École nationale des chartes | PSL

Clémence Beau

Étude actualisée sur l'archivage des bases de données. Expérimentations réalisées à partir d'une base de données sur les sportifs de haut niveau.

Mémoire pour le diplôme de Master
« Technologies numériques appliquées à l'histoire »
2023

Résumé

Ce mémoire résulte d'un stage de quatre mois effectué au sein du bureau de l'expertise numérique et de la conservation durable au service interministériel des Archives de France sur l'archivage des bases de données. Présentes à tous les niveaux de la vie quotidienne et administrative et renfermant des données à forte valeur patrimoniale, leur conservation est une problématique actuelle et récurrente au sein du milieu archivistique. Tandis que la France, héritière de la méthodologie de Constance recourt encore majoritairement aujourd'hui à l'archivage à plat pour conserver ses bases de données, d'autres initiatives suscitent l'intérêt de la communauté archivistique internationale. Parmi elles, le format SIARD (Software Independent Archival of Relational Databases), développé par les Archives fédérales suisses en 2004 et préconisé par plus d'une cinquantaine de pays ou encore l'émulation, utilisée pour la conservation des jeux vidéos et transposable aux données structurées.

À partir d'une documentation actualisée, d'expérimentations techniques réalisées à partir d'un cas d'étude concret - une base de données sur les sportifs de haut niveau - et d'échanges avec divers professionnels, cette étude a pour but de proposer aux services d'archives un nouvel état des lieux des solutions pour la conservation pérenne des bases de données.

Abstract :

This dissertation is the result of a four-month internship at the bureau de l'expertise numérique et de la conservation durable at the service interministériel des Archives de France about the archiving of databases. Present at all levels of daily and administrative life and containing data with high heritage value, their conservation is a current issue within the archival community. While France, heir to Constance's methodology, still mainly uses flat archiving today to preserve its databases, other initiatives are arousing the interest of the international archival community. Among them, the SIARD (Software Independent Archival of Relational Databases) format, developed by the Swiss Federal Archives in 2004 and recommended by more than fifty countries, or even emulation, used for the conservation of video games and transposable to structured data.

Based on updated documentation, technical experiments carried out from a concrete case study - a database on high-level athletes - and exchanges with various professionals, this study aims to propose for archive services a new inventory of solutions for the long-term conservation of databases.

Mots clés :

Bases de données - Bases de données relationnelles - Système de gestion de bases de données - SQL - XML - Documents nativement numériques - Archivage - Archivage à plat - Software Independent Archival of Relational Databases (SIARD) - service interministériel des Archives de France - Archives fédérales suisses - Archives nationales - VITAM - Émulation - Sportifs de haut niveau - Rétro-ingénierie - Expérimentations - Historien - Archiviste - Usages - Accessibilité.

Informations bibliographiques :

BEAU Clémence, « Étude actualisée sur l'archivage des bases de données. Expérimentations réalisées à partir d'une base de données sur les sportifs de haut niveau. », mémoire de master 2 *Technologies numériques appliquées à l'histoire*, dir. Emmanuelle Bermès, septembre 2023.

Remerciements

Ce mémoire, résultat d'un stage stimulant de quatre mois au service interministériel des Archives de France marque également la fin de six années d'études passionnantes. Synthèse d'une réflexion alimentée par diverses personnes, sa réalisation leur est en partie due et il convient donc de les remercier à juste titre.

Je tiens tout d'abord à remercier Emmanuelle Bermès pour son accompagnement, sa disponibilité, sa bienveillance et ses précieux conseils tout au long de l'année et plus particulièrement durant toute la durée de ce stage et la rédaction du mémoire.

Je remercie très chaleureusement Violette Lévy et Dominique Naud pour leur accueil enthousiaste au sein du bureau de l'expertise numérique et de la conservation durable du SIAF, leur disponibilité, le partage de leurs connaissances et de leur expérience et leurs encouragements. Ce stage fut, grâce à elles, une première immersion dans le monde professionnel à la fois très enrichissante et passionnante.

Je souhaiterais par ailleurs exprimer ma reconnaissance à Louis Vignaud, expert chargé de la politique nationale sur les métadonnées et référentiels archivistiques du SIAF pour le temps qu'il m'a accordé. Nos échanges hebdomadaires sur diverses notions informatiques et son appétence technique m'ont beaucoup aidée et rassurée. Je le remercie pour cela.

Je voudrais dire merci à Anne Lambert et Chloé Moser, respectivement cheffe et adjointe du bureau des archives des ministères sociaux pour leur accueil sympathique et leur dévouement.

Je remercie très sincèrement Edouard Vasseur, qui, malgré son emploi du temps chargé, a pris le temps de répondre à mes interrogations, de me conseiller et de me partager sa précieuse documentation.

J'aimerais aussi exprimer ma gratitude aux chercheurs historiens et sociologues, archivistes français et suisses et autres spécialistes du sujet qui ont pris le temps de discuter avec moi. Chacun de ces échanges stimulants a donné une impulsion à ma réflexion et a ajouté une véritable dimension de recherche à ce sujet pourtant originellement technique.

Je tiens à adresser un merci plus large à toute l'équipe du service interministériel des Archives de France pour son accueil, sa gentillesse et sa considération. C'est avec plaisir que je la réintègrerai pour quatre mois à la rentrée.

