Guide Méthodologique

Pour l'archivage des bases de données

Auteurs : Philippe Prat, Lorène Béchard
Version : 1.1
Date : 16/04/2013
Numéro Document : GM_archivage_BDD-1.1
Confidentialité : Public

SOMMAIRE

1. Objectifs

Les bases de données représentent un élément essentiel de nos systèmes d'information. Elles en sont la mémoire. Au delà de la simple sauvegarde, il est nécessaire de s'interroger sur leur conservation, afin de permettre à cette mémoire de perdurer et de rester exploitable sur le long terme. Mais cette tâche est d'autant plus ardue que les données sont fortement imbriquées au sein des systèmes d'informations, des organisations et des hommes qui sont eux, toujours en constante évolution.

Actuellement, il existe peu ou pas de cadre précis concernant l'archivage des bases de données, que ce soit au niveau technique, organisationnel ou administratif. Pourtant l'enjeu est de taille dans la mesure où les bases de données sont au cœur de nos systèmes d'informations et où les évolutions technologiques nous amènent à penser que cela devrait continuer dans ce sens.

L'objectif de ce guide est de proposer un support aux acteurs confrontés à la problématique de l'archivage des bases de données. Il s'adresse aux archivistes, aux informaticiens, aux décideurs et plus généralement à toute personne intéressée par cette question. Ce guide est un outil méthodologique et de dialogue afin que chaque acteur puisse comprendre au mieux les problématiques et les enjeux d'un projet d'archivage de bases de données. Il apporte par exemple des éléments aux archivistes pour comprendre les contraintes auxquelles sont assujetties les informaticiens et réciproquement. Il propose également une vision globale aux décideurs, leur permettant d'initier ou pas de tels travaux. Ce guide fournit des méthodes pour vous aider à identifier les points importants à prendre en compte lors du processus d'archivage des bases de données.
Une base de données est souvent étroitement liée à une application qui l'exploite afin d'effectuer des traitements pour une personne ou une organisation. Aussi il nous semble nécessaire d'étendre notre étude afin d'y inclure les contextes applicatifs, métiers et organisationnels.

Ce guide est le fruit d'une réflexion commune entre archivistes et informaticiens sur la base de l'expérience capitalisée par le CINES en archivage électronique. Confronté à la réalité du terrain, il évoluera en fonction des retours d'expériences que nous souhaitons nombreux.

1.1. Périmètre de l'étude

Cette étude se focalise sur l'archivage des bases de données de type relationnelles. Nous ne connaissons pas de solution simple, applicable de manière systématique pour les archiver. Il s'agit plutôt de savoir identifier les risques de perte d'information, les obligations de conservation et de prendre en compte les droits sur la donnée contenue dans une base de données.
Ensuite, il est important d'identifier les moyens disponibles (technique, matériel, financier) pour effectuer un archivage.
Toutes ces questions doivent être abordées en fonction de l'objectif de l'archivage. Comme toute action, le préalable reste donc de bien identifier l'objectif : pourquoi archive-t-on et quels seront les utilisations potentielles de cet archivage ?

2. Préambule

2.1. La conservation à long terme de documents numériques

Ces dernières années ont vu un accroissement considérable de l'information numérique produite. Dans le même temps, nous sommes témoins du caractère de plus en plus volatile de cette information, à l'image du nombre de liens morts rencontrés sur le web. Les risques liés au numérique sont aujourd'hui bien identifiés : l'obsolescence du matériel, la disparition du logiciel de lecture, l'obsolescence du format du fichier et la perte de la signification du contenu informationnel.
Face à cela, les bonnes pratiques sont connues : copies multiples sur différents types de supports, veille sur les technologies existantes, sélection des formats de fichiers avant archivage, utilisation de métadonnées pour documenter les informations archivées... A partir de là, il apparaît clairement que la seule sauvegarde des données, même sécurisée, n'apporte qu'une sécurité actuelle des données et ne constitue en aucune manière une garantie de pérennité dans le temps, tant sur la lisibilité du support que sur la capacité à restituer un contenu compréhensible.
Or, du fait notamment des évolutions législatives qui accordent au numérique la même valeur probante que l'écrit sous certaines conditions, l'enjeu de la conservation des données numériques est crucial. De plus, avec la dématérialisation des procédures, ces données nativement électroniques, n'existent d'ailleurs pas toujours sous la forme d'un fichier. Il s'agit souvent de données intégrées au sein d'une base de données.
En somme, assurer une conservation pérenne de ces données, c'est conserver l'information contenue dans son aspect physique comme dans son aspect intellectuel de manière à pouvoir la rendre accessible et compréhensible aussi longtemps que nécessaire.

2.2. L'organisation de l'information numérique

Une information est le sens que l'on peut tirer de l'exploitation d'une donnée. Une donnée n'a aucune signification en soi. Elle peut se présenter sous différentes formes : numérique, matérielle ou abstraite. Actuellement, toute donnée numérique dans son aspect le plus primitif est constituée d'une suite d'éléments binaires, c'est-à-dire pouvant posséder deux états distincts (0 ou 1). Cela pourrait en être autrement, comme le codage du vivant utilisant des chaines d'ADN composées de 4 éléments distincts ou le morse qui possède deux états, un son de courte durée et un long. Il existe également une infinité de manière de décrire quelque chose en mode binaire comme il existe de nombreux langages sur terre pour permettre aux hommes de communiquer. Aussi, comme il est nécessaire d'identifier et de connaître le langage de son interlocuteur, il est également nécessaire de connaître la manière dont une donnée est organisée, ce qui en informatique correspond à son format.

Un document numérique sert de support à des données, il en est le contenant. Il se présente généralement sous la forme d'un ou de plusieurs fichiers que l'on peut exploiter en utilisant un logiciel informatique.

On peut opposer à cette organisation sous forme de fichiers, relativement statique, une organisation plus complexe dans laquelle le contenu auquel on veut accéder n'est plus rassemblé dans un ou plusieurs fichiers directement exploitables par un logiciel mais sous une forme qui peut impliquer de nombreux processus et qui sont différentiables en plusieurs couches ; parmi lesquelles la couche « donnée » sur laquelle nous porterons notre attention.

L'obsolescence du fichier ? Il nous parait important de noter que les systèmes d'informations actuels sont de plus en plus organisés selon des architectures complexes. Il suffit pour cela d'observer l'essor des systèmes informatiques nomades (tablettes numériques, smartphones, etc.).
Afin de proposer de nouvelles fonctionnalités, les fichiers deviennent de plus en plus complexes et ont inévitablement tendance à être remplacés par des formes plus structurées comme des bases de données qui stockent leur contenu. Ceci pour permettre à des applications distantes d'y accéder selon différentes vues : l'album photo accède à la base de données des photos, le lecteur multimédia accède à la base de données des fichiers musicaux, etc.
Prenons l'exemple des données d'un listing de numéros de téléphone. Elles seront difficilement exploitables si elles sont saisies dans un fichier de type tableur. L'usage actuel tend plutôt vers

l’utilisation d’une application pour saisir le nouveau numéro, consultable ensuite depuis un PC, un smartphone ou via le web. De la même manière, il est possible de consulter ses mails sur un ordinateur ou sur un smartphone.
La vue, c'est-à-dire l’apparence graphique, est différente mais le contenu est le même, et la donnée n’est pas dans un fichier. D’ailleurs, où est-elle vraiment ?

