, Harpège et Paye, mais aussi des bases spécifiques à l'établissement, ou même extérieures à l'université et à son environnement.



Un ED présente une vue transverse de l'institution

Un projet d'élaboration d'ED doit prendre en compte l'objectif de production de tableaux de bord à partir du système d'information global de l'établissement, et doit aussi tenir compte des contraintes : hétérogénéité des bases sources, chargements historiques réguliers, base cible destinée à l'extraction de rapports. **C'est d'abord un projet de traitement d'information, donc informatique**, qui doit tenir compte de la complexité, de la variété, et des évolutions des bases sources, de la qualité variable (données manquantes notamment) de ces bases sources, et de la grande diversité des indicateurs à obtenir. Elaborer un entrepôt de données, c'est **éditer** à destination de la direction et des personnels de l'université concernés, un potentiel de tableaux de bord vérifiés et validés.

2. Les étapes du projet.

2.1 Le programme de l'année 2000.

Le programme de travail retenu pour l'année 2000 comportait les étapes suivantes :

a) **La définition du cahier des charges du projet.** Trois tableaux de bord ont été établis par le groupe de travail "Pilotage des Etablissements" de l'Agence (comportant des anciens présidents d'université, des secrétaires généraux, chargés de mission "statistiques" ou "pilotage", des informaticiens) et validés par la CPU.

b) **La constitution d'un méta-dictionnaire.** Le méta-dictionnaire d'un ED décrit les données à extraire des bases sources, leur nature, leur mode de transformation, leur mode de transfert vers la base cible. Il s'agit d'une étape importante, nécessitant des connaissances techniques approfondies de l'ensemble des applications sources. Ce travail a été validé par l'Agence à Montpellier et par le groupe de travail "pilotage" de l'Agence.

c) **La conception de la base cible**[4], et la programmation de deux types de procédures :

c1 les procédures logicielles d'extraction, transformation et chargement de données (ETL) des bases sources vers la base cible,

c2 les procédures logicielles d'extraction de données de la base cible et leur présentation.

Compte tenu des évolutions potentielles des bases sources, des chargements répétés de la base cible, et de l'aspect très souple que doit présenter le processus d'interrogation de la base cible, le choix a été fait de retenir des outils du commerce pour réaliser ces procédures logicielles :

- Data Stage d'Informix[5] pour les opérations **c1**,
- Business Objects (version 5) pour les opérations **c2.**

d) La livraison d'une première version V1 en novembre 2000.

Conformément au programme élaboré en janvier 2000, il a été adopté une démarche évolutive d'élaboration de l'ED. La version V1 livrée fin 2000 était la première version.

[4] Oracle 8i a été retenu pour la base cible.
[5] « Préconisation d'un outil ETL », J.-F. Desnos, 07/09/2000.

Les universités pilotes.

Après appel de candidatures, l'AMUE a pris en juin 2000 le parti de retenir 2 tranches d'universités pilotes :

- Tranche 1 (démarrage en juillet 2000) : Amiens, Paris VI, Rennes I, Strasbourg I, et Versailles.
- Tranche 2 (démarrage en janvier 2001) : Besançon, Lille I et Lille II.

Concernant la tranche 1, Amiens, Rennes I, Strasbourg I, et Versailles exploitent les quatre applications de l'Agence : Apogée, Nabuco, Harpège et Paye.

Paris VI n'exploite pas Apogée, et avait acquis un outil d'ETL préalablement au démarrage du projet. Paris VI n'utilise donc pas Data Stage, et doit donc adapter les procédures d'extraction à son application de gestion d'étudiants, Scolar, et à son outil d'ETL, Genio.

Par ailleurs, chacune de ces cinq universités possède son application spécifique de gestion des heures complémentaires, utilisée comme une des sources par le projet, et peut exploiter d'autres applications locales. Les procédures de chargement de l'ED seront donc adaptées par chacune des universités pilotes, mais la structure de la base cible ED sera semblable dans chacun des établissements, au moins pendant la phase pilote du projet.

L'installation du projet dans les universités pilotes.

L'ED a été installé par l'équipe de Grenoble de l'AMUE dans les universités d'Amiens, Rennes I, Strasbourg I, et Versailles entre le 20 novembre et le 4 décembre 2000, en fonction du calendrier d'installation des outils logiciels (Oracle 8i, B.O. v5, Data Stage) et matériels (serveur, postes NT) de chacun des sites.
Lors, ou à l'issue, de chacune de ces visites, une base ED prototype a été chargée, et des tableaux de bord édités.
Concernant Paris VI, documents et logiciels ont été livrés par internet.

Les documents techniques d'accompagnement, recensés dans le rapport de septembre 2000[6], ont été livrés à l'ensemble des sites. Ces documents sont les suivants :

- Estimation des moyens minimaux nécessaires aux sites pilotes,
- Recommandations aux sites pilotes,
- Méta-dictionnaire,
- Préconisation d'un outil logiciel ETL,
- Cahier des charges d'implantation,
- Scripts de chargement de la base cible ENTREPOT,
- Modèle conceptuel de données d'ENTREPOT,
- Univers et requêtes B.O.,
- Méthodologie et mise en oeuvre de l'ED.

[6] « Etat d'avancement du projet au 30/09/2000 » (JFD).

2.2 Le programme 2001.

La version V1 livrée fin 2000 est destinée à évoluer. En effet, comme il a été proposé[2] , nous avons retenu un développement en « spirale » comportant schématiquement les étapes suivantes :

a - un périmètre fonctionnel volontairement limité à un ensemble de besoins,
b - la réalisation d'une version d'ED correspondante,
c - des évolutions demandées par les utilisateurs : extension du périmètre fonctionnel,
d – retour en b, pour créer une version enrichie ou,
e – arrêt des évolutions.

Ce processus est particulièrement adapté à la réalisation d'un ED, en raison de l'évolutivité des besoins, et de la souplesse d'enrichissement de ce type de base de données.

Dans la phase pilote, l'ensemble des sites doit garder la même structure d'ED, quitte à ce que chacun développe des interrogations (avec Business Objects) spécifiques.

Des évolutions communes du méta-dictionnaire, des procédures de chargement, et de la structure de la base cible sont donc prévues. Lors de la phase pilote, ces initiatives sont coordonnées, et les sites doivent se tenir aux évolutions choisies.

L'objectif de la phase pilote est de disposer en novembre 2001 d'un ED livrable aux sites de généralisation.
Cet ensemble livrable a pour objet de rendre les sites autonomes.