Enfin, je remercie mes parents pour leur confiance et leur soutien durant tout mon parcours universitaire ainsi que Camille et Romain pour leur écoute et leurs encouragements qui furent d'une grande aide lors de la rédaction de ce mémoire.

Introduction

« La fragilité du support numérique est antinomique de l'objet même de l'archivage patrimonial qui est la pérennisation de l'information[1]. » évoque Céline Guyon, présidente de l'Association des archivistes français dans l'article intitulé « La pratique archivistique publique en France, entre adaptation et négociations. Expériences et réflexions d'une archiviste. » paru en 2015 dans *Les cahiers du numérique*. En effet, le numérique tend à rompre la stabilité du document d'archives puisque contrairement au papier, le support n'est plus en mesure de garantir la fixité des informations. Le concept de fixité, au coeur de la pratique archivistique est défini comme étant « la propriété de ce qui se maintient ou est maintenu dans un état donné[2] ». Or, avec le numérique, plusieurs risques sont aujourd'hui identifiés parmi lesquels nous pouvons mentionner l'obsolescence du matériel informatique et des formats de fichiers, la disparition des logiciels de lecture ou encore la perte de la signification du contenu informationnel[3]. Pour contrer ces risques, plusieurs bonnes pratiques sont mises en place telles que des copies multiples sur différents types de supports, une veille sur les technologies existantes mais également l'utilisation de métadonnées pour documenter les informations archivées[4].

Si les problématiques énoncées ci-dessus concernent tous les documents nativement numériques, à savoir les documents directement créés sous forme électronique[5], il convient dans un souci de définition de notre sujet, d'opérer une distinction au sein de cet ensemble entre documents bureautiques et bases de données. On entend par document bureautique « un document produit par un logiciel de traitement de texte ou un tableur[6] » tandis qu'une base de données est considérée comme étant une « collection de données unitaires, d'objets de données et de liens entre eux, structurée de manière à permettre à un certain nombre d'applications différentes d'y accéder[7] ». Bien que tous prennent place au sein de cet ensemble désigné sous les termes de « documents nativement numériques », des enjeux propres et distincts leur sont affiliés, directement corrélés à la manière dont ils structurent l'information. Tandis que des travaux tels que celui de

---

[1] GUYON Céline, « La pratique archivistique publique en France, entre adaptation et négociation. Expériences et réflexions d'une archiviste. », *Les cahiers du numérique*, n°11, 2015 (p. 77 à 114). Disponible à l'adresse suivante : https://www.cairn.info/revue-les-cahiers-du-numerique-2015-2-page-77.htm, consulté le 28 août 2023.

[2] GUYON Céline, « L'archivage comme dispositif de transformation de la nature intrinsèque des objets nativement numériques », *Balisages*, 2020, https://publications-prairial.fr/balisages/index.php?id=282, consulté le 28 août 2023.

[3] *Guide méthodologique pour l'archivage des bases de données* élaboré par le Centre d'informatique national de l'Enseignement Supérieur (CINES), https://www.cines.fr/wp-content/uploads/2022/05/GM_archivage_BDD-v1.1.pdf (consulté le 8 avril 2023).

[4] *Loc. cit.*

[5] Par opposition aux documents numérisés qui sont issus de la numérisation de documents papier.

[6] POIVRE Joël, *L'archivage des documents bureautiques. Manuel pratique*, Direction des Archives de France, Paris, 2004.

[7] Définition issue de la source suivante : BCS , 2013, BCS = BCS Academy Glossary Working Party (2013), BCS Glossary of Computing and ICT 13th edition. Disponible à l'adresse suivante : https://learning.oreilly.com/library/view/bcs-glossary-of/9781780171500/11_GlossaryofICT_partA9.xhtml, consulté le 29 août 2023.

Joël Poivre[8] ont permis depuis plusieurs années déjà, d'établir une méthodologie relativement précise sur l'archivage des documents bureautiques, la réflexion sur la conservation des bases de données s'inscrit dans une temporalité différente.

En effet, les données dites « structurées » sont les premières données numériques collectées par les Archives Nationales, et ce, dès le début des années quatre-vingt. Se pose alors une question : comment conserver durablement des données originellement produites pour être maintenues dans une structure physique ? En effet, tout l'enjeu de la conservation des bases de données réside dans le fait de pérenniser l'intelligibilité, l'authenticité et l'intégrité de données créées à l'intérieur d'une structure constituée généralement de tables, de colonnes et de lignes. De cette réflexion résulte la création en 1981 du service d'archivage électronique Constance (Conservation et Stockage des Archives Nouvelles Constituées par l'Électronique) reposant sur la méthodologie suivante : les données sont archivées à plat dans un fichier, c'est-à-dire qu'elles sont rendues accessibles indépendamment de la structure informationnelle dans laquelle elles ont été produites. En parallèle, sont stockées d'une part les informations de représentation explicitant la structure des fichiers et de l'autre, la documentation associée précisant le contexte de production et d'extraction des données.
Bien que cette technique ait fait ses preuves et soit encore largement utilisée aujourd'hui en France, d'autres méthodes ont parallèlement émergé depuis. En effet, les années 2000, marquées par l'arrivée d'internet, voient une véritable massification de la production de documents nativement numériques et particulièrement des bases de données. Celles-ci sont désormais considérées comme un véritable objet d'étude en soit et non plus seulement comme de simples composantes des documents bureautiques. Cette accélération entraîne une réflexion profonde des milieux archivistiques français et internationaux et débouche sur la création par les Archives fédérales suisses d'un format exclusivement dédié à l'archivage des bases de données relationnelles : le format SIARD. Le principe est le suivant : le contenu d'une base de données relationnelle est extrait et transposé en SIARD, permettant ainsi d'archiver durablement les données indépendamment de leur logiciel d'origine. Celles-ci pourront ainsi être réimplantées à l'avenir dans des systèmes de gestion de bases de données modernes[9].