3. Concepts archivistiques applicables

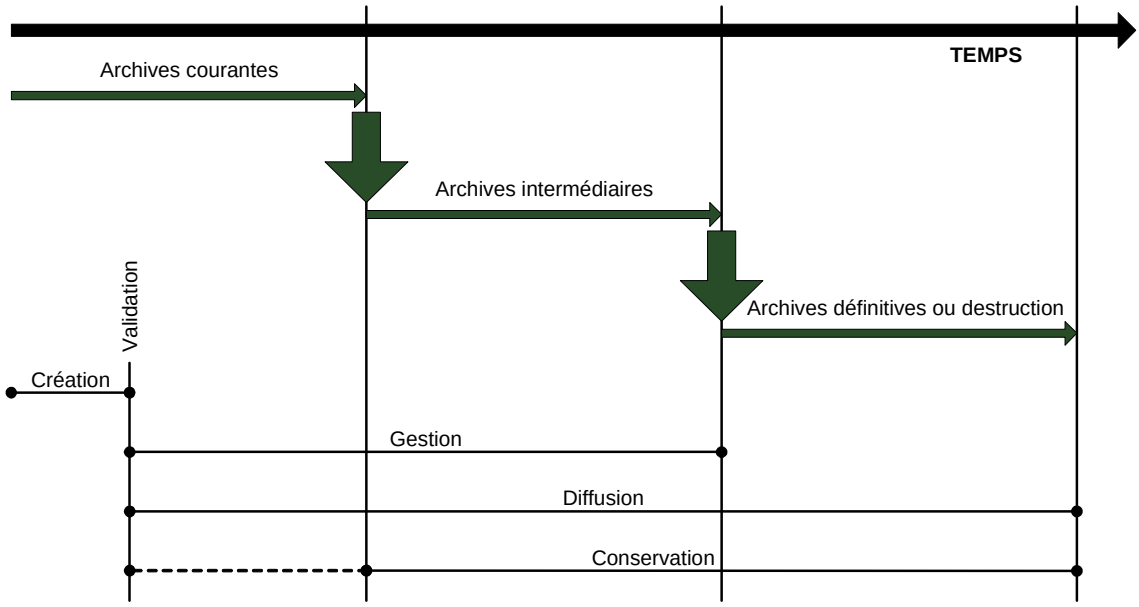
3.1. Le cycle de vie de la donnée

Le vocabulaire employé ici fait volontairement référence au terme de « donnée » plutôt qu'à celui de « document » ou d'« archives » au vu du contexte des bases de données que nous nous proposons d'étudier dans ce guide. Ces termes sont toutefois interchangeables dans le sens où une donnée, un document ou une archive contiennent une information. De manière plus générique, on pourrait alors parler du cycle de vie de l'information, qu'elle soit représentée sous la forme d'une donnée ou d'un document.

Toute donnée, quelle que soit son support, une fois qu'elle est créée, suit un cycle de vie en fonction de l'usage pour lequel elle a été créée et éventuellement d'exigences juridiques. Ce cycle de vie est matérialisé par des phases successives :

- La donnée est fréquemment utilisée pour les besoins pour lesquels elle a été créée, qu'elle soit en cours de création ou validée.
- La donnée a été validée, son utilisation est plus ponctuelle (elle sert de référence par exemple), elle possède parfois une valeur probante. Elle est dite « semi-active » ou « d'âge intermédiaire ».
- La donnée n'est plus utilisée pour les besoins pour lesquels elle a été créée, elle n'a plus de valeur probante, mais elle revêt toutefois un intérêt à être conservée au titre de l'histoire. Elle est dite « d'âge définitif ».

Dans le cas où elle n'aurait pas d'intérêt historique, elle peut être éliminée conformément aux procédures règlementaires en vigueur (pour les archives publiques, se référer au Code du Patrimoine). Toutes les données n'ont donc pas vocation à être conservées définitivement.



L'identification de ce cycle de vie est effectuée par l'archiviste en étroite collaboration avec le producteur de la donnée et conformément au cadre juridique applicable.

Il s'agit de déterminer pour chaque type de données une Durée d'Utilité Administrative (DUA), c'est-à-dire, « la durée légale ou pratique pendant laquelle un document est susceptible d'être utilisé par le service producteur ou son successeur, au terme de laquelle est appliquée la décision concernant son traitement final. Le document ne peut être détruit pendant cette période qui constitue sa durée minimale de conservation »[1]. La DUA couvre donc les âges courant et intermédiaire.

Un certain nombre de questions permettent d'évaluer cette DUA : au bout de combien de temps la donnée est-elle validée et n'est plus modifiée ? Pendant combien de temps utilise-t-on la donnée

[1] Définition extraite de l'Abrégé d'archivistique, publié par l'Association des Archivistes Français en 2007.

régulièrement ? Existe-t-il des recommandations de DUA pour ce type de données ? Existe-t-il des obligations légales de conservation ?

Identifier le cycle de vie de la donnée va permettre d'en organiser l'archivage, intermédiaire dans un premier temps, puis éventuellement définitif à l'issue de la DUA.
Or, le contexte du numérique bouleverse quelque peu les pratiques autour de la gestion du cycle de vie. En effet, dans le cadre traditionnel du papier, les documents d'âge courant ne font pas l'objet d'un archivage ; celui-ci intervient seulement lorsque le document passe à l'âge intermédiaire. Lorsqu'il s'agit de données ou de documents électroniques, la prise en charge doit s'effectuer au plus tôt, dès que la donnée est validée, afin de maintenir sa valeur probante et de limiter les risques d'obsolescence technologique et d'inintelligibilité.
Ainsi, l'intervention d'un archiviste est nécessaire le plus en amont possible afin de mettre en œuvre les moyens nécessaires à la préservation de l'intégrité, de l'authenticité, de la lisibilité et de la compréhension de la donnée.

3.2. Les métadonnées

Les métadonnées ne sont pas seulement la carte d'identité d'un document. Elles permettent entre autres de l'identifier, de le décrire, d'expliquer l'origine de sa création, son utilité et ses destinataires.
Il existe plusieurs types de métadonnées :

- les métadonnées de gestion, pour accéder au document ;
- les métadonnées de description, pour rechercher le document et en comprendre le contenu ;
- les métadonnées de préservation, pour garantir la pérennité de l'accès et de la compréhension du document.

Sans tous ces éléments, un document peut vite devenir incompréhensible et donc inexploitable.

Dans le cas des bases de données, la collecte et l'archivage de métadonnées est indispensable pour garantir la compréhension des données à travers le temps. Par exemple, comment comprendre une succession de chiffres dans un tableau, s'il n'est pas précisé à quoi ils correspondent, qui a fait ce document, pour qui, dans quel but ou encore à quelle date.

3.3. La règlementation autour de la confidentialité

En France, la protection des données et de leur contenu informationnel est un élément important traité à plusieurs niveaux :

✓ Tout d'abord, les informations personnelles ou nominatives font l'objet d'une attention particulière, notamment dans le contexte de l'informatique et de l'internet. Tout traitement informatique de données personnelles doit être encadré par la CNIL (Commission Nationale de l'Informatique et des Libertés) dans la mesure où il est obligatoire d'en faire la déclaration auprès de cet organisme. Le droit à l'oubli consacré par la « loi informatique et libertés » prévoit la suppression des données personnelles une fois que les besoins pour lesquels elles ont été collectées sont satisfaits. Bien que cela puisse a priori rentrer en contradiction avec l'obligation d'archivage, le droit à l'oubli doit être aménagé afin de ne pas nuire à la collecte des données historiques.

✓ Par ailleurs, dans le cas des archives publiques, le régime de communication du Code du Patrimoine (art. L213) instaure une communicabilité de plein droit. Des exceptions sont toutefois prévues en fonction de la nature des informations contenues dans les documents, notamment les informations pouvant porter atteinte à la vie privée. Cela se traduit par la mise en place de délais (25, 50, 75, 100 ans) pendant lesquels les documents ne sont communicables qu'au service producteur ou suite à une demande de dérogation.

Les bases de données contiennent souvent des informations nominatives. Il est donc important de respecter le cadre légal aussi bien lors de l'archivage que lors de la communication des informations archivées.

4. Les bases de données

4.1. Les types de bases de données

Une base de données est un ensemble de données suffisamment organisé pour en permettre une exploitation cohérente. Les données peuvent être organisées selon différents modèles dont voici quelques exemples :

Le modèle hiérarchique associe les données uniquement via des relations de composition de type « parent-enfants » : un livre est composé de chapitres, eux-mêmes composés de plusieurs paragraphes qui sont composés de plusieurs phrases, mots et enfin de lettres.

Le modèle relationnel organise les données sous forme de tableaux (appelés « tables ») qui peuvent être reliés les uns aux autres par des relations (dites « relations de jointure »). Des éléments particuliers de ces tables comme les clés primaires et clés étrangères permettent de manière préférentielle de créer des liens entre des tables. Les clés étrangères garantissent la cohérence du schéma général.