Etapes du projet en 2001 :

- livraison aux sites pilotes tranche 1 d'une V2 corrective et avec évolutions mineures en janvier 2001,

- livraison aux sites pilotes d'une nouvelle version V3 en mai 2001, comportant, en réponse à leurs demandes, des modifications de la base cible et de l'univers BO.

- tenue d'un séminaire, bilan de l'expérimentation des sites pilotes en octobre 2001

- stabilisation la V3, et préparation d'un « livrable » sur CD-ROM comportant l'ensemble des logiciels et documents du projet à destination des sites de généralisation en novembre 2001.

3. Le cahier des charges.

Le groupe de travail « Pilotage de Etablissements » déjà cité a proposé de retenir dans l'ED trois tableaux de bord. Il a donné du tableau de bord la définition suivante :

« Le **tableau de bord** doit donner une vision cohérente et représentative de la politique de l'établissement. Il doit permettre d'anticiper et de prévoir, il doit signaler les dysfonctionnements et mesurer les performances. Ce doit être une référence commune pour l'équipe de direction et les conseils : en ce sens c'est aussi un acte de communication. C'est un instrument d'action, synthétique, privilégiant les représentations graphiques plutôt que des tableaux de nombreux chiffres. C'est aussi un ensemble de clignotants qui doivent permettre de susciter une réaction immédiate
La construction d'un tableau de bord impose une réflexion approfondie sur les **indicateurs** nécessaires. C'est à chaque établissement de trouver et de combiner ses propres indicateurs, ce qui requiert une analyse très fine du fonctionnement de l'université, des objectifs prévus dans son projet d'établissement, des données dont l'établissement dispose dans son système d'informations. »

Le groupe de travail a retenu trois tableaux de bord :

Tableau 1 : évolution des emplois d'enseignants-chercheurs par rapport aux besoins de formation et de recherche .

Le but de ce tableau de bord est d'aider l'université à prendre des décisions sur les emplois.
Les indicateurs permettront de satisfaire les demandes suivantes :

- Concernant la formation : rapprocher, par section CNU, les charges de formation, les effectifs étudiants, le nombre et le type de formations (par cycle) du potentiel disponible (nombre d'enseignants par corps) ; figurer l'évolution sur les dernières années ; établir un histogramme des âges des enseignants-chercheurs.
- Concernant la recherche : observer l'évolution des laboratoires : les laboratoires « « labellisés », le nombre d'habilités, le taux d'encadrement des doctorants, les primes d'encadrement.

Tableau 2 : présentation synthétique des UFR

L'objectif est de fournir un certain nombre de ratios aux composantes.
Indicateurs retenus :

- Budget : D.G.F., report des crédits budgétaires non consommés, réserves et évolution
- nombre d'étudiants
- Rapport enseignants-chercheurs / Iatos, avec leurs attributs : grades, âges, catégories (A, B, C) pour les IATOS, nombre de contractuels (quel type de financement),
- nombre d'intervenants extérieurs, heures complémentaires payées,
- Offre de formation : nombre de diplômes, nombre de diplômés par filière par rapport au nombre d'inscrits, activité en formation continue.

Evolution de ces indicateurs sur plusieurs années.

Tableau 3 : l'attractivité internationale de l'établissement en matière d'offre de formation

L'objectif est d'aider l'établissement à maîtriser sa politique internationale.
Indicateurs retenus :

- Nombre d'étudiants inscrits à l'université, avec répartition suivant les secteurs géographiques d'origine (dont Union européenne), et par type de formation suivie,
- Nombre d'étudiants étrangers accueillis dans le cadre de programme d'échange,
- Nombre de français partant à l'étranger dans le cadre d'un programme,
- Nombre d'étudiants inscrits dans des formations intégrées conduisant à un double diplôme,
- Nombre de thèses en co-tutelles et évolution,
- Nombre d'enseignants étrangers accueillis,
- Nombre d'enseignants partant à l'étranger,
- Montants du budget international de l'établissement par type de financement (dont part et sources de financement provenant d'autres sources que le MENRT), et du budget des services « internationaux ».

4. La base cible.

4.1 La modélisation dimensionnelle.

Avant de fournir la structure de la base cible, nous allons présenter quelques notions et définitions concernant la modélisation d'un entrepôt de données et notamment :

- la structure dimensionnelle en tables de faits et tables de dimensions,
- et, au §7, la visualisation des données (analyse multi-dimensionnelle).

Ces notions ne seront qu'abordées dans ce document. Pour une information plus complète, le lecteur peut se référer à la bibliographie, en particulier aux ouvrages [K1], [K2] et [BO], à la présentation de l'INSA de Lyon à EUNIS 2001 [FL], et à une note de J.-B. Nataf sur la structure de l'entrepôt de données pour le pilotage des universités[7].

La structure dimensionnelle.

La structure des données d'une application transactionnelle comme Apogée est établie sur un modèle dit relationnel. Cette modélisation est adaptée aux transactions, beaucoup moins aux interrogations. Pour élaborer un entrepôt de données, on utilise plutôt un modèle dimensionnel. L'unité d'un modèle dimensionnel, c'est l'ensemble d'une table de faits et de tables de dimensions, dont chacune est reliée à la table de faits par une jointure. On obtient ainsi un schéma en étoile. Un entrepôt de données, c'est schématiquement un ensemble de tables de faits, donc un ensemble d'étoiles.

[7] Structure de l'entrepôt de données de pilotage : les objectifs de l'entrepôt de données de pilotage des universités et leurs conséquences sur sa structure et sa construction, note à l'Agence, Jean-Baptiste Nataf , avril 2001

Table de faits.

Une table de faits comporte des dimensions (reliées aux tables de dimensions) et des indicateurs. Exemple de table de faits :

dimensions : - composante
- année universitaire
- nationalité
- étudiant

indicateur : - inscrit (0 ou 1, Oui / Non)

Tables de dimension.

A la variable composante, on associe une table de dimension, où figureront par exemple l'adresse de la composante, le nom de son directeur et autres attributs de type "texte".

De même, année universitaire, nationalité et étudiant peuvent se voir associer chacun une table de dimension.

Base cible ED.

Pour répondre aux interrogations du cahier des charges, la base cible est structurée de la manière suivante :

- Cinq tables de faits :
  - Inscription_etape,
  - Inscription_CNU_elp,
  - Poste_affecte,
  - HeureComp,
  - Situation_financiere.
  -
- Tables de dimension.
  La plupart des tables de dimension associées sont limitées au libellé (mais ce n'est pas le cas de Identite_etudiant ou Elp, par exemple).