Il faut en réalité attendre les années 2010 et l'explosion des usages du numérique pour assister à une véritable accélération des travaux menés sur l'archivage des bases de données. Désormais présentes à tous les niveaux de la vie quotidienne, administrative et institutionnelle, elles renferment une quantité non négligeable d'informations dont l'intérêt patrimonial ne fait alors plus aucun doute et dont les enjeux de conservation sont clairement identifiés.
En France, les travaux affluent : entre rédaction de guides méthodologiques, recensement des méthodes existantes, études menées sur l'émulation et les expérimentations du format SIARD, la communauté archivistique française montre un réel attrait pour le sujet. On constate néanmoins un décalage avec les autres pays européens. En effet, malgré l'utilisation du format SIARD pour l'archivage de la matrice

8 POIVRE Joël, *L'archivage des documents bureautiques. Manuel pratique*, Direction des Archives de France, Paris, 2004.

9 Ces propos sont extraits de la page de présentation du format SIARD sur le site des Archives fédérales suisses, https://www.bar.admin.ch/bar/fr/home/archivage/outils-et-instruments/siard-suite.html, consultée le 23 août 2023.

cadastrale, la méthode préconisée par la France pour la conservation des bases de données reste l'archivage à plat contrairement à la plupart des pays européens qui semblent, dès le début des années 2010 recourir majoritairement au format suisse. Parallèlement, d'autres initiatives internationales voient le jour parmi lesquelles nous pouvons mentionner le format d'archivage CHRONOS ou le logiciel DBPTK.

C'est donc dans ce contexte d'effervescence intellectuelle marquée d'un côté par la floraison de diverses méthodes d'archivage des bases de données et de l'autre par les interrogations suscitées par le format SIARD que s'inscrit notre stage au service interministériel des Archives de France (SIAF). Créé par l'arrêté du 17 novembre 2009, le SIAF, anciennement connu sous le nom de Direction des Archives de France, est l'un des services patrimoniaux du ministère de la Culture. Il exerce sa tutelle sur les trois services à compétence des Archives nationales ainsi que sur les services d'archives des collectivités territoriales. En plus de définir, coordonner et évaluer l'action de l'État en matière de collecte, de tri, de classement, de description, de conservation et de communication des archives publiques, le SIAF exerce également un rôle de conseiller pédagogique auprès des différents services d'archives.
Ce stage, proposé à l'École nationale des chartes dans le cadre du master Technologies numériques appliquées à l'histoire, tend à répondre à un triple objectif. Il s'agit tout d'abord d'actualiser les travaux effectués en recensant les différentes alternatives proposées aujourd'hui pour la conservation des données structurées. Aussi, les projets incluant l'archivage de bases de données se multipliant, il est l'occasion de proposer à ses différents interlocuteurs des solutions d'archivage pérennes en soulignant les points d'attention de chacune. Enfin, le SIAF, de par sa place au sein de la communauté archivistique internationale vise à comprendre pourquoi le SIARD, bien que majoritairement adopté à l'étranger, n'ait été, jusqu'à présent que très peu utilisé en France.

Ainsi, entre documentation, expérimentations techniques et recueil de témoignages de professionnels, il s'agit à partir d'un état de l'art actualisé, de proposer aux services d'archives, des solutions pérennes pour la conservation des bases de données. Afin de donner une dimension concrète à cette étude, nous réfléchirons à partir de l'étude d'une base de données décommissionnée sur les sportifs de haut niveau.
Cette réflexion, s'articulera autour de trois parties. Dans un premier temps, nous réaliserons un état des lieux actualisé sur les pratiques en termes d'archivage de bases de données. Pour cela, il conviendra de redéfinir l'objet « base de données » mais également de comprendre les enjeux liés à leur archivage en rappelant l'historique de leur conservation. L'énonciation des étapes d'archivage nous conduira ensuite à énumérer les techniques existantes tant au niveau national qu'international en termes de préservation des bases de données. Puis, nous nous concentrerons sur notre cas d'étude concret : la base de données sur les sportifs de haut niveau. À travers l'exposition du projet d'archivage, la rétro-documentation avérée nécessaire et les expérimentations techniques réalisées d'une part sur l'archivage à plat, et d'autre part sur le format SIARD, nous établirons, à partir des résultats, des préconisations finales sur la solution d'archivage semblant la plus adaptée à ce cas.
Enfin, une dernière partie, plus réflexive viendra questionner l'objet base de données par le biais des usages et de l'accessibilité. Grâce au recueil de témoignages d'archivistes, de chercheurs en sciences humaines et sociales et d'experts informatique, nous réfléchirons à l'essence même des métiers d'archiviste et de

chercheur et en questionnerons les limites sous plusieurs angles : quand termine le travail de l'archiviste ? Quand commence celui de l'historien ? À partir de quand peut-on considérer qu'un archivage est abouti ?

