

ARCHIMED - Une base de données historicisée au quotidien

In : « Mémoire électronique, Archivage et travail des historiens du futur », Bulletin de l'association Histoire et Informatique (vol. 13/14, 2002/2003, pp. 89-102)
[<http://www.ahc-ch.ch/index.php?id=40&L=1>]

Jean-Daniel Zeller
Archiviste principal
Hôpitaux universitaires de Genève

Résumé

Les Hôpitaux universitaires de Genève, forment depuis 1995 un réseau d'établissements hospitaliers, comptant près de 7'900 collaborateurs travaillant sur 5 sites géographiques, comptant 2187 lits, et offrant annuellement 45'000 hospitalisations et 730'000 consultations ambulatoires (chiffres 2001). L'hôpital cantonal de Genève a été un des pionnier de l'informatique médicale en Suisse, voire en Europe. Dès 1976, le centre d'informatique hospitalière concevait le système DIOGENE, qui allait connaître de nombreux développements. Actuellement, la base de données patients rassemble les données d'environ un million de patients, avec un accroissement annuel de l'ordre de 40'000 patients.

Dès le départ, le système DIOGENE a été conçu comme un système orienté patient et non pas comme un système de gestion administratif. Cependant, une telle base de données ne pouvait pas ne pas être exploitée du point de vue statistiques, tant pour des interrogations de type médical qu'administratif. Durant les premières années d'utilisation, ce type d'interrogation s'effectuait en mode batch, durant les heures creuses de l'exploitation. Cependant, ce type d'exploitation s'est progressivement heurté à plusieurs obstacles, et l'on envisagea la possibilité de créer une base « d'archivage » qui permettrait le lancement de requêtes statistiques sophistiquées, paramétrables, et sur une base anonyme sans mobiliser les ressources de la base de données opérationnelle.

Ce projet vit le jour en 1993 par la réalisation de la base de données intégrée ARCHIMED,. Le concept clé de cette application réside dans la transcription des données liées aux patients en une série de « faits élémentaires », qui peuvent être manipulés ultérieurement de manière extrêmement performante. Pratiquement, les données préalablement sélectionnées des différents systèmes informatiques opérationnels, sont déversées quotidiennement dans une application qui les transforme en faits élémentaires puis les déversent dans la base de données ARCHIMED sur une base annualisée.

Après dix ans d'activité (1993-2003) les chiffres sont les suivants :
70 millions d'enregistrements (faits élémentaires)
Accroissement d'environ 1,5 Gigabytes par an (5MB * 365 j. = 1,8 GB)
(soit 200 tables annuelles)

Le concept qui a présidé à la naissance d'ARCHIMED a été d'ordre médical et opérationnel. Cependant, la nécessité d'harmoniser les données provenant de plusieurs systèmes initiaux, a forcé ses concepteurs à une réflexion ontologique qui a amené à une structuration qui rencontre les préoccupations d'une conservation des données à long terme. ARCHIMED représente donc un prototype de ce que pourrait être une base de donnée « historique », conçue dès le départ pour maîtriser un grand nombre de données à long terme.

Summary

The university Hospitals of Geneva, formed since 1995 with a network of hospitable establishments, with 7'900 collaborators working on 5 geographic sites, with 2185 beds, and offering annually 175'000 hospitalizations (HC: 52'000) and 620'000 ambulatory consultations (2001).

The cantonal hospital of Geneva was one of the pioneers of medical data processing in Switzerland, even in Europe. From 1976, a team of IT specialists-doctors, designed the DIOGENE system, which went through numerous developments.

At present, the data base of patients DIOGENE gathers data for a little more than a million of patients, with an annual increase of the order of 45'000 new patients. From the beginning, the DIOGENE system was designed as a patient oriented system and not as an administrative system of management.

Such a data base could be exploited from the statistical point of view, for example for queries of a medical type or administrative data. During the first years of use, this type of query was made in batch mode, during the 'slack' periods.

However, this type of exploitation gradually met with several obstacles and we envisaged the possibility of creating an "archival" data base with the possibility of directly interrogating an anonymised data base, by using browsers, without using the operational data bases resources. This project lives by the implementation of the data base integrated ARCHIMED system in 1993. The key concept of this application lies in the transcription of the data linked to the patients in a series of "elementary facts", which can later be manipulated in an extremely powerful way.

Practically, the pre-selected data from the various operational computer systems, are filtered daily by a series of interfaces, which transforms them into elementary facts then places them in the ARCHIMED data base on an annualised base.