La manière par excellence d'exploiter les données d'un système de gestion de bases de données (SGBD) relationnel est le « langage structuré de requêtes », le Structured Query Language ou SQL. Il permet d'interagir avec les données d'un SGBD relationnel. Cette interaction peut se faire en lecture et en écriture. Le SQL utilise les jointures des différentes tables pour obtenir un résultat. Il permet donc d'écrire sous une forme informatique compréhensible par le SGBD la demande d'un utilisateur pour répondre à un besoin donné.
Exemple de requête SQL pour une base de données d'une bibliothèque contenant une table « Livres » avec un champ « Auteurs » et un champ « Titre » : SELECT AUTEURS FROM LIVRES WHERE TITRE IS « Guide sur les bases de données » qui va permettre de trouver les auteurs du livre « Guide sur les bases de données ».
De plus en plus de SGBD peuvent gérer de manière spécifique certains types de données comme des dates ou des données XML à des fins de comparaison ou d'exploitation au moyen de langages tels que XPath et XQuery.

Le modèle Objet ne stocke pas des données de type simple comme un entier, une date ou un texte mais des objets issus du monde de la programmation orienté objet. Un objet est un ensemble de données et de fonctions représentant une entité cohérente (ex : une voiture possède 4 roues, avance, recule etc..).

Il existe d'autres modèles de SGBD qui peuvent être spécialisés pour une fonction précise. Notons par exemple le **RDF** (Resource Description Framework) qui est un formalisme relativement récent permettant de positionner des informations sur un objet à l'aide d'un triplet : sujet, prédicat, objet.

Le sujet permet d'identifier la ressource, le prédicat d'attribuer une propriété ou une relation à la ressource et l'objet de donner une valeur à cette relation.
Par exemple : L'archive x *(sujet)* a été produite par *(prédicat)* l'administration y *(objet)*.
Il existe un langage, le SPARQL, permettant au même titre que le SQL pour un SGBD classique, d'interagir avec ces données. Ce type d'organisation de l'information peut être qualifié de modèle de graphe ou modèle descriptif.

Enfin, les **entrepôts de données** sont de véritables bases de données de bases de données. Ils permettent de rassembler au sein d'un même système d'information, des données issues d'entités ou de métiers très différents. Les données concernées sont souvent peu volatiles car elles n'ont pas ou plus vocation à être modifiées. Ces systèmes privilégierons donc l'aspect recherche et recoupement de données plus que l'aspect transactionnel (modification, suppression). Cela permet par exemple de faire du « Data Mining » pour rechercher des « métaInformations », c'est-à-dire des informations

découlant de corrélations non évidentes comme la relation entre une maladie ou un comportement et une séquence d'ADN.
Dans la mesure où de tels systèmes gèrent une grande quantité de données hétérogènes, il est nécessaire de les standardiser pour faciliter ou rendre possible leur exploitation. Aussi, lors de leur versement dans l'entrepôt, il faudra par exemple transformer un champ date préalablement représenté par une chaîne de caractères en un champ date au format ISO 8601.
Un service d'archivage à long terme de base de données pourrait donc proposer un service de type entrepôt de données dans lequel le dépôt des tables, la sélection des champs et le mode d'interrogation seraient négociés entre le service d'archives et le producteur.

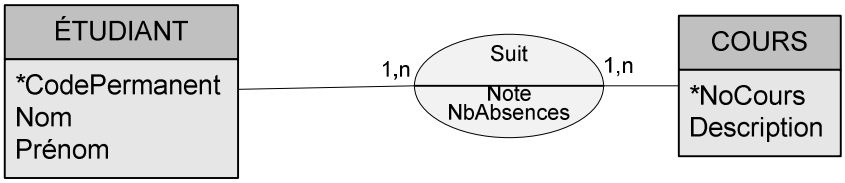
4.2. Décrire des bases de données

Il est possible de décrire une base de données selon les règles de description des documents d'archives. Cependant, du fait de sa complexité, elle demandera davantage de précision. Il existe en effet plusieurs niveaux de description : de son utilisation générale jusqu'à la description technique de chaque élément (table, champ, relations...).
Pour cela, nous nous inspirons de la méthode Merise[2] qui distingue 3 niveaux de description :

Le modèle conceptuel des données (MCD) :

Il s'agit de décrire des entités (ensemble d'objets ayant des attributs identiques) et des relations (association ou actions entre les entités). Chaque objet doit être identifiable.

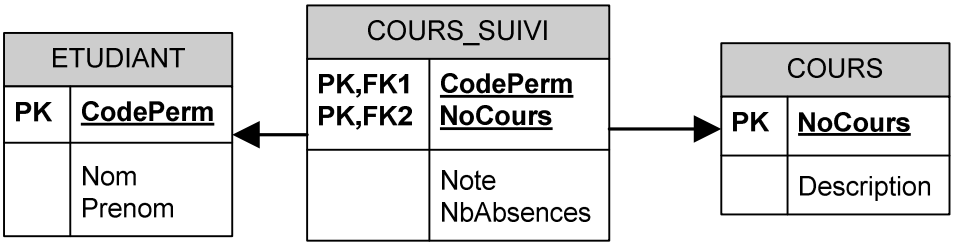
Ci-dessous, la représentation conceptuelle d'un étudiant qui possède un nom et un prénom et qui suit des cours. Le suivi de chacun de ces cours donne lieu à des notes et un nombre d'absences.



Le modèle logique des données (MLD) :

Une fois le MCD établi, on peut le traduire en différents systèmes logiques, comme un ensemble de fichiers binaires, un fichier XML particulier ou au sein d'un SGBD.
Il est relativement simple de transposer un modèle conceptuel de données en un modèle logique dédié au SGBD (ou MLBD) : les entités correspondent en fait à des tables, les attributs des entités à des attributs de tables et les relations à des associations entre tables.

Ci-dessous, la représentation du modèle logique des données pour les cours suivis par un étudiant. On notera que ce schéma ajoute une précision par rapport au modèle conceptuel en décrivant la relation entre l'entité COURS et l'entité ETUDIANT à l'aide de clés primaires (PK) et étrangères (FK). Ces dernières sont le moyen de relier les tables entre elles.



2

Le modèle physique des données (MPD) :

Le MPD est une implémentation particulière du MLD pour un SGBD particulier. Il s'exprime en SQL avec des déclinaisons spécifiques au SGBD. On s'appliquera alors à choisir les bons types d'attributs pour les données identifiées dans le MLD, les bonnes relations, etc.…
L'intérêt du modèle conceptuel des données est de voir quels sont les concepts de haut niveau auxquels la base doit répondre. La séparation entre le MLD et le MPD permet entre autres le portage d'une base de données d'un SGBD vers un autre. Par exemple, on peut traduire un MPD MySQL en un MLDR puis traduire ce MLDR en un MPD postgres.

Le script ci-dessous représente une instruction de création de la table "étudiant" destinée à contenir les données relatives à l'étudiant. Cette instruction est directement interprétable par le SGBD.

```
CREATE TABLE ETUDIANT {
        CodePerm INT NOT NULL PRIMARY KEY
        NOM CHAR(20) NOT NULL
        PRENOM CHAR(20) NOT NULL
}
```

L'archivage doit s'intéresser à ces 3 niveaux de description car chacun apporte un niveau de représentation nécessaire à la compréhension des données conservées.

Au delà de cette méthode d'analyse, on peut dire que les données sont **:**

- Quantifiables : volume, nombre total de données stockées.
- Qualifiables : Elle contient des informations publiques, privées, confidentielles, stratégiques.
- Interdépendantes : la compréhension d'une donnée dépend de la présence d'autres données.

Ces éléments seront autant d'arguments qui nous aiderons dans le choix de notre stratégie d'archivage.

4.3. Les caractéristiques d'une base de données

Dans un contexte d'archivage, il est important de pouvoir caractériser l'usage qui est fait de la base de données. Ces caractéristiques seront de précieux indicateurs pour permettre par la suite d'identifier la meilleure stratégie d'archivage. Dans la mesure où il n'existe pas réellement de vocabulaire standard, nous utilisons ici des termes qui nous sont propres.