Ce travail, résultat d'un stage de quatre mois au service interministériel des Archives de France tend finalement à poser les problématiques régissant l'archivage des bases de données en France aujourd'hui. Réalisé sans prétention, il tend à mettre en exergue les points d'attention constatés lors des expérimentations et espère apporter aux archivistes, quelques réponses sinon un nouveau support de réflexion.

PREMIÈRE PARTIE - ÉTAT DE L'ART : L'ARCHIVAGE DES BASES DE DONNÉES D'HIER À AUJOURD'HUI

Chapitre 1 : Besoin de définition et définition des besoins

1. 1 : Définitions liminaires

Chaque jour, lorsque nous faisons nos courses, que nous utilisons notre téléphone ou que nous consultons un site web, sans nous en rendre compte, nous faisons fonctionner de multiples bases de données. Une base de données est une « collection de données unitaires, d'objets de données et de liens entre eux, structurée de manière à permettre à un certain nombre d'applications différentes d'y accéder[10]. » Cette structuration des données[11] permet de représenter lisiblement et ainsi, de pouvoir exploiter efficacement des informations relatives à une activité humaine.

Une base de données est un objet d'étude complexe d'abord parce qu'il induit de préserver à la fois l'information numérique et la structure de la base dans un format spécifique mais aussi parce que les modèles de bases de données sont en constante évolution. Archivistes, bibliothécaires et ingénieurs informaticiens s'accordent tous pour dire qu'il n'y a effectivement pas un seul type de bases de données. On catégorise actuellement les bases de données de plusieurs manières : par usage, par type de contenu ou encore par modèle. Ainsi, afin de refléter au mieux les réalités actuelles, nous exposerons ici deux classifications distinctes. La première, élaborée par le Centre d'informatique national de l'Enseignement Supérieur (CINES) tend à différencier les bases de données en fonction de la nature des relations entre les données. On distingue alors le modèle hiérarchique du modèle dit « objet » et du modèle relationnel[12]. Créé dans les années 1960, le modèle hiérarchique est le plus ancien. Il repose sur une structure arborescente à l'intérieur de laquelle les données sont associées uniquement via des relations de type « parent-enfants ».
Le modèle objet stocke, quant à lui, non pas des données de type simple comme des entiers, des dates ou du texte mais des objets issus de la programmation orientée objet. On entend par objet, un ensemble de données et de fonctions représentant une entité cohérente.
Enfin, le modèle relationnel, inventé en 1970 par le mathématicien Edgar Frank Codd, organise les données sous forme de tables reliées entre elles par des jointures. Ce modèle, créé pour optimiser les traitements et éviter la redondance, est très normalisé et donc largement utilisé encore aujourd'hui.

Une autre catégorisation, présentée par Emmanuelle Bermès dans le cadre de son cours sur le traitement de la donnée dispensé à l'École nationale des chartes, s'appuie sur la distinction entre le modèle en arbre (ou hiérarchique), le modèle en tables (ou relationnel) et le modèle en graphe (ou sémantique). Ici, le classement repose tant sur la nature des relations des données que sur la modélisation de celles-ci. Si les deux premiers modèles correspondent aux définitions données précédemment, le modèle en graphe se

---

10 Définition issue de la source suivante : BCS , 2013, BCS = BCS Academy Glossary Working Party (2013), BCS Glossary of Computing and ICT 13th edition. Disponible à l'adresse suivante : https://learning.oreilly.com/library/view/bcs-glossary-of/9781780171500/11_GlossaryofICT_partA9.xhtml, consulté le 29 août 2023.

11 « Donnée » est un « terme utilisé, en particulier en informatique, pour désigner une information », Association des archivistes français, *Abrégé d'archivistique, Principes et pratiques du métier d'archiviste*, Paris, 2020, p. 330.

12 Définition issue du *Guide méthodologique pour l'archivage des bases de données* élaboré par le Centre d'informatique national de l'Enseignement Supérieur (CINES), disponible à l'adresse suivante : https://www.cines.fr/wp-content/uploads/2022/05/GM_archivage_BDD-v1.1.pdf (consulté le 8 avril 2023).

concentre quant à lui tant sur les relations entre les données que sur les données elles-mêmes. Il permet une visualisation plus évidente des liens entre les données, ce qui facilite leur exploitation. Cette représentation des données est particulièrement utile pour toute personne désireuse d'effectuer des recherches pointues dans les données, qu'il s'agisse d'un chercheur ou d'un particulier.

Une base de données relationnelle est composée d'une ou plusieurs tables (ou classes ou entités) contenant chacune un ensemble de données relatives à un même sujet. Chaque table est constituée d'un ensemble de colonnes (ou attributs) décrivant le contenu des lignes (ou enregistrements) de la table.
La création d'une base de données relationnelle est soumise à diverses contraintes basiques que l'on appelle « déclarations natives ». Il est d'abord nécessaire de typer les données de chaque colonne afin que leur visualisation soit cohérente et exploitable. Pour cela, il suffit d'indiquer si les données inscrites dans chaque colonne seront textuelles ou numériques[13] ainsi que leur caractère obligatoire ou facultatif (*NULL*[14]).
Puis, vient le moment de définir les clés candidates de chaque table, c'est-à-dire tout attribut d'une table dont la valeur unique permet de retrouver un élément dans cette table. L'administrateur choisit ensuite parmi ces clés candidates, l'attribut utilisé comme clé primaire pour chaque table. Celle-ci permettra alors d'établir des relations entre les tables en garantissant que chaque enregistrement n'est stocké qu'une fois et qu'aucune information n'a été perdue. La clé primaire doit être unique et immuable pour une entité donnée. L'attribut concerné ne peut être ni facultatif, ni vide. Enfin, pour connecter les tables entre elles, on définit une clé étrangère dont le rôle est d'indiquer que les données d'une table sont reliées à celles d'une autre table.