After ten years of activity (1993-2003) the annual evolutions are the following :
Increase in the neighborhood of 1,5 Gigabytes a year ($5\text{MB} / \text{j} * 365 \text{j} = 1,8 \text{GB}$)
70 million entries (elementary facts)
(i.e. the addition of 200 tables annually)

The concept which presided over the birth of ARCHIMED was of a medical and operational order. However, the necessity of harmonizing the data resulting from several different operational systems, forced its designers to an ontological reflection leading to a structuralization meeting the concerns of a conservation of the long-term data.

ARCHIMED represent therefore a prototype "historical" data base, designed at the beginning for the management of large amounts of data for long term preservation.

1. Contexte institutionnel

Les Hôpitaux universitaires de Genève, forment depuis 1995 un réseau d'établissements hospitaliers, comptant près de 7'900 collaborateurs travaillant sur 5 sites géographiques, comptant 2185 lits, et offrant annuellement environs 175'000 hospitalisations (HC : 52'000) et 620'000 consultations ambulatoires (chiffres 2001). Il est issu du regroupement des établissements suivants :

Etablissement	Créé en*	Type de soins	Lits
Hôpital cantonal (HC)	1856	Soins aigus (i.c. pédiatrie et maternité)	1176
Hôpital psychiatrique	1900	Psychiatrie	343
Hôpital de Loëx	1900	Soins de longue durée	268
Hôpital gériatrique	1970	Gériatrie	293
Centre de soins continus	1980	Soins palliatifs	105
			2185

* sur leur site actuel

Cette répartition, issue de l'héritage historique, tend à être remplacée par une organisation plus répartie, en fonction des spécialités médicales, regroupées actuellement en 11 départements médicaux, correspondants également aux activités académiques que se doit d'assumer un hôpital universitaire.

2. Contexte informatique

L'hôpital cantonal de Genève a été un des pionnier de l'informatique médicale en Suisse, voire en Europe. Dès 1976, l'équipe d'informaticiens-médecins emmenée par le professeur Jean-Raoul Scherrer, concevait le système DIOGENE, qui allait connaître de nombreux développements, dont les principales étapes (applications patients uniquement) sont résumées ci-dessous :

Année	Application	Remarques
1978	Démarrage de l'application hospitalière DIOGENE.	Au départ, uniquement pour les patients hospitaliers
1982	Intégration des policliniques dans le système DIOGENE.	Bases de données en parallèle. L'intégration sera achevée seulement en 1993.
1988-1998	Déploiement progressif des applications UNI-LAB.	Gestion des examens de laboratoire
1994	Mise en application d'UNI-IMAGE.	Gestion des examens d'imagerie médicale
1993-1996	Mise en application d'UNI-DOC.	Production de documents à partir des données DIOGENE
1995	Migration de DIOGENE sur des machines et applications standards. Les stations de travail sont remplacé par des PC.	Restructuration des bases de données. Changement d'OS et de DBMS.
1998	Création du concept DOMED (médical), DOSSI (soins infirmier) puis DPI (Dossier du Patient Intégré).	Navigateur médical permettant l'accès aux données/documents des patients issus de diverses applications
2003	Début de l'analyse de la fusion des applications de gestion administratives des patients (DIOGENE+PHILOS)	Intégration des systèmes de l'hôpital cantonal avec ceux des autres sites hospitaliers
2005	Fusion des bases de données IMPACT (site HC) et PHILOS (autres sites)	Dans une nouvelle application nommée DPA (Dossier Patient Administratif)

Actuellement, la base de données patients de DIOGENE rassemble des données pour un peu plus d'un million de patients, avec un accroissement annuel de l'ordre de 45'000 nouveaux patients.

3 Problématique ...

Dès le départ, le système DIOGENE a été conçu comme un système orienté patient et non pas comme un système de gestion administratif. Cela est particulièrement bien illustré par la représentation constamment affirmée de la « roue » DIOGENE (voir figure 1). Cela apparaît également dans la structure fonctionnelle des application dont le point d'entrée est également le patient (voir schéma des tables de données de DIOGENE, figure 2).