  Certaines d'entre elles sont associées à plusieurs (voir à toutes les) tables de fait :
  - Correspondance_ufr, ou liste de « composantes ED » auxquelles on associe les composantes correspondantes dans chacune des applications sources.
  - Annee_universitaire (indique le début d'année universitaire),
  - Lib_CNU (libellé des sections CNU),
  - Dates_observations. L'ED est structuré en « couches de données » dont chacune représente une photographie de l'université à une date donnée. Le rythme de ces photographies (chargements de l'ED) est choisi par l'université. Celle-ci peut décider d'éditer des données recueillies à des dates différentes pour les étudiants, les enseignants, le budget...ces dates différentes sont consignées dans la table des dates d'observation.

Nous présentons en annexe le modèle de données de la version V3 de l'ED, suscité par le groupe de travail technique de l'Agence.

5. Le méta-dictionnaire.

Le méta-dictionnaire est constitué des méta-données du projet. Il comprend les définitions générales et détaillées sur les notions utilisées (agrégats, tables de faits) dans la base cible de l'entrepôt de données.
Le méta-dictionnaire précise la définition des données, la façon dont elles sont calculées (règles de gestion), ainsi que leur méthode de calcul à partir des bases de données de l'AMUE.
Le méta-dictionnaire est un document technique de 55 pages. On ne donnera ici que les notions générales et un exemple de quelques données chargées à partir d'Harpège.

Notions générales.

- **Année universitaire** : c'est une notion clé dans l'entrepôt de données, car dans la majorité des cas les calculs sont faits par année universitaire. En ce qui concerne le tableau 1 et toutes les données qui lui sont rattachées, l'année est l'année universitaire, considérée au 2 septembre (pour le personnel). Dans le tableau 2 et le tableau 3, l'année pour les données budgétaires est l'année civile, pour les autres données c'est l'année universitaire considérée au 2 septembre de l'année civile précédente.
  Par ailleurs, pour les données extraites de la base HARPEGE, une liste des années de « validité » des données de l'entrepôt sera à paramétrer (une liste d'années sera à saisir, dans une table de la base cible à la mise en service du prototype sur chaque site pilote) et seules ces années permettront de faire des calculs, sur le personnel. En effet le personnel pouvant être en poste depuis de nombreuses années, il est nécessaire de sélectionner les années sur lesquelles on souhaite faire les calculs.

- **Elément pédagogique** : les calculs utilisant la notion d'élément pédagogique sont basés sur les éléments pédagogiques « feuilles », et uniquement sur ceux-là. Ce sont les éléments pédagogiques terminaux dans la structure des enseignements.

- **UFR, Composante ou Structure** : les données APOGEE sont calculées par rapport aux « composantes » existantes dans la base, les données issues de NABUCO le sont par rapport à un niveau 1 passé en paramètre au chargement et aux niveaux 2 (actuellement le niveau 3 est conservé dans l'entrepôt et sert pour le calcul du budget de la « cellule internationale »). Les données de la base HARPEGE sont extraites en tenant compte de la structure d'affectation et si celle-ci n'est pas de type composante ou établissement on prend en compte la structure de type composante ou l'établissement dont dépend la structure d'affectation.

  Pour effectuer des rapprochements entre les données des différentes bases sources, une liste de correspondance entre les codes des composantes APOGEE, structures de type composante d'HARPEGE, niveaux 2 de NABUCO et composantes des heures complémentaires sera à définir, sous la forme d'un fichier texte au moment de l'installation du prototype, par les sites pilotes. Cette liste est chargée dans une table de la base cible qui sert au moment de l'édition des tableaux avec BO.

- **CNU** : le code de la section CNU est dans les bases sources codé sur 2, 4 ou 5 caractères. Dans l'entrepôt de données, il est codé sur 2 caractères, ce qui correspond à la section sans la sous-section. La présentation des résultats se fera en considérant ce code CNU sur 2 caractères par agrégation des sous-sections correspondantes.

Le libellé de la section CNU sera celui de la base HARPEGE, table CNU.
CNU et APOGEE : le code CNU des éléments pédagogiques « feuilles » devra être saisi pour assurer la validité des calculs.
CNU et HARPEGE : le code CNU utilisé est celui associé au poste et non celui associé à la carrière de l'agent (en fait ce sont les postes occupés qui sont classés par CNU dans l'entrepôt et non le personnel). Le code CNU associé à la carrière du personnel est conservé dans l'entrepôt mais non utilisé.

- **Nombre d'heures équivalent TD** : le nombre d'heures est calculé en utilisant les facteurs de correspondance habituels suivants : 1 pour les heures de TD, 1,5 pour les heures de CM et 0,667 pour les heures de TP. L'existence de groupes de TP TD ou CM est intégrée dans le calcul des heures équivalent TD par l'entrepôt de données (s'il y a 2 groupes de TP, les heures sont multipliées par 2).
  Le nombre d'heures correspond aux heures réellement effectuées. Le calcul par diplôme fournira donc le nombre d'heures effectuées par les enseignants et non le nombre d'heures suivi par un étudiant inscrit à ce diplôme.
  APOGEE : les volumes horaires des éléments pédagogiques « feuilles » devront être renseignés pour assurer la cohérence des calculs dans l'entrepôt.

- **Corps** : les enseignants statutaires et les heures complémentaires sont calculés par corps. Pour la base HARPEGE on prendra le corps de l'enseignant. Pour la base HELICO (base de gestion des heures complémentaires utilisée à Grenoble), les heures complémentaires sont calculées en utilisant les codes grade-profession, CNU et composante de la base HELICO.

- **Nombre d'étudiants** : l'entrepôt de données extrait les étudiants ayant une inscription administrative à une étape. Un étudiant inscrit à plusieurs étapes est extrait autant de fois que d'inscription. Le témoin étape première est chargé dans l'entrepôt pour pouvoir faire des agrégats uniquement pour les inscriptions principales.
  Le nombre d'étudiants « physiques » est calculé uniquement dans le tableau 3 au niveau de l'université. En effet un étudiant peut s'inscrire à plusieurs formations correspondant à différents diplômes, composantes ou cycles.
  Tous les calculs par section CNU et par éléments pédagogiques, et les calculs d'heures de cours, ne concernent que les étudiants ayant en plus de l'inscription administrative à une étape une inscription pédagogique à cette étape, cette dernière devant obligatoirement être décomposée en éléments pédagogiques.
  Les étudiants « doctorants » sont les étudiants inscrits aux diplômes suivants : Thèse, Diplôme de recherche technologique et Habilitation à diriger des recherches. Pour ceux-ci des informations complémentaires sont extraites de la table APOGEE THESE_HDR_SOUT. Pour les calculs d'agrégats, seuls ceux inscrits en Thèse sont pris en compte.