Une base de données est dite « **vivante »** si les éléments qui la constituent sont modifiés ou que de nouveaux éléments sont ajoutés. On parlera de base de données « **figée »** si aucune modification, ajout ou effacement n'ont été effectués récemment.

Une base de données est fortement **« consultée »** si un grand nombre de consultations est fait sur les données qu'elle contient.

Une base de données est dite « **cumulative »** si on ne fait qu'ajouter de nouveaux éléments sans en modifier et sans en effacer. De manière inverse, on parlera de base de données « **dynamique** » si l'ajout et la modification sont autorisés et utilisés.

Voici un tableau récapitulatif des points que l'on vient d'exposer :

	Lecture	Modification	Ajout	Suppression
Vivante	?	oui		
Figée	?	non		
Dynamique	?	oui		oui
Cumulative	?	non	oui	non
Consultée	oui	?		

Au delà des ces notions, il est important d'avoir conscience de l'environnement d'exploitation de la base de données. Une base de données n'a lieu d'être que si elle est liée à des applicatifs, à des établissements et au final à des personnes. Il sera nécessaire d'analyser et de cerner l'impact de cet environnement pour en intégrer tous les éléments nécessaires à l'exploitation future de la donnée.
Ces éléments constitueront alors des informations de représentation (au sens du modèle OAIS) qui aideront le futur utilisateur à comprendre et exploiter la base.

4.4. Inventaire des principaux SGBD existants

Il reste très difficile de pouvoir classer les SGBD : certains sont propriétaires comme le SGBD Oracle et d'autres libres de droit comme MariaDB, qui est la suite non propriétaire de MySQL racheté par Sun Microsystem.
Certains sont des applications à part entière et d'autres des composants que l'on doit greffer à des outils ou langages de programmation. Il existe également des SGBD, comme *OpenOffice Base,* construits pour être utilisés comme un logiciel bureautique par un utilisateur sans grande compétence en SGBD.

Dans une perspective d'archivage, on privilégiera les SGBD libres de droit même s'il reste difficile de passer d'une base relationnelle à une autre. Il est même probable que les SGBDR ne seront jamais totalement interchangeables dans la mesure où leurs fonctionnalités sont de plus en plus évoluées et s'éloignent de la norme SQL.

Vous trouverez ci-dessous une liste des principaux SGBD disponibles :

Nom	Année	Editeur	Caractéristiques	Type	Licence
MariaDB	2009	*Monty Program Ab*		serveur	GPL
OpenOffice.org Base	2002	Oracle Corporation		Logiciel applicatif	LGPL
HSQLDB	2000	*Thomas Mueller*	relationnel, embarqué, centralisé, pour groupes de travail et particuliers	Composant logiciel	BSD
SQLite	2000	*D. Richard Hipp*	embarqué	composant logiciel	Domaine public
Caché	1997	InterSystems	objet, pour entreprises, distribué	serveur	propriétaire
Apache Derby	1996	Apache Software Foundation	embarqué, relationnel, centralisé	Composant logiciel	Apache
MySQL	1995	Oracle Corporation et MySQL AB	centralisé, embarqué, distribué, pour entreprises, groupes de travail et particuliers	serveur	GPL
HyperFile	1993	PC Soft		composant logiciel	propriétaire
Microsoft Access	1992	Microsoft	relationnel, pour particuliers et groupes de travail	L4G	propriétaire
Microsoft SQL Server	1989	Microsoft	entreprises, groupes de travail, particuliers, relationnel, distribué	serveur	propriétaire
Paradox	1987	Corel		logiciel applicatif	propriétaire
FileMaker Pro	1985	*FileMaker*	relationnel, pour groupes de travail	logiciel applicatif	propriétaire
PostgreSQL	1985	Michael Stonebraker		serveur	BSD
DB2	1983	IBM	pour entreprises, groupes	serveur	propriétaire

			de travail, particuliers		
Firebird	1981	Firebird Foundation	relationnel, centralisé, embarqué, pour groupes de travail et entreprises	serveur	Interbase
Informix	1981	IBM	pour entreprises, groupes de travail, distribué	serveur	propriétaire
Progress	1981	Progress Software Corporation		L4G	propriétaire
Oracle Database	1979	Oracle Corporation	entreprises, groupes de travail, particuliers, relationnel, spatial, distribué	serveur	propriétaire
dBase	1978	Ashton-Tate	relationnel, pour particuliers	L4G	propriétaire
MaxDB	1977	SAP AG et MySQL AB	objet-relationnel, pour entreprises et groupes de travail, centralisé	composant logiciel	GPL
Ingres	1974	*Ingres Corporation*	relationnel, spatial, centralisé, distribué	serveur	GPL
Pick	1968	Pick System		serveur	propriétaire

Source : Wikipédia

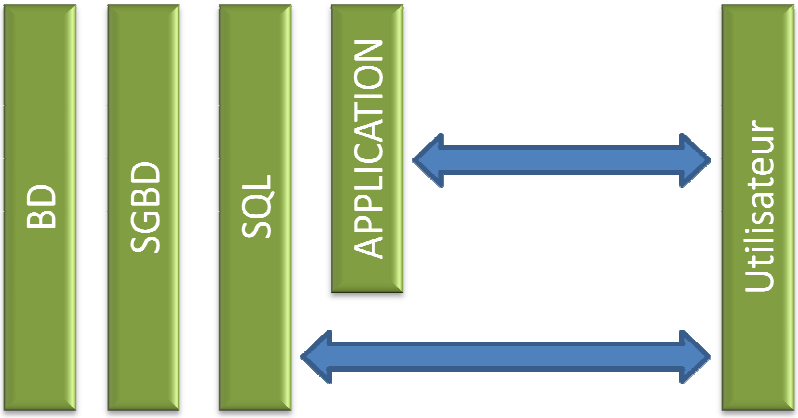
Les SGBD se renouvellent souvent plus vite que les données qu'ils contiennent. En effet, le cycle de vie d'une donnée n'est en rien corrélé avec celui du système informatique qui en assure la gestion. De manière générale, les SGBD proposent périodiquement une version améliorée (disons tous les 6 mois). Une donnée par contre peut avoir une pertinence qui s'étend sur une période de quelques millisecondes à la pérennité absolue.
Il n'y a donc aucun lien à faire entre la donnée et le SGBD qui la contient, si ce n'est de s'assurer que le SGBD à qui l'on souhaite confier la gestion de nos données offre des fonctionnalités et des garanties permettant d'être en accord avec les services que l'on souhaite avoir sur ces données.
Si une donnée a un cycle de vie plus long que la version du SGBD sur lequel elle est implémentée, alors il faudra se munir des compétences nécessaires pour enclencher un changement de version. Il reste important d'avoir un SGBD qui soit à jour afin de pouvoir bénéficier de la meilleure qualité de services apportée par les derniers correctifs.

4.5. Les bases de données au sein des systèmes d'information

Les bases de données sont au cœur de la plupart des systèmes d'information des entreprises et des administrations. Elles constituent, avec la documentation bureautique numérique et papier, la majeure partie de la mémoire d'un établissement.

Il n'y a pas de règles spéciales sur le nombre de bases de données dans une entreprise. Il peut y en avoir aucune, une ou plusieurs. Il n'est également pas rare de voir des bases de données externes à l'entreprise.

Ces bases de données peuvent être exploitées par des logiciels ou directement par des personnes via des commandes SQL.