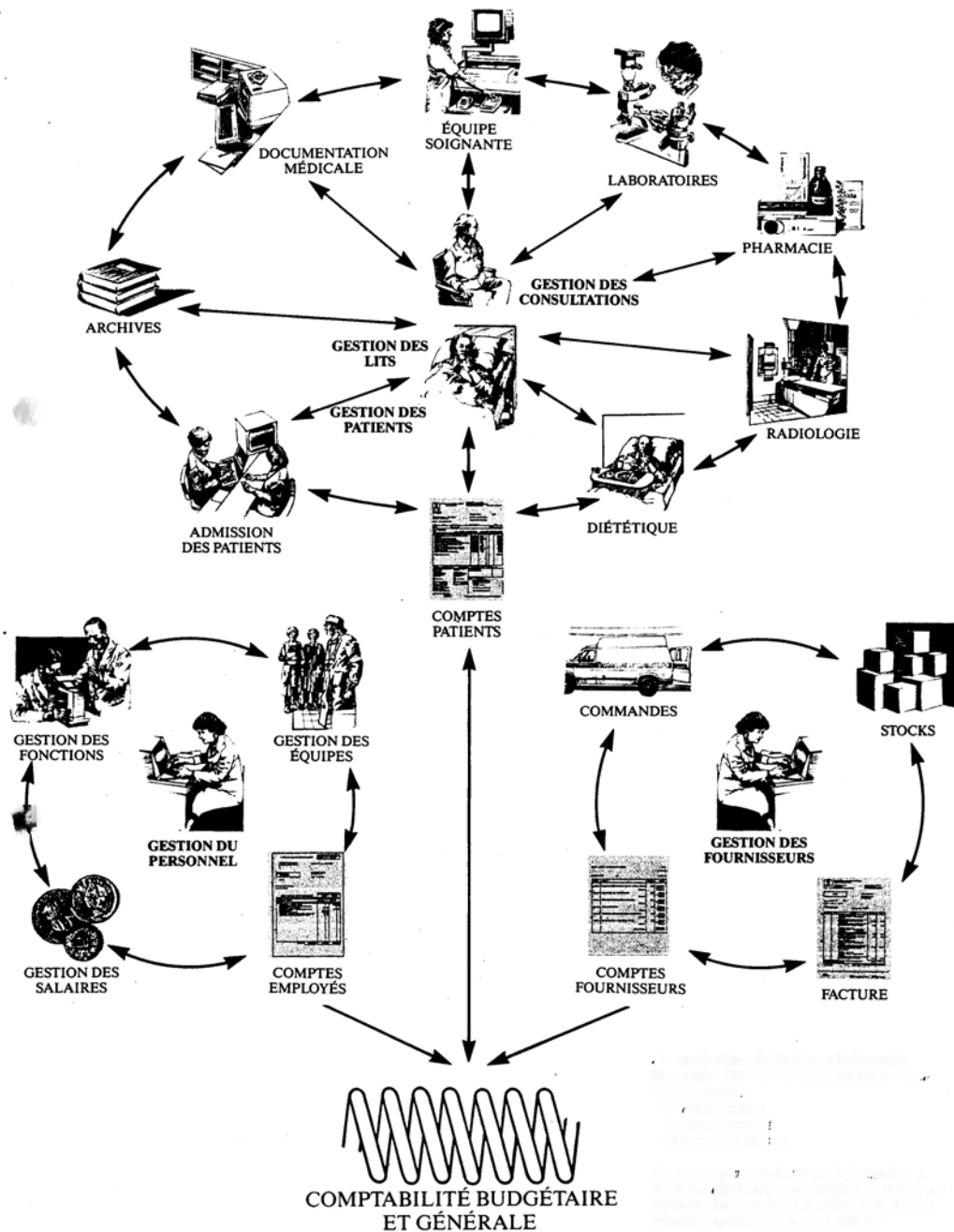


Figure 1 : Structure globale des applications informatiques des HUG

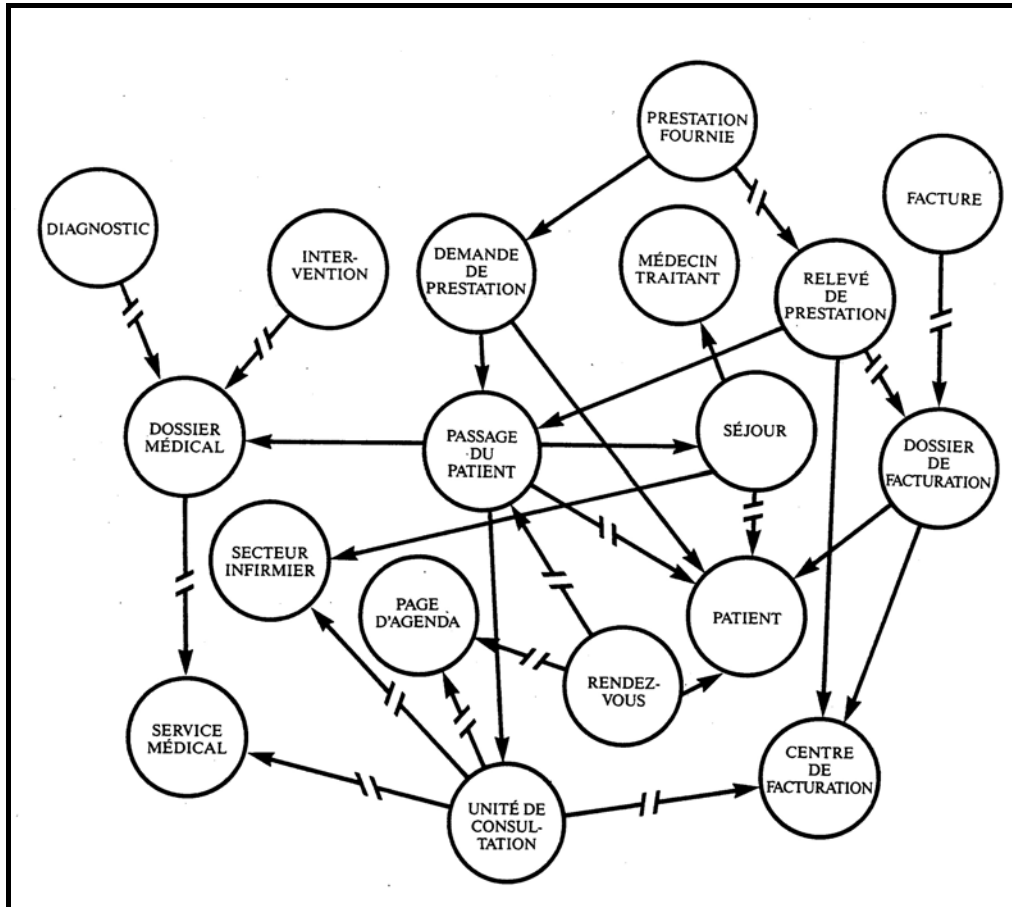


Figure 2 : Schéma des relations des données dans le système DIOGENE

Les flèches simples indiquent une relation fonctionnelle, les flèche barrées indiquent une application.

Cependant, une telle base de données ne pouvait pas ne pas être exploitée du point de vue statistique, tant pour des interrogations de type médical (charge des services, épidémiologie) qu'administratif (statistiques annuelles, coût par patient, etc.).

Durant les premières années d'utilisation de DIOGENE, ce type d'interrogation s'effectuait en mode batch, durant les heures creuses de l'exploitation, entre 22 heures et 4 heures du matin. Cependant, ce type d'exploitation s'est progressivement heurté à plusieurs obstacles :

- L'accroissement continu de la base et les besoins de sécurité élevés nécessitaient de plus en plus d'opérations de maintenance devant s'effectuer pendant ces mêmes heures creuses.
- L'accroissement de la base rendait également les analyses statistiques de plus en plus intéressantes pour les gestionnaires, qui prirent l'habitude d'en solliciter de plus en plus.
- L'organisation hiérarchique de la base ne facilitait pas le lancement de requêtes statistiques, dont les critères n'étaient pas directement liés au patients mais plutôt à des groupes d'événements, qui pouvaient se situer à des niveaux hiérarchiques très divers dans la base, et nécessiter la constitution de fichiers temporaires très volumineux, perturbant d'autant le fonctionnement opérationnel courant de la base de données.

Pour toutes ces raisons, le responsable de ces éditions statistiques, envisagea dès les années 90, la possibilité de créer une base normalisée unique, qui permettrait :

- d) Le lancement de requêtes statistiques sophistiquées et paramétrables sans mobiliser les ressources de la base de données opérationnelle.
- e) La possibilité d'interroger une base de données anonymisée, directement par les utilisateurs finaux au moyen de navigateurs (en limitant la réalisation à la demande de requêtes SQL par les informaticiens).
- f) La possibilité d'intégrer des données relatives aux patients issues de systèmes d'information divers. Cette option allait s'avérer cruciale quelques années plus tard lorsqu'il s'agira de préparer la fusion des HUG, dont les établissements possédaient des systèmes informatiques divers et non intégrés.

Ce projet vit le jour par la réalisation de la base de données intégrée ARCHIMED, en 1993. Le concept clé de cette application réside dans la transcription des données liées aux patients en une série de « faits élémentaires », qui peuvent être manipulés ultérieurement de manière extrêmement performante. Avec l'ouverture du service intranet à l'Hôpital cantonal, cette version « dynamique » des atlas statistiques remplacera définitivement en 1995 la version papier produite entre 1983 et 1993.