- **Nombre d'enseignants** :
  Par section CNU : le calcul se fait en considérant la section CNU du poste, seul les enseignants (contractuels ou statutaires) occupant un poste sont comptés.
  Les enseignants du second degré sont calculés en fonction de leur spécialité et non de la section CNU puisqu'ils n'en possèdent pas.
  Par corps : seuls les enseignants statutaires sont comptés (les contractuels n'ayant pas de corps).

- **Tranche d'âges** : les tranches d'âges sont dans l'entrepôt de données d'une durée d'un an, le calcul par tranche plus longue se fera au moment de l'interrogation.
- **Nombre d'IATOS** : les IATOS correspondent aux personnes présentes dans la base HARPEGE et n'étant pas enseignants.
- **Diplômés** : les diplômés sont les étudiants et les doctorants ayant un résultat positif à un diplôme ainsi que l'autorisation de délivrance de diplôme à *Vrai*, quelle que soit la session, seul le résultat d'admission étant pris en compte. Le ratio diplômés/inscrits correspond au rapport des étudiants (y compris les doctorants) qui ont réussi le diplôme sur ceux qui sont inscrits à l'étape diplômante.
- **Budget** : montant total du budget voté avec les Décisions Budgétaires Modificatives.
- **Dépenses** : montant total des mandats moins le montant total des ordres de reversements.
- **Recettes** : montant total des titres validés moins montant total des réductions.
- **Disponible** : montant du reste à dépenser (budget moins dépenses).

Exemple de données chargées (élément du méta-dictionnaire).

Le tableau ci-dessous figure quatre données de la base cible ED ayant trait au personnel (colonne de gauche). Ces données sont extraites d'Harpège. La colonne centrale indique la table d'Harpège concernée s'il y a lieu (on constate que les contractuels n'ont ni code_corps, ni carrière). La colonne de droite précise le champ source extrait et les conditions d'extraction (règles de gestion). On ne retient que les codes structure de type « établissement » ou « composante ».

Colonne cible	Table source	Colonne source
NOM	INDIVIDU	NOM_PATRONYMIQUE
CODE_CORPS	ELEMENT_CARRIERE. Non renseigné pour les contractuels	CODE_CORPS
CODE_STRUCTURE_AFFECTATION	STRUCTURE	C_STRUCTURE avec C_TYPE_STRUCTURE=C ou E
DATE_DEBUT_CARRIERE	CARRIERE ou ELEMENT_CARRIERE ou AFFECTATION. Non renseigné pour les contractuels	D_DEB_CARRIERE ou, si elle est non renseignée, le minimum de D_DEB_AFFECTATION et D_EFFET_ELEMENT

Quelques exemples de données chargées dans la base cible.

6. Exemple de procédure Data Stage : le chargement du budget.

Paramètres de la procédure : niveau 1 (suppression de composantes), intervalle d'années.

- Extraction des éléments de dépense de la table E_DEP_GES de Nabuco et création du fichier temporaire *Hash1*.
  - Clé : Exercice, Code N1, Code N2, Code N3, Code destination, Code rubrique,
  - Montant budget F, Cumul mandats F, Cumul reversements F
  - Calculs en Euro.

- Extraction des éléments de recette de la table E_REC_GES de Nabuco et création du fichier temporaire *Hash2*.
  - Clé : Exercice, Code N1, Code N2, Code N3, Code destination, Code rubrique
  - Cumul titres validés F, Cumul réductions validées F
  - Calculs en Euro.

- Chargement de la table Budget à partir de *Hash1* et *Hash2* en deux étapes :
  - Création des lignes de dépense, avec recette correspondante s'il y a lieu (sinon, titres et réductions de titres sont mises à zéro), date de chargement, composante ED.
  - Création des lignes de recette sans correspondance de dépense, date de chargement, composante ED.

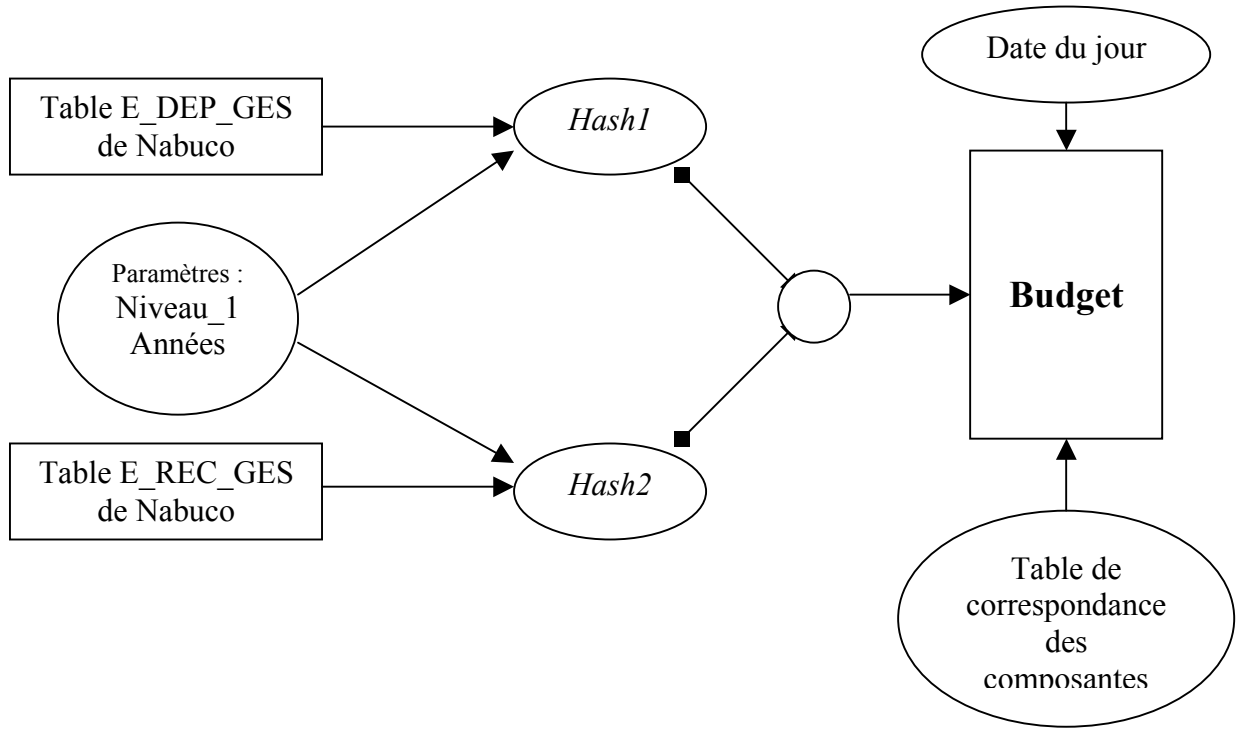


Schéma de principe de la création de la table BUDGET de l'ED

Pourquoi utiliser un outil d'ETL ?