Une base de données est indissociable d'un système de gestion de bases de données (SGBD). En effet, lors de sa création, la base est enregistrée sur un support de stockage. Pour stocker, créer, gérer et manipuler les données contenues dans le support, un logiciel système est requis : c'est le système de gestion de bases de données (SGBD). Celui-ci gère tous les aspects primaires de la base tels que la manipulation des données, l'authentification des utilisateurs, la protection des données contre les accidents ou encore l'insertion et l'extraction des données. Microsoft Access, Oracle et MySQL font partie des systèmes de gestion de bases de données les plus utilisés. Si les deux premiers sont des systèmes propriétaires, MySQL est quant à lui, un logiciel libre.

Pour interagir avec une base de données qu'elle soit relationnelle ou non, il est nécessaire de recourir à un langage que le SGBD comprenne. Pour cela, on utilise le langage de requête SQL (Structured Query Language). Créé en 1974 et normalisé depuis 1986, il permet de créer, modifier et interroger la structure d'un système de gestion de bases de données.

1. 2 : Historique

[13] D'autres types sont possibles : nombre, entier, date…

[14] Lorsque *NULL* est affiché dans un champ, cela signifie qu'aucune valeur n'est indiquée.

Une base de données est avant tout un objet conceptuel, le résultat d'une structuration intellectuelle de l'information. Conçues comme le regroupement en un même endroit d'informations structurées de telle manière à faciliter leur lisibilité, on peut tout à fait imaginer que les bases de données aient existé au format papier sous forme d'inventaires ou de listings. Transposée au numérique, cette organisation de l'information est simplifiée et systématisée. On relate ainsi les premières bases de données informatiques dès les années soixante. Afin de saisir au mieux les enjeux actuels posés par l'archivage des bases de données sans pour autant prétendre à une exhaustivité parfaite, il semble nécessaire de revenir à la genèse.

Pour cela, et avant tout propos, il convient de définir de ce que sont les archives. L'article 211-2 du Code du Patrimoine définit les archives comme étant « L'ensemble des documents, y compris les données, quels que soient leur date, leur lieu de conservation, leur forme et leur support, produits ou reçus par toute personne physique ou morale et par tout service ou organisme public ou privé dans l'exercice de leur activité[15]. »

Si l'on dissocie aisément les archives papier de celles numériques, il est néanmoins nécessaire de rappeler qu'au sein même des archives numériques, une distinction entre documents numérisés et documents nativement numériques est requise. En effet, tandis que les premiers sont issus de la numérisation de documents papier, les seconds sont directement créés sous forme numérique[16].

Enfin, parmi les documents nativement numériques, il convient de distinguer les données que l'on considère comme étant non structurées, tels que les fichiers bureautiques ou les courriels, des données dites « structurées » issues de bases de données ou d'outils de gestion électroniques de documents (GED)[17]. Cette dissociation doit être rappelée car la temporalité, les méthodes et les problématiques en découlant sont différentes et strictement liées à leur nature.

L'archivage des données dites « structurées » n'est pas un sujet récent. En effet, dès la fin des années soixante-dix, les Archives Nationales collectent des données issues des grandes bases statistiques produites par l'Institut national de la statistique et des études économiques (INSEE), l'Institut national d'études démographiques (INED) et les ministères de l'Agriculture et de l'Équipement. Ces collectes importantes et régulières poussent l'institution à réfléchir à un moyen d'archiver ces données. En 1981, Constance (Conservation et Stockage des Archives Nouvelles Constituées par l'Électronique) voit le jour[18]. Cette méthodologie pionnière pour l'archivage des données nativement numériques est en premier lieu appliquée à la conservation de données structurées. Elle repose sur deux éléments : d'un côté, l'archivage à plat des données, c'est-à-dire selon des méthodes permettant de les rendre accessibles en dehors des systèmes qui les avaient produites et de l'autre, le stockage des informations de représentation explicitant la structure des fichiers et de la documentation associée, qui précise pour sa part le contexte de production et d'extraction des données. Avec Constance, les fichiers de données sont réduits à des exports à plat selon un principe simple :

---

15 « Archives », article 211-2 du Code du Patrimoine. https://www.legifrance.gouv.fr/codes/id/LEGISCTA000006129161/ (consulté le 17 juillet 2023).

16 MAGNIEN Agnès (dir.), GUICHARD-SPICA Hélène, LAPERDRIX Marie, LOPEZ Magalie, LEBLANC Marie-Noëlle, GALLET-MENAGER Isabelle, *Vade-mecum de l'archivage des documents électroniques*, Janvier 2012.

17 Distinction entre données structurées et données peu ou non structurées. https://archives.paris.fr/a/531/archivage-numerique/ (consulté le 17 juillet 2023).

18 Propos issus d'un échange mené avec Martine Sin Blima-Barru, Mathias Fereira et Emeline Levasseur (département de l'administration des données aux Archives Nationales), le 6 juin 2023.

chaque table correspond à un fichier de données encodé en ASCII et dont le texte n'est pas délimité[19]. Ce sont les informations de représentation et la documentation associée qui permettent de rendre ces données, au premier abord inintelligibles, compréhensibles par l'homme.