4 Le concept d'ARCHIMED¹

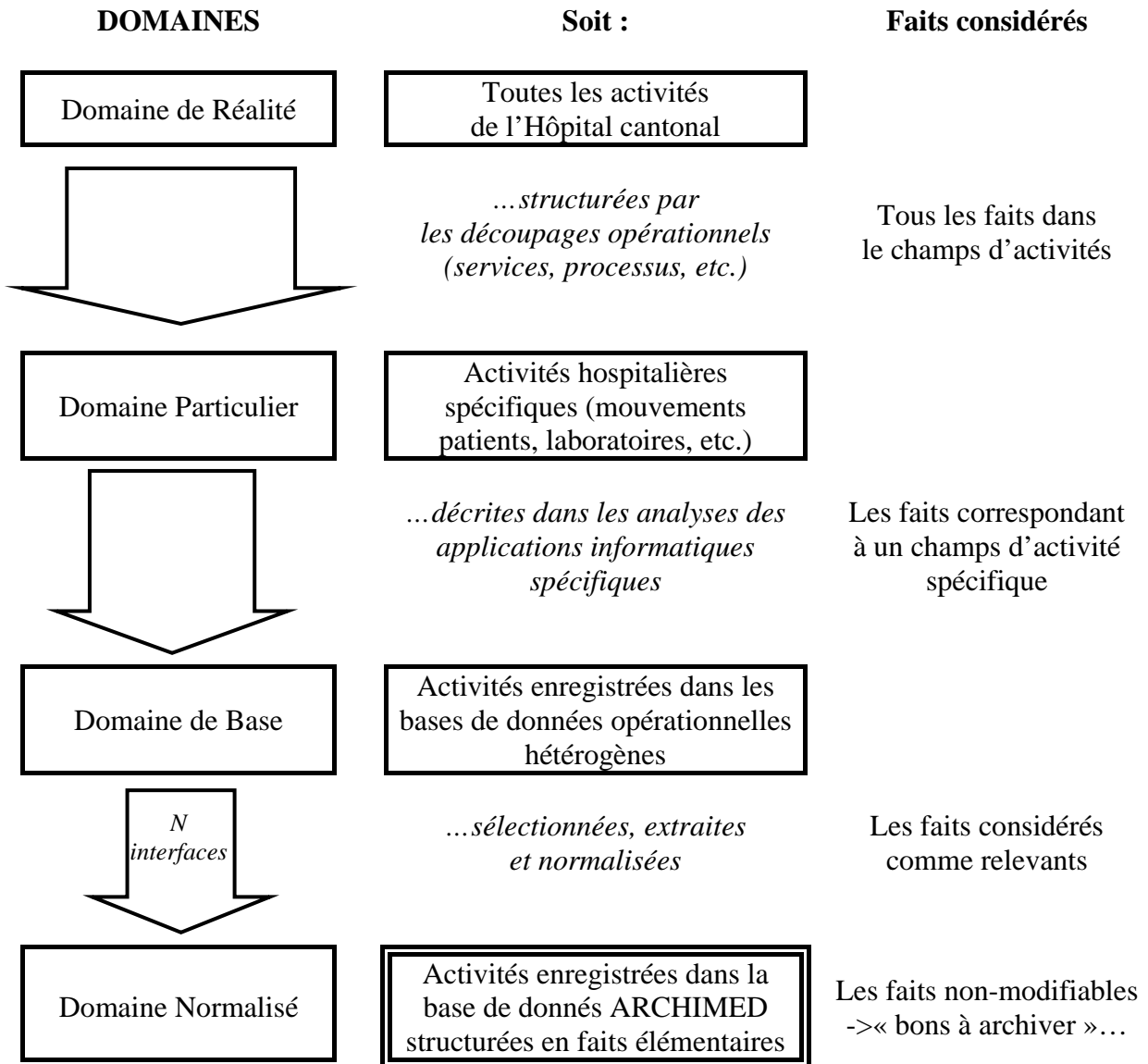
Si le concept se rapproche des notions actuellement bien connues des « entrepôts de données » (Data Warehouses) il s'en distingue dans la mesure où les systèmes de data warehousing s'occupent plus d'offrir une **présentation comparable** des données sélectionnées issues de différents systèmes opérationnels, en vue de leur consultation par les gestionnaires, tandis qu'ARCHIMED effectue de surcroît un travail **d'harmonisation et d'homogénéisation des données au niveau syntaxique, sémantique et ontologique** (voir plus bas chap. 4.1) pour constituer une base de données intégrée. Cette base de données constitue le noyau d'un **système** complété par une série d'outils d'aide à la décision.