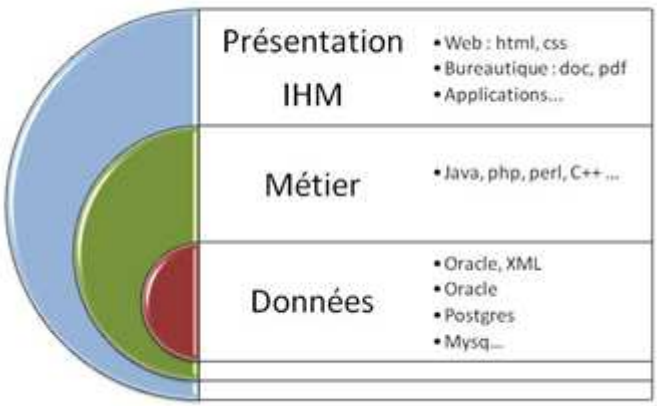
Les données dans les systèmes d'information

Les doubles flèches représentent les interactions possibles entre les données et l'utilisateur. Ces interactions peuvent se faire en lecture et modification, ce qui explique le double sens. Dans le cas où l'utilisateur n'interagit pas directement avec la base, l'application sera un élément essentiel pour comprendre les données. Il sera donc nécessaire lors de l'archivage de prendre en compte cet élément.

Pour qu'une personne puisse utiliser une base de données, il faut qu'elle connaisse la structure de la base, l'organisation et le sens des données contenues. Lorsque la base est utilisée uniquement par une application, il peut être très difficile de comprendre cette organisation. Dans ce cas, seul le résultat apporté par l'application permettra d'avoir une information intelligible. Par exemple, les sites Web de type CMS proposent des pages Web facilement compréhensibles dont les informations sont dans une base de données. Retrouver ces informations directement depuis la base de données est souvent très difficile, voire impossible du fait de l'importance de la couche de présentation.
On voit donc que le niveau qui fait sens peut se situer dans la couche applicative. C'est le cas par exemple d'un fichier au format PDF correspondant au salaire d'un employé ; les informations sur ce salaire pouvant par ailleurs être très complexes, réparties sur plusieurs bases de données et donc très difficilement archivables. On se posera alors la question de l'archivage des résultats produits par l'application plutôt que la base de données elle-même.

Les logiciels de création de documents bureautiques proposent de plus en plus d'interfaces avec des bases de données pour créer des rapports, faire du publipostage, etc. La part des pages web mises à jour à l'aide de bases de données prédomine sur internet alors qu'il y a dix ans elles étaient encore construites de manière statique.

De ce fait, il est important de prendre en compte dans un processus d'archivage l'environnement applicatif de la base de données. Ce qui nous amène à porter notre attention sur la configuration du système d'information et notamment sur les architectures informatiques appelées « 3-tiers », très utilisées actuellement.



Les 3 couches applicatives

Les applications interagissent sur des données pour produire un résultat. Ces traitements sont effectués par la couche dite « métier » car elle est en charge de la logique à mettre en œuvre pour satisfaire un service particulier. Cette couche métier va prendre et mettre à jour des informations qui sont situées au niveau de la couche dite de « données ». Elle constitue la mémoire morte de l'application qui lui permet par exemple de se souvenir du nom et du prénom d'un étudiant qui suit un cours à l'université. Enfin, pour interagir avec les différents utilisateurs de l'application, pour permettre à une personne de saisir un nom et un prénom et de pouvoir obtenir la fiche descriptive d'un étudiant, il est nécessaire d'avoir une couche dite de « présentation ».

Les réflexions autour de l'archivage du web ou des bases de données conduisent irrémédiablement à la problématique de l'archivage des systèmes avec des architectures complexes et multicouches.
Toutefois, dans ce contexte le choix d'archiver les éléments faisant partie de la couche de présentation peut vite montrer ses limites.

5. Les méthodes de sauvegarde des données d'une base

Plusieurs méthodes sont applicables pour sauvegarder des données. Nos propos s'attacheront à parler de sauvegarde et non d'archivage en tant que tel. En effet, l'archivage s'inscrit dans une démarche beaucoup plus large qui va de la sélection des données à la mise en place d'un plan de pérennisation. Ces points sont abordés dans la partie 6.

Le terme d'export est souvent employé pour remplacer celui de sauvegarde. Il souligne bien le fait que l'on va extraire d'un système actif (le SGBD), en cours de fonctionnement, des données dans un format particulier (XML, SQL, CSV ou autre) pour conserver une image de toutes les données du système à un instant T. Il nous semble important de discerner deux types d'exports, l'export partiel ou total.

L'export total est souvent réalisé à l'aide de fonctions du SGBD (fonction DUMP). Il permet d'avoir une image de la base de données en conservant les données et les relations entre les tables.

Il n'est pas toujours utile de conserver la totalité de la base. Parfois même, pour des raisons de confidentialité, il est fortement déconseillé de le faire. Il s'agit alors, en ayant connaissance du modèle conceptuel de la base et du contenu, de faire une sélection des informations éligibles à la conservation (cf. partie 6.4). On pourra alors produire un ou plusieurs fichiers dans un format cible SQL (via éventuellement la fonction DUMP, CSV, XML, etc.). C'est ce que nous appelons l'export partiel.

Export total	Export partiel
Conservation du modèle relationnel	Destruction du modèle relationnel ce qui peut aboutir à une perte de cohérences des données.
Simplicité via la fonction DUMP des SGBD	Réalisable via la fonction DUMP que l'on paramètre mais plus complexe dans la mesure où il faut connaître le modèle et les données que l'on souhaite conserver.
Obligation de conserver toutes les données même les données sensibles ou inutiles.	Possibilité de sélectionner les données pertinentes pour l'archivage (épuration de la base)

Caractéristiques de l'export total et partiel

Au-delà de cette notion d'export total ou partiel, il est important de bien choisir la méthode et le format d'export. Nous vous présentons ici les quelques pistes qui nous semblent les plus opportunes pour réaliser cette fonction :

Le format XML

Fortement utilisé pour les échanges de données, le format XML permet de définir des structures complexes de données. Cela peut par exemple permettre d'indiquer des types de données, des références de tables, etc.

A noter : l'initiative des Archives fédérales suisses pour l'archivage de bases de données relationnelles avec la **solution SIARD** (Software Independent Archiving of Relational Databases). Cette solution permet de décrire à la fois les fonctions de gestion spécifiques à un SGBD ainsi que les données qu'il contient, grâce à un format libre spécialement conçu pour l'archivage de bases de données. A chaque base correspond un fichier SIARD, au format ZIP, contenant des fichiers XML dont un fichier qui décrit toutes les métadonnées de la base de données.
Cette solution propose un paquet freeware avec le logiciel "SIARD Suite" qui permet de convertir une base de données relationnelle au format SIARD. SIARD Suite se base sur les standards XML, SQL:1999 et UNICODE et permet actuellement la prise en charge des formats des SGBD Oracle,

Microsoft SQL Server, MySQL et Microsoft Access. Les Archives fédérales suisses mettent gratuitement SIARD Suite à la disposition de tous[3].

Avantages	Inconvénients
Le XML permet de décrire des entités complexes. Aussi, il est possible de créer ses propres fichiers XML avec des caractéristiques répondant à des problématiques particulières.	La création de fichiers XML à partir des données nécessite des compétences informatiques.
Il existe une solution SIARD pour certains SGBD	La solution SIARD est restreinte à quelques SGBD

La fonctionnalité DUMP des SGBD

La plupart des SGBD proposent une fonction de « déchargement » (dump) qui permet de transférer toutes ou partie des données contenues dans un fichier sous forme de requêtes SQL. Chaque SGBD possède des caractéristiques qui lui sont propres et qui vont impacter la manière dont les informations vont être écrites dans le fichier d'export. Aussi un fichier d'export provenant d'un SGBD ne sera pas forcément compatible avec un autre SGBD.

Avantages	Inconvénients
Possible dans tous les SGBD	Incompatibilité possible entre les SGBD. Il sera nécessaire de réinjecter les données dans le même SGBD ou un SGBD compatible.
Export simple à faire	
La plupart des fonctionnalités du SGBD peuvent être conservées.	
Possibilité de sélectionner les données	

Le format CSV

Le Comma Separated Values représente des données tabulaires sous forme de valeurs séparées par des virgules. La simplicité de ce format est telle qu'il n'a jamais vraiment fait l'objet d'une spécification formelle. Dans un SGBD, les données étant incluses dans des tables, il est alors possible de les migrer d'une table vers un fichier CSV. Cependant, il ne sera pas possible de conserver les types des données (une date, un entier, un booléen) ni les relations entre les tables (clés primaires et étrangères). Il sera donc opportun d'apporter des informations de représentation sous forme de métadonnées pour expliquer ces éléments manquants.