Si la mise en place de Constance marque le début de la collecte de données structurées aux Archives nationales, celle-ci reste d'abord circonscrite aux versements émanant des grands services nationaux de statistiques. Cette tendance semble se poursuivre jusqu'au début des années 2000. La généralisation de l'informatique tendant à faire évoluer les pratiques administratives, l'usage des bases de données se systématise. Dès lors, des informations auparavant consignées sur papier sont entrées dans ces conteneurs structurés. Les Archives nationales assistent à une massification des versements ainsi qu'à une diversification de nature des objets à archiver. Alors qu'en 2010, le fonds Constance était constitué à 80% de données structurées, celles-ci, supplantées par les documents bureautiques, ne représentent plus que 20% des versements effectués entre 2010 et 2015[20]. Cette massification conduit la communauté archivistique à recentrer sa réflexion sur l'archivage des documents bureautiques afin d'élaborer une méthodologie de conservation satisfaisante. Parallèlement, les travaux sur l'archivage des bases de données connaissent un ralentissement. Il faut attendre 2020[21] pour assister à une reprise de la réflexion sur le sujet sans doute explicable par différents facteurs.

Tout d'abord, bien loin de diminuer, l'usage des bases de données dans les administrations croît de manière exponentielle ces dernières années car comme l'évoque Hélène Bégon-Tavera dans son ouvrage *La transformation numérique des administrations*, les années 2010 sont marquées par l'explosion des usages du numérique[22]. Par conséquent, les versements de données structurées sont de plus en plus fréquents, plus seulement restreints au niveau central mais désormais également dans les territoires.
Par ailleurs, les premières collectes de boîtes mails survenues à la toute fin des années 2010 confrontent les archivistes à devoir conserver un nouvel objet complexe. Ces données, pourtant nativement numériques, ne s'archivent pas comme des documents bureautiques. Générées dans des logiciels informatiques tels que Gmail, Outlook ou Thunderbird dont on ne peut actuellement pas garantir l'ouverture d'ici plusieurs dizaines d'années, elles doivent donc pouvoir être conservées indépendamment de ceux-ci. Le début des années 2020 est finalement l'occasion de remettre au centre des préoccupations du monde archivistique la dissociation entre les différents types de données nativement numériques et les enjeux de conservation posés par chacun d'eux. Si des travaux tels que celui de Joël Poivre permettent à l'archiviste de savoir comment archiver un document bureautique, des divergences subsistent tant au niveau national qu'international sur la manière de

---

19 *Loc. cit.*

20 SIN BLIMA-BARRU Martine, VAN DE WALLE Thomas, L'archivage numérique aux Archives Nationales : de Constance à ADAMANT, *La Gazette des archives*, 2015, n°240, pp. 73-74.

21 Propos issus d'un échange conduit avec Martine Sin Blima-Barru, Mathias Fereira et Emeline Levasseur (département de l'administration des données aux Archives Nationales), le 6 juin 2023.

22 BÉGON-TAVERA Hélène, *La transformation numérique des administrations*, La documentation française, 2021. Dans cet ouvrage, l'autrice opère une séparation en quatre âges : les années 1960-1970 correspondent à l'âge de l'informatique, les années 1990-2000 sont l'âge d'internet et de la société de l'information, les années 2010 marquent l'explosion des usages du numérique tandis que les années 2020 équivalent au début de l'intelligence artificielle et des doutes.

conserver les bases de données, tout l'enjeu résidant dans la conservation concomitante des informations et de leur structuration.

1. 3 : Besoins du service interministériel des Archives de France : contextualisation du stage

C'est de par son rôle d'administration centrale que le Service interministériel des Archives de France (SIAF) fut amené à réfléchir sur l'archivage des bases de données. En effet, nous l'avons vu précédemment, ces dernières années marquent un regain d'intérêt unanime pour les questions y ayant trait. Avant d'exposer ici les besoins identifiés, présentons cette institution et ses différentes missions.
Le Service interministériel des archives de France est créé par l'arrêté du 17 novembre 2009[23]. Il fait partie de la direction générale des patrimoines et de l'architecture (DGPA), l'une des quatre composantes principales du ministère de la Culture. Son organisation fut fixée par l'arrêté du 31 décembre 2020 relatif à l'organisation de la direction générale des patrimoines[24]. Le SIAF définit, coordonne et évalue l'action de l'État en matière de collecte, de tri, de classement, de description, de conservation et de communication des archives publiques, à l'exception de celles conservées par le ministère des Armées et le ministère de l'Europe et des Affaires étrangères, qui disposent d'une autonomie dans ce domaine. Il exerce sa tutelle sur les trois services à compétences des Archives nationales mais aussi sur les services d'archives des collectivités territoriales ainsi que sur les organismes autorisés, à titre dérogatoire, à conserver eux-mêmes leurs archives publiques définitives. Au-delà de son action de coordination et de contrôle, le SIAF exerce également un rôle de conseiller pédagogique auprès des divers services d'archives en organisant des journées d'études, des réunions diverses ou encore des cycles de formation.