Les procédure d'extraction / transformation / chargement telles que l'exemple décrit ci-dessus pourraient être programmées sans outil spécifique, à l'aide des langages SQL et C par exemple.
Qu'apporte un outil d'ETL du commerce, comme Data Stage, qui a été retenu ?

Concernant le développement et la maintenance de l'ED :

- une plus grande rapidité de développement,
- une maintenance adaptative et corrective plus sûre,
- une évolutivité beaucoup plus aisée (bases cibles et bases sources),
- une adaptabilité à des systèmes d'information qui évoluent.

Et en matière d'exploitation :

- des chargements programmés (couches), par un outil de planification intégré.

7. Les interrogations de la base cible.

En version V3 du projet, les interrogations sont proposées à partir de deux univers Business Objects (BO). Un certain nombre de requêtes pré-établies y sont associées.

Le progiciel Business Objects est utilisé pour extraire, mettre en forme et analyser des données provenant d'une base de données.

Une requête BO extrait les données, puis les présente dans des tableaux ou des documents dynamiques auxquels il est possible d'associer des graphiques, et dont il est possible de détailler les niveaux. Ces documents peuvent être publiés sur l'intranet de l'établissement.

Schématiquement, BO est manipulé par deux catégories de personnes :

- Les spécialistes de la base de donnée concernée, qui définissent les univers, c'est à dire l'architecture des informations extraites de la base de données. Pour le projet ED, deux univers seront livrés par l'Agence. Cela n'est pas limitatif. La cellule « Pilotage » de l'université aura notamment pour rôle de créer d'autres univers, spécifiques à l'établissement, et les requêtes associées.
- Les utilisateurs manipulent un ou des univers :
  - de manière avancée en créant de nouvelles requêtes,
  - de manière simple en paramétrant et en exécutant des requêtes existantes, par un simple « clic » sur une icône.

Rotation des dimensions et tranchage.

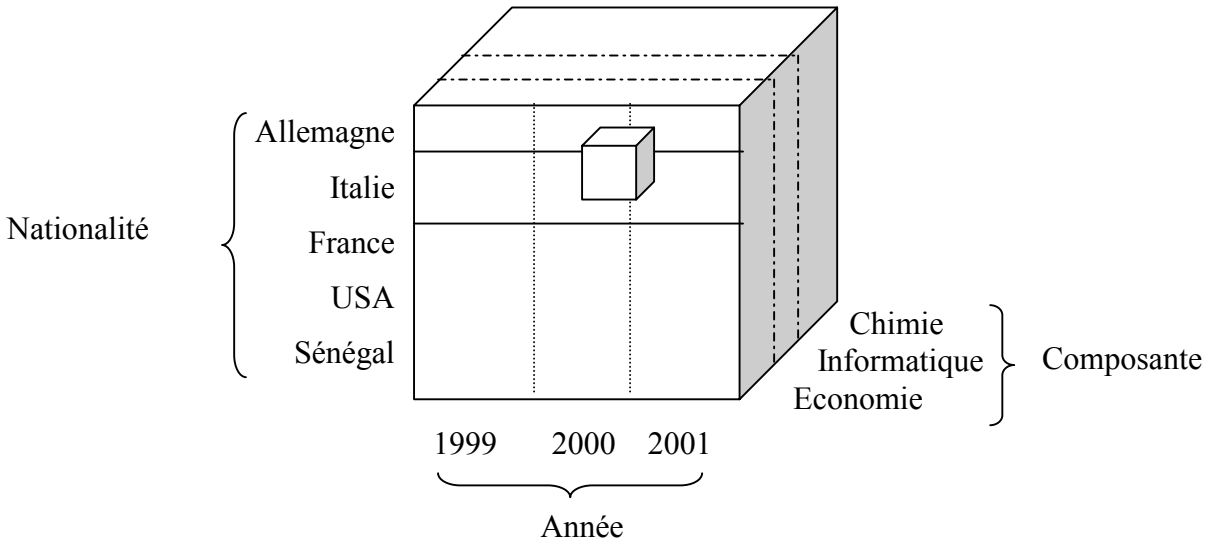
Les variables d'un univers BO, appelées objets, sont essentiellement de deux types :

- ♦ dimension : une dimension prend une liste de valeurs,
- o indicateur : un nombre.

Reprenons notre exemple du §4 de schéma en étoile très simple.
Un table de faits avec :

- ♦ 4 dimensions { - composante, - année universitaire, - nationalité, - étudiant
- o 1 indicateur : - inscriptions

On peut représenter le nombre d'inscriptions sur un tableau, ici un cube puisqu'il y a 3 dimensions :



En rendant constante une des dimensions, par exemple en fixant la discipline à la valeur "informatique", on obtient un tableau à 2 dimensions (de type tableur), et dans chaque élément du tableau le nombre d'inscriptions par année et nationalité dans la composante "informatique".
De la même manière, on peut fixer l'année, ou la nationalité. On dit par commodité que l'on effectue une rotation des dimensions et des "tranches" de cube. Il y a 3 possibilités de tranchage parallèlement aux côtés du cube.

Dans le cas général, un indicateur est associé à n dimensions (n > 3) et le cube est alors appelé hypercube. Si l'on veut en avoir une représentation lisible, on doit fixer n-2 dimensions.

A ces n dimensions, sont en général associés plusieurs indicateurs.
Ici, un deuxième indicateur pourrait être le nombre de thèses.

Hiérarchie de dimensions.