Pratiquement, les données préalablement sélectionnées des différents systèmes informatiques opérationnels, sont filtrées quotidiennement par une série d'interfaces (8 à ce jour), qui les transforment en faits élémentaires puis les déversent dans la base de données ARCHIMED sur une base annuallisée (voir schéma en annexe). L'annuallisation a été imposée au départ par des contraintes techniques de place mémoire et de gestion des applications. Avec les moyens actuels elle ne serait plus nécessaire, et les tables annuelles pourraient être regroupées en une seule grande table. (*note de l'archiviste : cependant cette organisation, si elle implique certaines contraintes lorsque l'on effectue des requêtes sur des données pluri-annuelles, représente un mode de stockage intéressant sur le long terme car il permet une maîtrise prévisible des volumes de données et une rupture temporelle claire et explicite si l'on devait mettre certaines données off-line. Voir plus loin Chap 6).*

¹ On trouvera une description historique de ces travaux dans : **ARCHIMED : A Network of Intergrated Information systems**, G. Thurler, F. Borst, C. Bréant, D. Campi, J. Jenc, B. Lehner-Godinho, P. Maricot, J.R. Scherrer, in : *Methods on Information in Medecine*, Schattauer Verlag, 2000, vol. 39, no 1, pp. 36-43.

4.1. Le modèle ontologique d'ARCHIMED

Le modèle ontologique conçu pour construire l'application repose sur structure de domaines se superposant en couche d'abstraction successive, tel qu'illustré sur le schéma ci-dessous



Ce schéma montre que s'il s'agit bien, sur le plan technique, d'extraire des données des bases de données opérationnelles, la cohérence de ces données doit être assurée en amont. D'une part en fonction des définitions adoptées lors de la constitution des applications opérationnelles et d'autre part en fonction des « découpages » des unités opérationnelles définies indépendamment des applications informatiques mais dont celles-ci doivent tenir compte. Il existe par ailleurs d'autres découpages, liés à la gestion budgétaire ou à la localisation physique par exemple.

A partir de ce modèle, on a défini des *liens de base* qui ont été déduits en examinant les relations ou les similarités entre les faits contenus dans les différents domaines de base (par ex : patient, service médical, centre budgétaire, etc.)

A ces liens de base sont attachés des *propriétés*, caractérisées par un type, une valeur, un temps. La connexion entre un lien de base et une propriétés forme un *fait élémentaire*. On a alors l'équation suivante :

$$\text{Lien de base} + \text{propriété} = \text{fait élémentaire}$$

A ce niveau, il faut noter que chaque faits élémentaire est estampillé d'un **temps**. Celui-ci peut être un temps ponctuel, quand le fait n'a lieu qu'un fois (date d'examen par ex.), ou une durée avec ses deux limites de début et de fin (durée de séjour par ex.), celle-ci est alors représentée par un processus avec un fait de début et un fait de fin (ce qui formalise la notion de trajectoire hospitalière, mais permet, plus génériquement, de définir n'importe quelle chronologie de manière stricte).

Il faut également remarquer que les faits ne sont sélectionnés que lorsqu'ils ne sont plus modifiables. Cela correspond partiellement à la notion de clôture de dossier en archivistique classique, avec un nuance de taille : c'est le « fait » qui est clos et cette clôture s'effectue indépendamment d'une notion de dossier, bien que celle-ci existe par ailleurs dans certains domaines de base. Ceci est rendu possible sans perte d'information et de contexte parce que chacun des faits élémentaires est d'une part daté pour lui-même et d'autre part « liés » à un domaine de base.

4.2. *Pratique quotidienne*

Dans la pratique, une fois les domaines de base définis, des routines d'extraction sont appliquées aux différentes bases opérationnelles hétérogènes, en général sur un rythme quotidien pendant les heures creuses nocturnes. Ces tables, issues pour la plupart de base INGRESS sont extraites en fichiers ASCII et envoyées par protocole FTP vers l'application ARCHIMED. Elles sont alors distribuées dans des « paniers », en fonction de leur domaine de provenance. Ces paniers sont traités pour transformer les données en faits élémentaires normalisés, qui sont déversés dans les tables de la base de données ARCHIMED. Par ailleurs, un traitement parallèle sur les faits élémentaires permet de produire des « données réduites » qui sont en fait des indicateurs dérivés. (voir schéma de la constitution de la base de données ARCHIMED en annexe).