Avantages	Inconvénients
Format CSV simple et facilement utilisable pour ingérer les données dans un SGBD	Les types des données des tables à migrer doivent pouvoir être représentées sous forme de texte. A défaut, ils seront perdus. Il faudra alors créer des tables et retrouver des types compatibles avec chaque jeu de données pour pouvoir exploiter de manière convenable les données exportées.
	Les liens entre les tables seront perdus. Il faudra alors reconstituer les liens de clés primaires et étrangères pour permettre d'avoir un modèle relationnel cohérent et représentatif de l'ancienne base.

[3] Pour plus d'informations visitez le site : http://www.bar.admin.ch/

Le gel du SGBD

Un SGBD est composé à la fois de fichiers dédiés à l'exécution de ses processus internes et de fichiers permettant de contenir les données. Ces derniers sont alors en général uniquement exploitables par le SGBD et sont relativement localisables. Le gel du SGBD consiste donc à arrêter les processus en cours et à faire simplement la copie de l'ensemble des fichiers contenant les données.
Lorsqu'on le souhaitera, on pourra alors ré-exploiter les données, réinstaller l'application initiale et mettre les fichiers de données à l'emplacement dédié. Ainsi les données seront prises en charge lors de cette nouvelle installation du SGBD.

Avantages	Inconvénients
Simple à faire	Il faut réinstaller le même SGBD pour pouvoir réutiliser les données
Ne requiert que peu de ressources humaines	

L'entrepôt de données

Comme nous l'avons vu dans la partie 4.1, un entrepôt de données est une sorte de base de données de bases de données, dans la mesure où les données de multiples bases vont venir s'agglomérer dans ce système global.
Les données alimentant un entrepôt sont hétérogènes, issues de différents SGBD ou fichiers plats (Excel, CSV). Pour les intégrer dans ce système global, il faut alors les homogénéiser et leur donner un sens unique compréhensible par des utilisateurs éloignés de la donnée initiale. Pour cela, lors de leur intégration, un processus de normalisation et de rationalisation doit se faire pour améliorer la qualité de la donnée et donc sa capacité à être exploitée facilement.
Du fait de sa grande volumétrie, il doit y avoir une parfaite maîtrise de la sémantique et des règles de gestion des données.

Avantages	Inconvénients
La donnée est directement exploitable	Décrire l'entrepôt de données
Les données sont normées	
Possibilité de conserver les contraintes principales du SGBD (clés primaires, étrangères).	

Ces éléments nous montrent que l'entrepôt de données est une solution intéressante pour l'archivage des bases de données. L'intégration des données pourrait être assimilée à la phase de versement d'un OAIS[4] dans laquelle la définition des règles de gestion relève de la compétence de l'archiviste, à même de connaître les besoins des futurs utilisateurs, les données à intégrer, le format qu'elles doivent avoir et le modèle de description à appliquer pour que la compréhension future de l'information soit possible.

[4] ISO 14721 - Open Archival Information System (OAIS).

6. Les étapes de l'archivage d'une base de données

6.1. Pourquoi archiver une base de données ?

Avant d'envisager une étude détaillée au sein d'une organisation, une des premières questions à se poser est : pourquoi archiver une base de données ?

Il est crucial de bien identifier en amont de toute action les raisons qui poussent à l'archivage. D'une part parce qu'elles auront certainement un impact sur la manière d'archiver la base de données, et d'autre part parce que la complexité et le nombre des bases de données peuvent représenter un frein. Dans ce cas, les motivations qui conduisent à la mise en place d'un tel projet seront autant d'arguments pour convaincre les réticences que ce soit aussi bien vis-à-vis de la hiérarchie que du service informatique.

De manière générale, l'intérêt de la conservation d'un document ou d'une donnée fait l'objet d'une évaluation en fonction de son utilité administrative, juridique et de son intérêt historique, scientifique ou patrimonial (cf. partie Concepts archivistiques applicables).

Ce travail est effectué par l'archiviste, ou à défaut la personne qui endosse cette fonction, en collaboration étroite avec le producteur du document. Ceci dans le respect des prescriptions juridiques notamment dans le cas des documents à valeur probante. Nous reviendrons plus en détail sur cette évaluation dans la partie 6.4 « Évaluer une base de données ».

Selon les situations, les raisons qui motivent l'archivage peuvent varier :

- Premièrement, il faut se poser la question du statut de cette base de données ; les obligations pesant sur les archives publiques n'étant pas les mêmes que pour les archives privées.
En effet, tout organisme public ou exerçant une mission de service public produit dans le cadre de son activité des documents ou des données qui sont des archives publiques (cf. Code du patrimoine, livre II). La conservation de telles archives est encadrée par des procédures bien définies qui rappelons le sont passibles de peines d'emprisonnement et d'amendes si elles ne sont pas respectées. Ainsi, toute organisation a l'obligation de conserver les documents et donc les données qu'elle produit et ne peut les détruire qu'avec l'accord du service des archives responsable du contrôle scientifique et technique.

Les bases de données produites par les services publics contiennent donc des données publiques[5] et à ce titre, doivent être prises en compte dans la politique d'archivage.
Ces obligations ne s'appliquent pas dans le domaine privé ; chaque entreprise étant responsable de ses archives, dès lors qu'elle n'assure pas une mission de service public.

- Deuxièmement, une base de données peut être menacée de disparition lorsque le logiciel qui interagit avec elle n'est plus exploité ou exploitable. Cela peut se produire s'il n'est plus maintenu par le fournisseur, que la licence n'est plus valide, que le logiciel n'est plus utilisé ou que les compétences pour l'utiliser sont devenues suffisamment rares pour s'alarmer. Il faut alors se poser la question de l'archivage de cette base de données si l'on ne veut pas en perdre le contenu.

- Enfin, la volonté d'archiver certaines bases de données peut relever de la politique générale d'un organisme en matière de conservation, souvent en raison de la forte valeur probante des documents ou données qu'elles contiennent. Dans certaines entreprises, il peut également y avoir une volonté de conservation à des fins patrimoniales, pour constituer la mémoire de l'entreprise.

[5] A noter : il ne faut pas confondre la notion d'archives ou données publiques avec le fait que les données soient librement accessibles. Il est fréquent que des données publiques soient confidentielles au tire du respect de la vie privée notamment.

6.2. Faire un état des lieux de l'existant

Cette étape d'état des lieux suppose que dans un premier temps l'archivage des bases de données est envisagé d'un point de vue global au sein d'un organisme. Il est en effet préférable de mener une réflexion générale, en prenant en compte l'ensemble des bases de données utilisées par une structure plutôt que de se focaliser, souvent dans l'urgence et/ou au cas par cas, sur une seule base de données, parce qu'elle risque de disparaître par exemple. Cette façon de procéder aura l'avantage de permettre d'identifier plus facilement les redondances d'informations.

6.2.1. Cartographie du système d'information : inventaire des bases de données

Dans la mesure où l'on souhaite avoir une vue d'ensemble, il est nécessaire de recueillir des informations minimales permettant d'identifier toutes les bases de données utilisées par un organisme. Ces renseignements seront fournis à la fois par le service informatique chargé du système d'information et par les utilisateurs. Il est important d'identifier les interlocuteurs pertinents.