Considérons notre dimension "composante" (au sens composante de l'entrepôt). Si la composante peut être décomposée en cycles et étapes[8], à la place de la dimension composante, on aurait trois dimensions hiérarchisées (relation d'inclusion) :

- composante
  - cycle
    - étape

Dans ces conditions, nous souhaitons obtenir des détails sur les indicateurs :

- nombre d'incriptions
- nombre de thèses
-

c'est à dire que les éléments du tableau soient, à la demande, détaillés par étape, ou, au contraire, sommés par cycle ou pour l'ensemble des cycles d'une composante.
On dit alors que l'on fait de l'analyse dimensionnelle, descendante (vers les détails) ou montante (vers les regroupements).

De la même manière, les années pourraient être composées de périodes de temps plus courtes (semestre par exemple), et les nationalités pourraient être groupées par continent ou autres sur-ensembles (union européenne par exemple). On manipulerait alors 3 ensembles de dimensions hiérarchiques, avec de nombreuses possibilités d'observation, en faisant de l'analyse multi-dimensionnelle.

En résumé, lorsqu'on a organisé les données en dimensions hiérarchiques et indicateurs, il y a deux types d'analyse classiques, notamment proposés par B.O.[9] :

- La rotation des dimensions et tranchage,
- L'analyse descendante ou ascendante.

Ces possibilités sont exploitées dans l'entrepôt de données.

[8] la hiérarchie d'Apogée pourrait être : cycle, diplôme, version de diplôme, étape, version d'étape.
[9] Avec l'option "explorer" pour l'analyse ascendante ou descendante.

8. Conclusion.

Dans le cadre du projet global "Pilotage des Etablissements" de l'Agence, le projet Entrepôt de Données a pour objet de fournir aux universités une base de départ pour un système informatisé de pilotage qui soit à la fois indépendant des applications de production, globalement cohérent, et fortement évolutif. Cette nouvelle application pourra alors être enrichie au sein de chaque établissement pour progressivement constituer le système d'information de l'université. Le projet Entrepôt de Données est une entreprise de proposition aux universités d'un embryon, répondant à un cahier des charges proposé par un groupe de travail de l'Agence. Son élaboration est rendue possible en raison de l'exploitation depuis 1995 par les universités d'un noyau commun d'applications nationales de gestion.
Les universités pilotes vont tester et faire évoluer en commun le prototype jusqu'en septembre 2001. A partir de novembre 2001, une version stabilisée de l'entrepôt de données sera enregistrée à l'Agence de Modernisation, sous forme de document électronique contenant les logiciels et documents nécessaires son implantation.

9. Remerciements.

Ce projet est un travail d'équipe, effectué en collaboration et avec le soutien des personnalités et groupes de travail suivants :

- Josette Soulas, initiatrice du projet, et Suzanne Maury-Silland, Directrices Générales successives de l'Agence de Modernisation des Universités et Etablissements,
- Christian Charrel, Directeur de Département, et Sibylle Rochas, Coordonnateur du pôle domaines, responsables du projet Pilotage à l'Agence,
- Hélène Brochet-Toutiri, Régine Waltener (Agence), Yves Chaimbault (Lille I), Bernard Fradin (Lyon II), Michel Issindou (Université Grenoble II), Guy Bailleul (Lille II), Frédérique Cazejoux (Versailles), Yves Chaimbault (Lille I), Michel Daumin (Amiens), Annie Julien, Elisabeth Lagente, Solange Nédelec (Université Rennes I), Martine Stern (Versailles), Emmanuelle Salzard (Nancy II) membres du groupe de travail "Pilotage" et responsables des sites pilotes ,
- Roselyne Casteloot (Paris-VI), Jean-Pierre Finance, Paul Personne (Conférence des Présidents d'Université), Jacques Francois (Lyon-I), Sylvie Robert (Grenoble-II), Michel Roignot (Franche-Comté), Marylène Oberlé (Strasbourg-I), (Amiens), Jean-Emmanuel Rudio (Strasbourg-II), Danièle Savage (Lille-III), membres du groupe de travail "Pilotage"et concepteurs du cahier des charges,
- Bernard Barbez (Lille I), Emmanuelle Cravoisier (Amiens), Mostafa Al Haddad, Claude Derieppe (Lille II), Brigitte Perrigault, Anne Routeau (Rennes-I), François Cadé (SIIG Strasbourg), Elisabeth Flenet (Besançon), Olivier Raunet, Valéry Vaillant (Strasbourg-I), Dominique Fiquet, Nicolas Courtay, Robert Rivoire (Versailles), informaticiens des sites pilotes, et du groupe technique de projet qui travaillent à l'évolution du projet et à son implémentation,
- Olivier Morelle et Jean-Baptiste Nataf (Paris-VI), qui ont de plus activement contribué à la modélisation de la base cible et la conception d'univers B.O.,
- Marie-Hélène Glénat, Virginie Garnier et Nicolas Maume (Agence), développeurs du projet à Grenoble.

10. Bibliographie.

Livres :
[BO] Manuel du designer V5, Business Objects, 1999.
[F] Piloter l'entreprise grâce au data warehouse, J.-M. Franco et al., Eyrolles 2001.
[G] La construction du datawarehouse, J.-F. Goglin, Hermès 1998.
[I] Building the Data Warehouse, W. H. Inmon, Wiley 1996.
[K1] Entrepôts de données, guide pratique du concepteur, R. Kimball, Wiley 1997.
[K2] Concevoir et déployer un data warehouse, R. Kimball et al., Eyrolles 2000.

Article :
[FL] "A Design and Implementation of a Data Warehouse for Research Administration in Universities", André Flory, Pierre Soupirot, Anne Tchounikine, INSA de Lyon.
Actes d'EUNIS 2001, The 7th International Conference of European University Information Systems, Berlin, Humboldt-University, March 28-30, 2001

Site Web:
http://www.dw-institute.com/ Cours, conférences, séminaires, ouvrages sur le data warehouse.