4.3. *L'accès aux données ARCHIMED*

Dans un premier temps, les gestionnaires d'ARCHIMED ont développé des requêtes d'interrogation similaires à celles préalablement utilisées pour produire les statistiques à partir des applications opérationnelles. Celles-ci étaient éditées sous la forme d'atlas statistiques annuels en mensuels, ainsi que d'autres atlas particuliers, en fonction des demandes. Ces éditions ont été remplacées par un navigateur, disponible sur l'Intranet des HUG, qui permet aux collaborateurs de consulter ces données communes de manière rétrospective (ce qui était difficile avec les atlas publiés sur papier) et avec une actualisation quotidienne (ce qui était impossible avec des éditions annuelles et mensuelles). Ces *statistiques d'activités* sont affichables à différentes échelles temporelles (années, mois, jour en général) et à différents niveaux d'agrégation (des HUG en entier jusqu'à l'unité de soins). Les tableaux sont exportables vers un tableur excel pour un traitement localisé.

La véritable valeur de la base de données intégrée d'ARCHIMED n'apparaît cependant qu'avec les différents outils d'aide à la décision développés ultérieurement par l'Unité d'information médico-économique (UIME) qui permettent diverses interrogations sophistiquées. La plupart se présentent sous forme de navigateur, nécessitant un droit d'accès.

Les outils disponibles sont, par exemple :

- le **calcul d'indicateurs** (DRG :diagnosis related groups, ré-admissions)
- les **archives des faits des patients** (recherche de cas similaires, recherche des faits d'un patients, données de laboratoires)
- les **statistiques médicales** (code diagnostic, interventions, statistique selon normes OFS, etc.)
- les **services Archimed** (analyse des mouvements d'urgence, calcul des scores de gravité, suivi des patients sur plusieurs années, etc.)²

5 Etat actuel de la base ARCHIMED

A fin 1998 la base de données ARCHIMED était constitué de:

- 5 MB de données (faits) incorporés quotidiennement dans la Base de Données Intégrée (BDI).
- Distribuées dans 750 tables, incluant 50 millions d'enregistrement, pour un total de 8 GB.
Ces données couvrent les activités hospitalières depuis 1990, la plupart des activités de laboratoire depuis 1993 et les activités ambulatoires depuis 1996.

Après dix ans d'activité (1993-2003) les évolutions annuelles sont les suivantes :

- Accroissement d'environ 1,5 Gigabytes par an (5MB/j * 365 j. = 1,8 GB)
- 70 million d'enregistrements (faits élémentaires)
(soit l'ajout de 200 tables annuelles)

6. Une base de données historique(s) ?

Comme expliqué plus haut, le concept qui a présidé à la naissance d'ARCHIMED a été d'ordre médical et opérationnel. Cependant, la nécessité d'harmoniser les données provenant de plusieurs systèmes opérationnels différents, a forcé ses concepteurs à une réflexion ontologique qui a mené à une structuration qui rencontre les préoccupations d'une conservation des données à long terme.

Actuellement, ARCHIMED représente non seulement un outils d'aide à la décision, raison par laquelle il a été conçu, mais également une source de données historique sans égale, car les données consolidées dans cette base unique ne pourraient être rassemblées autrement, chaque système opérationnel (et il y en a environ une centaine au sein des HUG) ayant sa propre structure et ses propres dictionnaires.

Cependant, ARCHIMED n'est pas une base de données historique « idéale » pour les raisons suivantes :

a) *Un outil avant d'être un système*

ARCHIMED a été initialement conçu comme un outil technique, dont l'objectif était de simplifier l'accès à des données provenant de bases de données hétérogènes. Bien que très vite ses concepteurs aient appréhendés ses possibilités d'outil d'aide à la décision, ce n'est qu'après la migration des systèmes informatiques en 1995 et la mise en place d'un plan d'investissement de grande envergure en 1998 qu'ARCHIMED apparaît pour la première fois comme une application identifiée au sein de la division informatique et de l'institution.