Le tableau ci-dessous donne un exemple du type d'informations à consigner :

Caractéristiques de la base	Description
Nom courant	Nom usuel utilisé dans l'entreprise pour identifier la base
Application utilisant la base	Liste de tous les logiciels ou programmes informatique en interaction avec la base de données. Il peut s'agir par exemple d'un site web utilisant la base ou d'une application spécifique à l'établissement.
Liste des utilisateurs	Liste des utilisateurs ou des groupes d'utilisateurs de la base et des droits associés. Pour chacun on notera l'application utilisée pour interagir avec la base.
Système informatique	Identification et localisation des SGBD. Information sur les paramétrages particuliers, les sauvegardes etc.
Date de mise en service	Depuis quand utilise-t-on la base ?
Date de fin d'utilisation	Le SGBD est-il en fin de vie ou a-t-il une fin d'utilisation programmée ?
Taille de la base	Nombre d'enregistrements ou volume total en Mo/Go/To ?
Objectifs et fonctionnalités	A quoi sert la base ? Que permet-elle de faire ?
Type de contenu	Que contient la base : des données, des documents ? Y a-t-il des données confidentielles ? Ces données sont-elles également disponibles sur papier ? Y a t-il eu des récupérations de données d'anciennes bases ? Y a-t-il des contenus à valeur probante ?
Cycle de vie global des données (DUA)	Fréquence des mises à jour ? Y a-t-il des contraintes juridiques qui nécessitent de conserver les données un certain temps ?
Questions diverses	
Voyez-vous un intérêt à l'archivage de tout ou partie de la base ?	
Qu'est-ce qui motive l'archivage de cette base de données ?	

6.2.2. Inventaire des moyens mis à disposition

Il faut également recenser les moyens dont on dispose au sein de l'organisme pour réaliser cet archivage. En effet, les actions que l'on pourra mettre en œuvre ne seront pas les mêmes selon les ressources mises à disposition.

Le tableau ci-dessous présente les questions à se poser, réparties par types de moyens :

Moyens	Description
Financier	A-t-on un budget dédié pour cet archivage ? De combien ? Pourra-t-on acheter du matériel spécifique ? Recruter du personnel qualifié ? Faire appel à un hébergeur, à un tiers archiveur ?
Humain	Dispose-t-on de personnel consacré à ces tâches, ou de temps mis à disposition ? Combien ?
Compétence informatique	A-t-on des connaissances en bases de données ? en XML ? en développement ?
Compétence archivistique	A-t-on des connaissances sur les pratiques de tri et de traitement des archives ? en documentation ? juridiques ?
Matériel	Possède-t-on du matériel informatique disponible pour archiver ces bases de données ? Si oui, quelles en sont les caractéristiques ?
Établissement potentiellement partenaire (Tiers archiveur)	Des partenariats avec d'autres structures sont-ils envisageables / possibles ? Quelles perspectives de mutualisation ?

L'inventaire des moyens passe également par la définition des rôles de chacun des intervenants afin de bien en délimiter le périmètre d'intervention. Trois compétences sont essentielles : celle de l'archiviste, de l'informaticien et du producteur. Elles doivent pouvoir communiquer afin d'éviter toutes incompréhension ou ambigüité qui pourraient impliquer des erreurs souvent irréversibles.

✓ Rôle de l'archiviste : conseil pour la détermination du cycle de vie afin de respecter les contraintes juridiques (ex. : il faut conserver tel document pendant x années en cas de recours juridique) et sélection des données à archiver définitivement. Responsable, en collaboration avec l'informaticien, de la manière dont on va conserver les informations et surtout de la documentation (métadonnées) qu'il va falloir y associer pour en garantir l'intelligibilité. Il doit transposer les pratiques archivistiques classiques dans le domaine de l'archivage des bases de données.

✓ Rôle de l'informaticien : expertise sur l'architecture et la gestion des bases de données, sur la manière de faire des exports et des requêtes, sur le matériel nécessaire et les mesures techniques à mettre en œuvre pour répondre aux exigences de conservation.

✓ Rôle du producteur et/ou utilisateur de la donnée : expertise sur le contenu informationnel de la donnée (de quoi s'agit-il ? Y-a-t-il des obligations légales de conservation des données ?). Il donne également des indications sur l'usage qu'il fait du contenu de la base : qu'est-ce qu'il utilise ? Sous quelle forme (données brutes / résultats de requêtes) ? Pendant combien de temps (DUA) ?
→ C'est lui qui a la pratique de la base et la connaissance métier, il est donc crucial de l'associer dès le début à la démarche d'archivage.

Il est important de préciser que les étapes de mise en œuvre de l'archivage d'une base de données, qu'elles soient pilotées par un archiviste ou un informaticien, sont réalisées dans le cadre d'un groupe de travail avec une validation commune.

6.3. Sélectionner les bases de données à archiver

L'étape suivante consiste à analyser l'état des lieux réalisé afin de sélectionner les bases de données nécessitant un archivage ou du moins d'affecter des priorités pour l'archivage. Ce travail se fait en fonction des caractéristiques de la base, du cycle de vie des données qu'elle contient et comme nous l'avons vu, des moyens dont on dispose pour effectuer l'archivage. Cela peut prendre la forme d'une liste de critères pertinents préalablement identifiés, dont voici un exemple :

- ✓ Le SGBD est en fin de vie. Cela signifie qu'il va être de moins en moins utilisé par le personnel et/ou qu'il ne dispose plus d'une maintenance informatique, d'où des risques multiples de perte de données, de perte de connaissance de son fonctionnement par le personnel, ou d'inaccessibilité à la base.

- ✓ Le volume des données présentes dans la base et qui ne sont plus utilisées est relativement important et une purge est à envisager.
- ✓ Les informations contenues dans la base sont uniques. Elles ne se retrouvent pas sous la forme de fichiers informatiques ou sur papier.
- ✓ Les données contenues dans la base ont une valeur légale / probante qui nécessite la mise en place de procédures d'intégrité et d'authenticité, ainsi qu'une conservation sur du moyen ou long terme (plus de 5 ans).
- ✓ Les données et/ou la base de données dans son ensemble ont un intérêt historique à être conservées pour témoigner d'une activité, en raison de leur nature, et/ou de leur contenu informatif.
- ✓ Les données et/ou la base de données dans son ensemble ont un intérêt scientifique à être conservées pour effectuer des nouveaux traitements, des travaux de recherche, etc.

Il est à noter qu'une partie de ces critères renvoie à la question du choix du moment de l'archivage : quand décider d'archiver une base de données ?
Outre la disponibilité des moyens nécessaires, cela va principalement dépendre des données qu'elle contient, de leur état dans leur cycle de vie ou de leur valeur probante. Mettre en œuvre un archivage trop tôt, alors que les données sont encore vivantes et donc sujettes à modifications, rend la tâche beaucoup plus complexe puisqu'il faut gérer les mises à jour. En revanche, face à des données probantes, il faut être en mesure de sécuriser les informations dès leur validation pour garder cette valeur de preuve. Cela peut se faire par l'intégration de procédures spécifiques au sein de la base de données. Selon les cas, il peut être préférable d'attendre que les données soient figées ou que le SGBD arrive en fin de vie ; ce qui n'empêche pas en attendant de sauvegarder régulièrement la base et de commencer à la documenter pour en préparer l'archivage.

Il faut ensuite analyser les résultats en regard des contraintes que l'on a, en termes de délais et de complexité de la base de données, et prioritiser les actions à entreprendre.
Lorsque le logiciel qui permet d'exploiter une base de données arrive en fin de vie et que le serveur qui l'hébergeait est dans le même cas ou doit être récupéré pour un autre usage, si les données ont un intérêt légal ou historique à être conservées, il est nécessaire d'intervenir rapidement.

Prenons l'exemple d'une base de données de gestion du personnel dont le logiciel arrive en fin de vie et pour laquelle seules les données encore actives (dossiers des personnels encore en poste) ont été migrées vers le nouveau SGBD. Elle doit être traitée en priorité au vu des délais de conservation légaux pour ce type de données (90 ans à partir de la naissance de l'employé), d'autant plus que souvent le sort de l'ancienne base de données n'est pas très bien défini (déplacement vers un serveur inutilisé, diminution de la fréquence d'utilisation jusqu'à la perte de la connaissance de son fonctionnement, ou même la suppression complète).
De la même manière, il peut être utile d'évaluer l'intérêt historique de l'archivage au regard de la complexité de la base de données, selon la logique de « la fin justifie-t-elle les moyens ? ». Au vu des structures actuelles de SGBD de plus en plus complexes, il est possible que la mise en œuvre de la conservation ne soit pas en adéquation avec les moyens identifiés dans la partie 6.2.2 à tel point qu'elle devienne inenvisageable malgré l'intérêt historique des données.