Il se divise en deux sous-directions : d'une part celle chargée du pilotage, de la communication et de la valorisation des archives et d'autre part, celle en charge de la collecte, de la conservation et de l'archivage électronique[25]. Cette dernière est composée de trois bureaux : celui chargé du contrôle, de la collecte, des missions et de la coordination interministérielle, celui de la protection du patrimoine archivistique et enfin, celui de l'expertise numérique et de la conservation durable, à l'initiative du stage sur l'archivage des bases de données. Piloté par Violette Lévy et composé de quatre experts aux compétences diverses et complémentaires, il se charge d'animer la politique nationale en matière d'expertise numérique et de conservation durable des archives. À ce titre, il intervient tant sur l'accompagnement des administrations mettant en œuvre l'archivage électronique que sur les projets de construction de bâtiments publics d'archives et de définition de normes de conservation préventive. Par ailleurs, dans le contexte du cadre stratégique de

[23] En réalité, l'institution existe depuis 1897 et est connue jusqu'en 2009 sous le nom de Direction des archives de France.

[24] Page dédiée au service interministériel des Archives de France sur le site de FranceArchives : https://francearchives.gouv.fr/article/26287441, consultée le 7 juillet 2023.

[25] Organigramme du SIAF, sur la page du SIAF de FranceArchives ; https://francearchives.gouv.fr/file/9a480c69b465f8f616abff287f27d84a00e9cb11/2021_octobre_organigramme_SIAF.pdf, consulté le 7 juillet 2023.

modernisation des archives 2020-2024[26], il travaille avec l'ensemble des directions d'archives ministérielles pour identifier des solutions permettant l'industrialisation de l'archivage numérique, la conservation et l'accès durable aux ressources patrimoniales, qu'elles soient matérielles ou numériques.

Centralisant les besoins des services d'archives et présent sur la scène internationale via divers sujets, le SIAF a depuis longtemps déjà, perçu les enjeux autour de l'archivage des bases de données. Ainsi, entre 2010 et 2012, deux stages sont réalisés par deux étudiants du master Technologies numériques appliquées à l'histoire de l'École nationale des chartes. Le premier, effectué en 2010 par Baptiste Nichele vise à étudier un nouveau format d'archivage des bases de données relationnelles développé en 2004 par les Archives fédérales suisses : SIARD. En effet, si, comme nous l'avons vu, la méthode Constance visant à mettre à plat les données d'un côté et à conserver la documentation de l'autre, en place depuis les années 80, semble avoir fait ses preuves en France, cela n'empêche pas la mise au point d'autres stratégies d'archivage. Le travail de Baptiste Nichele a donc consisté à expertiser la composition et la structure de ce format grâce à sa documentation afin de savoir s'il était applicable au contexte français et surtout, si son utilisation présentait des avantages par rapport à un archivage à plat, notamment pour la prise en compte des métadonnées. De cette étude ont été tirés plusieurs constats dont le fait que le format SIARD permet de récupérer de manière automatisée aussi bien les données que leur structure ainsi que la totalité des contraintes et autres métadonnées. Cependant, il semble que soit constaté à cette époque un décalage entre les fonctionnalités promises par la documentation et celles réellement disponibles[27].

Un deuxième stage, effectué en 2012 par Marion Ville est venu compléter cette première étude[28]. En effet, après un premier travail théorique réalisé sur le format, il semblait important pour le SIAF de l'expérimenter sur un cas concret, or, en France, la matrice cadastrale est archivée au format SIARD. Ainsi, toujours dans l'optique de savoir si ce format pouvait être utilisable en France, des tests ont été réalisés sur les données migrées en SIARD.

Notre stage, placé dans la continuité des deux précédents, répond pour sa part, à un triple objectif. Tout d'abord, une nécessité d'actualisation : étudier les ressources et les outils méthodologiques existants aujourd'hui pour l'archivage de bases de données. Il s'agit également de faire face à cette massification des projets incluant l'archivage de bases de données en proposant aux différents interlocuteurs des solutions pérennes, tenant compte de tous les enjeux que l'objet base de données implique. En effet, dans le cadre de son rôle d'accompagnateur, le SIAF initie et pilote deux grands projets : le dispositif Interministériel d'Accompagnement aux Missions pour l'Archivage électronique[29] (DIAMAN) et Archivage numérique en

---

26 Cadre stratégique de modernisation des archives, 2020-2024 pris sur le site du Gouvernement : https://www.gouvernement.fr/sites/default/files/contenu/piece-jointe/2020/04/cadre-strategique-archives_2020-2024_affiche-a3.pdf, consulté le 7 juillet 2023.

27 D'après ce stage a été réalisé le mémoire suivant : NICHELE Baptiste, *Interopérabilité et pérennisation des archives électroniques. L'exemple du SEDA/EAD et du SIARD (étude de cas),* mémoire de master Technologies numériques appliquées à l'histoire, École nationale des chartes, 2010.

28 VILLE Marion, *La matrice cadastrale : archiver et exploiter une base de données*, mémoire de master Technologies numériques appliquées à l'histoire, École nationale des chartes, 2012.

29 Appel à projets DIAMAN 2023-2024, France Archives : https://francearchives.gouv.fr/fr/article/224585369, consulté le 7 juillet 2023.

Territoires[30] (ANET). Tandis que le premier tend à soutenir les administrations centrales qui s'engagent dans des actions visant à développer des solutions d'archivage numérique, le second vise à épauler les collectivités territoriales engagées dans des actions permettant de déployer des plateformes d'archivage numérique intermédiaire et/ou définitif.
Enfin, le SIAF, de par sa présence sur la scène internationale a été amené à s'interroger sur les méthodologies mises en oeuvre par ses voisins européens pour l'archivage de données structurées. En effet, alors que le format SIARD était récent et peu connu dans les années 2010, il est désormais maintenu par un consortium international et utilisé par plus d'une cinquantaine de pays. Le SIAF souhaitait donc réfléchir, d'après une documentation actualisée et au regard des autres méthodes d'archivage, à d'éventuelles possibilités d'application étendue au contexte français[31].