Annexe : les tables de la base cible

Tables de fait concernant les étudiants et tables associées

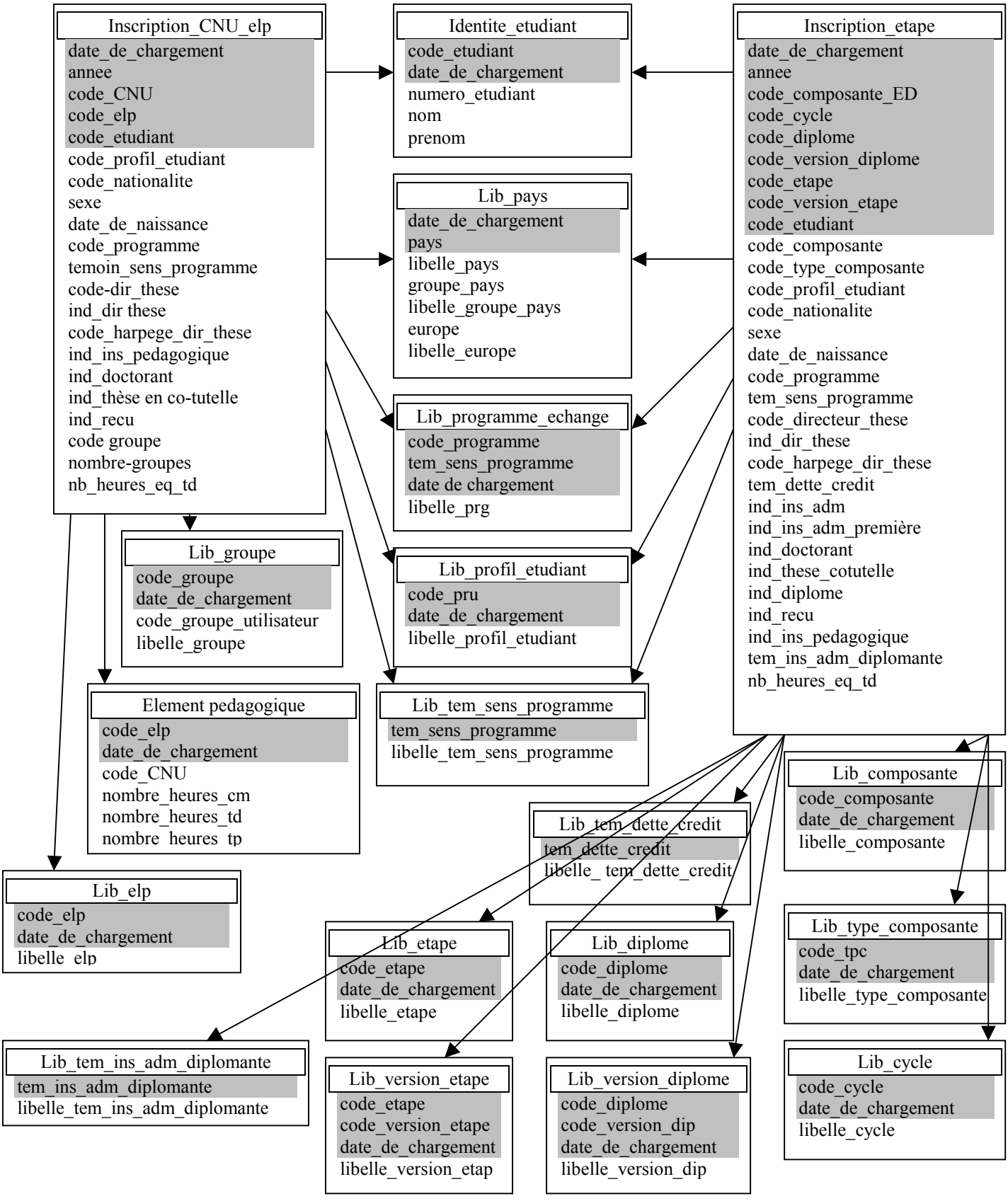


Table de fait concernant le personnel et tables associées

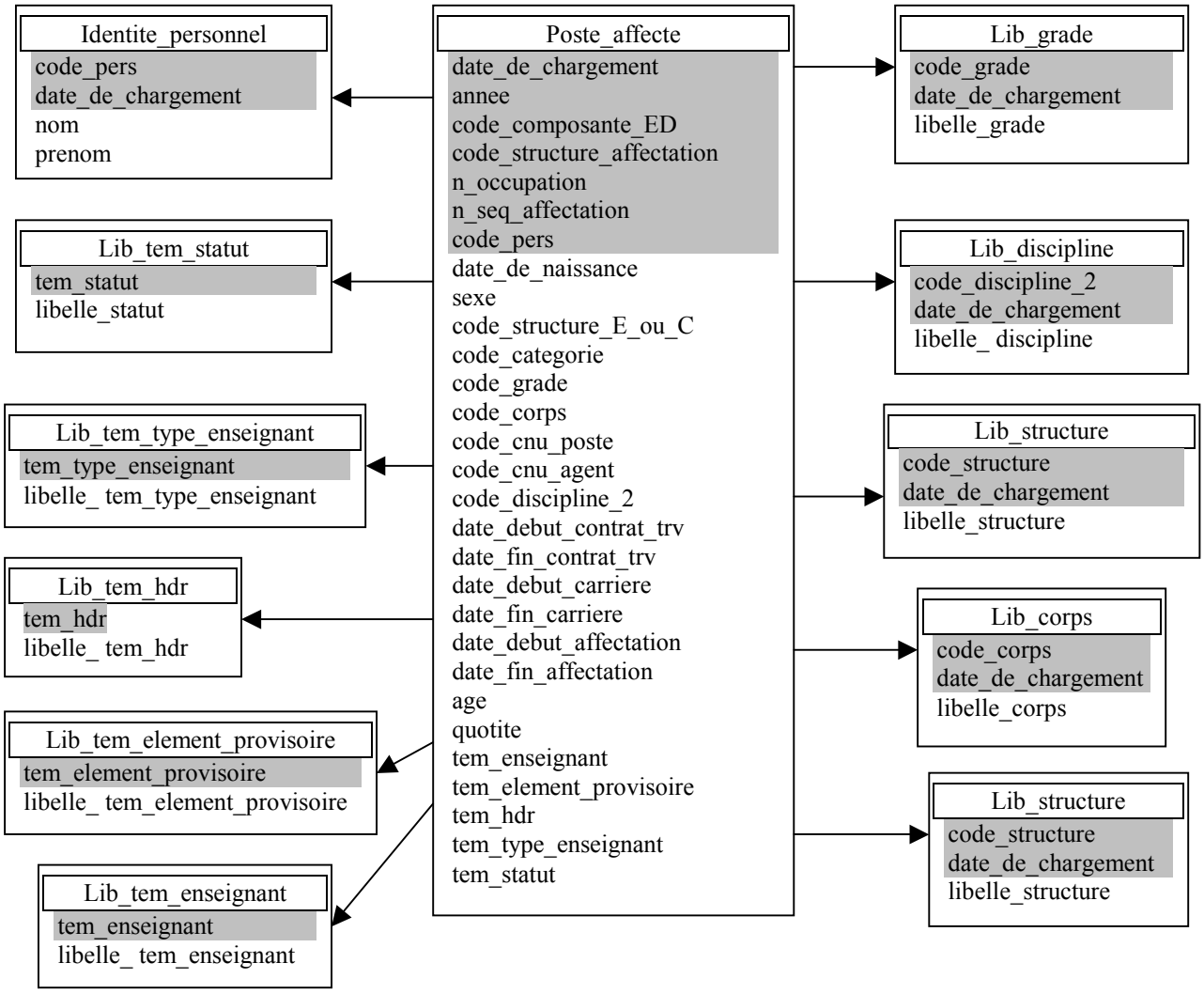