Les résultats de cette analyse doivent être reportés dans un document de synthèse basé sur l'annexe 1 qui présente la stratégie retenue sur la base de l'état des lieux de l'existant et liste la ou les bases de données à archiver selon un ordre de priorité. Ce rapport doit être soumis à validation hiérarchique avant de passer à l'étape suivante.

6.4. Évaluer une base de données

Lorsque l'on a identifié une base de données à archiver, il est important d'évaluer précisément son contenu afin de sélectionner ce que l'on va archiver (toutes les données ne sont pas forcément candidates à un archivage) et la solution la plus adaptée pour le faire. Les informations recueillies lors de l'état des lieux de l'existant (cf. partie 6.2.1) seront le point de départ de cette évaluation, complétées si nécessaire.

A noter : La démarche proposée dans la partie 3 d'ICA Req (Principes et exigences fonctionnelles pour l'archivage dans un environnement électronique – Module 3 : Recommandations et exigences

fonctionnelles pour l'archivage des documents dans les applications métier) est une aide possible pour l'audit d'une application.

6.4.1. Identifier le cycle de vie des données de la base

Comme cela a pu être expliqué précédemment, une base de données est composée d'éléments très hétérogènes et contient généralement des données courantes et intermédiaires, voire même définitives.
L'étape de l'état des lieux de l'existant a permis de dresser un aperçu du cycle de vie global des données de la base. Il convient maintenant de reprendre cette analyse plus en détail :

- Toutes les données suivent-elles le même cycle de vie ? ont-elles la même DUA ?
- Certaines sont-elles figées dès leur création ? ou au contraire constamment modifiées ?
- La base fonctionne-t-elle de manière cumulative ou au contraire dynamique ?

Par exemple, une donnée peut être présente dans une base dès le début de sa création et rester figée par la suite. A l'inverse, elle peut aussi être insérée dans la base pour une période relativement courte, de l'ordre de la seconde ou la milliseconde.

La volatilité des données doit également être mesurée en fonction du risque de non-disponibilité de l'information, ce qui passe par l'évaluation de la valeur probante des données.
Il est important d'identifier qu'elles sont les exigences en matière de traçabilité et de preuve afin d'identifier ensuite quels sont les éléments pertinents à archiver et comment procéder.

6.4.2. Évaluer la confidentialité des données

Une base de données contient généralement des informations confidentielles, soit parce qu'elles relèvent de la vie privée (identification de personnes), soit parce qu'il s'agit d'identifiants et de mots de passe utilisés pour l'accès notamment.
L'identification de ces différents types d'informations va permettre de prendre les mesures nécessaires :

- Dans le cas d'informations personnelles, d'avertir la CNIL des démarches entreprises si les données ont vocation à être archivées.
- Dans le cas d'identifiants et de mots de passe, de paramétrer les exports pour ne pas les rendre visibles lors de l'archivage.

6.4.3. Sélectionner les données à archiver

Sur la base de l'analyse menée, il faut alors sélectionner ce que l'organisme a réellement besoin d'archiver (quelles sont les données qui ont besoin d'être préservées ?) et préciser le moment et la fréquence de l'archivage de ces données.
Il faudra ensuite identifier toutes les données et leur localisation : données brutes, issues de requêtes, représentées dans des formulaires, etc.
Au-delà des données, on peut se poser la question de l'intérêt de l'archivage de la couche métier et de la couche présentation (cf. partie 4.5). Cette réflexion est étroitement liée à l'utilisation future des données.

6.5. Choisir sa stratégie d'archivage

Le choix du processus que l'on va appliquer pour archiver les données d'une ou plusieurs bases va donc dépendre de nombreux critères aussi bien techniques qu'au niveau des ressources humaines ou financières. Il s'avère difficile de dresser un tableau simple des critères et des stratégies d'archivage à appliquer car le niveau de complexité est tel qu'une expertise au cas par cas semble nécessaire.
Nous allons cependant essayer de parcourir quelques critères qui nous paraissent significatifs sous la forme d'un tableau synthétique. Les solutions d'archivage abordées ci-dessous sont expliquées dans la partie 5. Elles ne sont pas forcément exclusives et peuvent s'envisager notamment dans le cadre de partenariats entre institutions.

Critères	Action à privilégier
Peu de moyens : compétences archivistiques et/ou informatiques minimales.	Le gel de la base ou un dump de la totalité de la base peut être un moyen a minima. Il faut alors s'assurer que le SGBD utilisé pourra être réactivé lorsqu'on voudra ré-exploiter la base.
Les compétences archivistiques et informatiques sont présentes mais peu de budget.	L'archiviste peut faire un inventaire des données à garder et proposer un modèle logique des données à exporter. L'informaticien effectuera ensuite les requêtes nécessaires pour les exporter vers un format de type ods ou CSV par exemple. L'informaticien et l'archiviste prendront soin d'accompagner l'export réalisé d'une documentation sur les modèles conceptuel, logique et physique des données.
Les compétences archivistiques et informatiques sont présentes, avec un budget et une infrastructure importante.	On mettra en place un système informatique de type entrepôt de données. L'archiviste fera une analyse des données pertinentes à archiver et enclenchera avec l'informaticien un processus de versement des données dans l'entrepôt. On pourra se focaliser sur la description conceptuelle des données dans la mesure où les dimensions logique et physique seront supportées par l'entrepôt et seront d'autant plus simples que les données auront été réorganisées et normalisées.
On souhaite archiver une base de données peu complexe, avec peu de moyens.	On pourra faire un export au format CSV par exemple dans lequel chaque fichier CSV représentera une table. On accompagnera ces fichiers d'une description du modèle physique (MPD) pour mentionner les formats des données pour chaque colonne (dates, entier, clé primaire ou étrangère). Il faudra en outre décrire les modèles logique et conceptuel. Cette description pourra être restreinte compte tenu de la faible complexité de la base initiale.
Il existe des données que l'on ne souhaite pas archiver (confidentielles, inutiles etc..).	Obligation de faire un export ciblé.
La base de données est uniquement utilisée par une application et elle est trop complexe pour être directement comprise et exploitable par les personnes concernées.	Archivage des données sous forme d'états générés par l'application qui l'exploite (exemple : page PDF d'un contrat client au lieu des données brutes du client).
On souhaite pouvoir réutiliser la base de données avec l'application.	Il est alors nécessaire de conserver la totalité de la base de données, soit à l'aide de la fonction dump du SGBD (option à privilégier), soit en sauvegardant directement les fichiers du SGBD contenant les données. Il faudra conserver les sources du SGBD et de l'application et garantir le fait qu'une installation sur un autre système d'exploitation est possible tout au long de la durée de l'archivage.

Quelle que soit la stratégie d'archivage retenue, il faudra ensuite mettre en œuvre les moyens nécessaires pour assurer une conservation numérique. Cela peut notamment prendre la forme d'un système d'archivage électronique.

7. Annexe 1 : Grille d'évaluation d'une base de données

Cette grille est une aide à l'évaluation d'une base de données pour décider de son archivage en fonction de plusieurs critères préalablement identifiés. Un coefficient est attribué pour chaque critère. Il indique l'importance ou la pertinence du critère pour l'organisme. Une note est ensuite donnée en fonction du niveau de satisfaction de ce critère par la base de données. Les résultats obtenus (produit du coefficient et de la note) vont permettre d'affecter un indice de priorité à chaque base de données afin de hiérarchiser les priorités d'archivage.
Les critères et les coefficients renseignés ci-dessous sont là à titre d'illustration et doivent être adaptés en fonction du contexte de l'organisme.

Critères	Coef. (de 1 à 5)	Note (0 = non rempli 1 = à moitié rempli 2 = rempli)	Résultat	Remarques
Données figées (ne feront plus l'objet de mise à jour)	3			
Base de données consultée	1			
Fin de maintenance du SGBD	3			
Présence d'informations uniques	5			
Contraintes légales de conservation sur du moyen ou long terme	5			
Données à valeur probante	5			
Données ayant un intérêt historique	3			
Données ayant un intérêt scientifique	2			
La base de données est relativement simple : peu de tables, peu de traitements, peu de vues	2			
Données récupérées d'anciennes bases	1			