² On trouvera des descriptions plus détaillées de ces outils dans les articles suivants: **Retrieval of Similar Cases using the ARCHIMED Navigator**; Lehner B, Thurler G., Bréant C.; Tahintzi P. Borst F.; MIE, 2003, et **Toward a Systemic Approach to Disease**; Thurler G., Bréant C., Lehner B., Bunge M., Samii K., Hochstrasser D., Nendaz M., Gaspo J.M., Tahintzi P., Borst F.; ComPlexUs, Karger 2003.

b) *Un manque de vision institutionnelle*

Conséquence du point précédent, la hiérarchie, tant informatique qu'administrative, n'a réalisé l'intérêt de ce système d'information que lorsqu'il a été rendu visible sur l'intranet institutionnel. N'ayant pas participé directement à sa conception, les décideurs ont mis longtemps à reconnaître sa valeur.

c) *Une normalisation à posteriori*

Bien que certains dictionnaires de données soient communs à toutes les applications opérationnelles, comme les découpages, certaines harmonisations ne sont effectuées que lors du transfert des données dans ARCHIMED. Il manque à l'institution une instance qui définirait certains référentiels communs de manière univoque et uniforme, permettant une normalisation en amont.

d) *Une absence de politique de conservation à long terme*

Le système ARCHIMED représente surtout aux yeux de ses concepteurs et de ses utilisateurs un outil d'aide à la décision basé sur des périodes longues (10 ans) mais qui sont relativement courtes en terme d'archivistique. Jusqu'à présent, la place mémoire n'ayant pas fait défaut, la question de la conservation ne s'est pas posée, l'intérêt étant d'offrir en ligne le plus de données possible. Ce contexte est également valable pour les données des bases opérationnelles. La nature même de l'activité hospitalière portant sur la totalité de la vie des patients, cette tendance restera une constante. De plus en plus de données se trouvant nativement dans les systèmes informatiques des HUG, nous entamons seulement maintenant, secteur par secteur, des discussions sur la conservation (ou la non-conservation) à long terme de ces données.

Nonobstant ces réserves, ARCHIMED représente un intérêt historique considérable, pour les raisons suivantes :

a) *Une validation préalable des données*

L'analyse ontologique et sémantique des données déversées dans ARCHIMED implique que les données ainsi conservées ont été considérées comme pertinentes à la base. Ceci évitera un travail d'évaluation supplémentaire lors d'un futur archivage historique. On devra cependant veiller à ce que ces critères de validation soient explicitement documentés.

b) *Une structure simple et documentée*

La gestion des données en faits élémentaires et en tables annuelles rendent leur manipulation très indépendante des logiciels de gestion de base de données. Chaque type de relation étant documenté dans des dictionnaires il ne serait par exemple pas difficile de conserver ces tables sous forme XML. *Dans la perspective d'une conservation à très long terme, on doit cependant se poser la question de la conservation des navigateurs (qui donnent une image de l'usage actuel de ces données) et la possibilité de construire à long terme d'autres navigateurs, répondant à des questions d'ordre historique plutôt que médico-économique. Les HUG ne se sont pas encore prononcés à cet égard.*

c) *Une structure indépendante du temps*

Comme chaque fait élémentaire est daté et qu'il n'est transféré dans la base de donnée intégrée qu'une fois qu'il n'est plus susceptible de changement, on évite un problème récurrent dans les entrepôts de données courants, qui est celui de la mise à jour de données. De ce fait, la base est parfaitement cohérente dans le temps. La structuration en tables annuelles permet potentiellement une mise off-line par tranche chronologique sans aucune manipulation supplémentaire (les trajectoires de soins qui « passent » d'une année sur l'autre sont signalées

dans les tables par un drapeau, ce qui permet leur identification et la concaténation des données entre les tables annuelles).

Conclusion

Dans un article de 2001, Edward Atkinson défend la proposition selon laquelle les Data Warehouses (et ARCHIMED peut y être assimilé) sont des records, et sont à ce titre digne d'être conservés³. Dans son article la justification de sa position n'est pas très étayée. L'exemple de la base de données ARCHIMED peut fournir au moins un argument de taille. Si les entrepôts de données sont strictement documentés chronologiquement au niveau du fait élémentaire, ils représentent une source historique de première qualité.

En conclusion on peut affirmer que les entrepôts de données sont des records (historiques) si :

- ils sont strictement documentés (ontologies)
- ils sont strictement délimités (faits élémentaires)
- ils sont strictement datés (attributs temporels)
- ils permettent la construction de nouveaux critères de navigation

Auteur:

Jean-Daniel Zeller

Archiviste principal

Hôpitaux universitaires de Genève

Rue Micheli-du-Crest 24

CH-1211 Genève 14

Tel: +41 (0)22 372.60.31

Fax: +41 (0)22 372.60.30

Mail: jean-daniel.zeller(at)hcuge.ch

³ Dans: **Data warehousing-a boat records managers should not miss**, Records Management Journal, vol. 11, no. 1, avril 2001, pp 35-43.

Schéma de constitution de la base de données ARCHIMED