Table de fait concernant les heures complémentaires et table associée

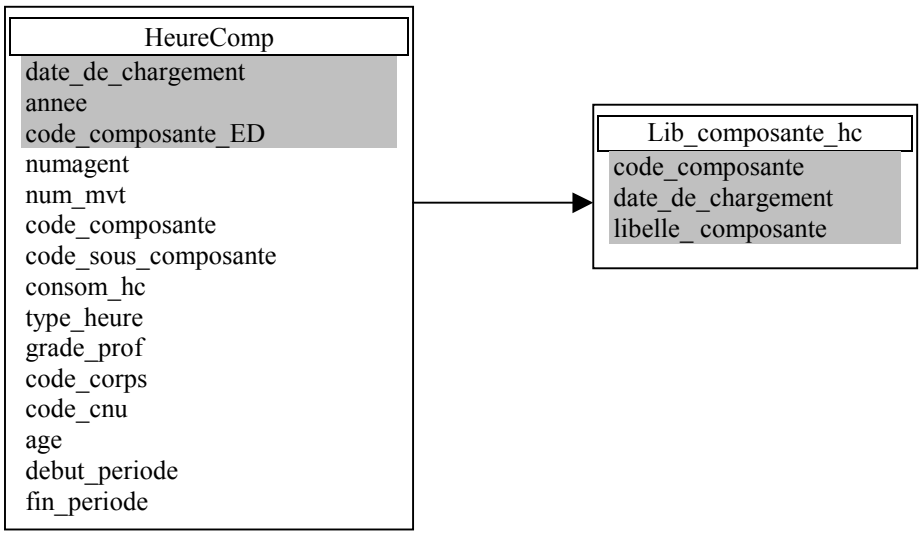
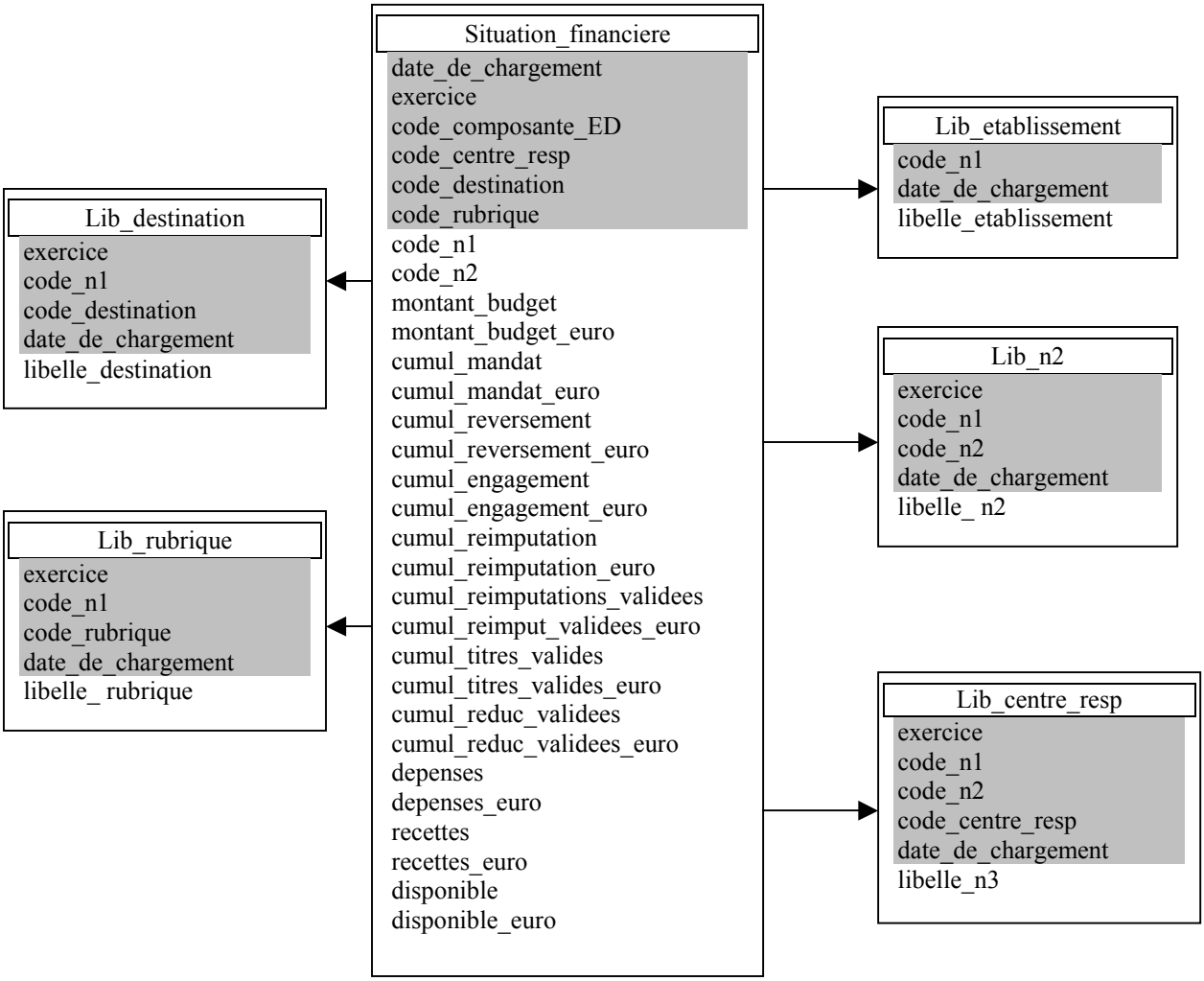


Table de fait concernant le budget et tables associées



Tables associées à toutes les tables de fait.