Finalement, l'archivage des bases de données vu sous le prisme du SIAF prend sens à la fois dans son rôle de centralisateur des besoins de services d'archives centraux et territoriaux mais aussi dans celui de pédagogue tendant à faire connaître à ses interlocuteurs les différentes solutions d'archivage existantes. Pleinement inscrite dans le cadre stratégique de modernisation des archives 2020-2024, cette réflexion soutient trois vocations majeures du SIAF : identifier et lever les obstacles culturels techniques et organisationnels à la massification de l'archivage numérique, archiver au niveau central les données des services déconcentrés de l'État issues d'applications développées et maintenues au niveau central et mieux coordonner l'action archivistique internationale.

30 Appel à projets ANET 2023-2024, France Archives : https://francearchives.gouv.fr/fr/article/171593987, consulté le 7 juillet 2023.

31 Le 21 septembre 2010, le SIAF publie une note d'information relative à l'étude du format SIARD pour l'archivage des bases de données relationnelles et au logiciel SIARDSuite mettant en oeuvre ce format, disponible sur FranceArchives, https://francearchives.gouv.fr/fr/file/257d593cdd2b9adfe46428e558559f57655ccd70/DGP_SIAF_2010_017.pdf, consulté le 24 avril 2023.

Chapitre 2 : Préalables archivistiques

2. 1 : Notions clefs

Quand on veut archiver une base de données, il est important d'avoir en tête certaines notions incontournables. Tout d'abord, l'archivage est une « activité qui consiste à gérer et organiser l'information dans le temps, quel que soit son support, pour la rendre accessible durablement, bien au-delà de la durée de vie des supports[32]. » Réduire l'archivage aux notions de préservation et de conservation reviendrait donc à considérer qu'il ne s'agit que d'un travail de maintien en état des documents. Or, archiver nécessite idéalement tant un travail en amont de la collecte qu'un travail en aval. L'archivage de documents papiers comme numériques repose avant tout sur deux concepts corrélés : l'intégrité et l'authenticité. L'intégrité est définie dans le Référentiel Général de Gestion des Archives comme étant la « qualité d'un document ou d'une donnée qui n'a pas été altéré[33]. » L'authenticité quant à elle, est entendue comme la garantie que le document archivé est exactement conforme au document originel produit et utilisé par le producteur dans le cadre de son activité[34]. Il revient donc à l'archiviste de veiller à ce que les informations archivées n'aient subi aucun traitement susceptible d'en avoir modifié le caractère d'exactitude originelle. Le document donné à voir dans le futur doit être fidèlement identique à celui produit. Mais alors, quels documents archive-t-on et pourquoi ?

Pour qu'un document ait un intérêt archivistique, il faut qu'il réponde à un ou plusieurs des critères suivants : il a une valeur de preuve juridique, engageant l'entité morale ou physique concernée ou bien il a un intérêt historique, permettant de constituer ou d'alimenter la mémoire des activités dont il est question[35]. Tous les documents produits par une organisation n'ont donc pas vocation à être archivés.
L'archivage des documents numériques s'inscrit dans un cadre conceptuel qu'il semble important d'évoquer ici : l'Open Archival Information System. Mis au point par le Consultative Committee for Space Data Systems et normalisé par l'ISO, il fournit un cadre théorique applicable à toutes les organisations conservant de l'information. Conçu pour être indépendant de toute évolution technologique, il définit un vocabulaire commun et modélise les responsabilités, les fonctions et l'organisation nécessaires pour assurer la pérennité de l'information numérique[36]. Le modèle de gestion qu'il propose est bâti autour de la notion de paquets d'informations contenant à la fois les données à archiver mais aussi les métadonnées associées. L'OAIS définit quatre acteurs : l'archive qui est l'acteur principal, présent à toutes les étapes du processus

---

[32] Définition issue de l'ouvrage suivant : Association des archivistes français, *Les archives électroniques*, Paris, 2020.

[33] Définition issue du Référentiel Général de Gestion des Archives, https://www.gouvernement.fr/sites/default/files/contenu/piece-jointe/2014/07/r2ga_document_complet_201310.pdf, consulté le 25 juillet 2023.

[34] Définition issue du travail suivant : « Conditions requises pour évaluer et maintenir l'authenticité des documents d'archives électroniques », International Research in Permanent Authentic Records in Electronic Systems (InterPARES), http://www.interpares.org/display_file.cfm?doc=ip1_authenticity_requirements(french).pdf, consulté le 25 juillet 2023.

[35] Ces critères sont définis dans un document interne à la Mission des archives auprès du ministère des Affaires sociales intitulé « Le cycle de vie des données d'une applications », février 2019.

[36] Association des archivistes français, *Abrégé d'archivistique, Principes et pratiques du métier d'archiviste*, Paris, 2020, p. 62.

mais également le management, les producteurs et les utilisateurs qui gravitent autour d'elle. On considère que le management est le décideur politique. Le producteur lui, fournit l'objet à archiver qu'on appelle aussi le « submission information package » (SIP). Une fois versé, le SIP devient un « archival information package » (AIP) dont des copies, ou « dissemination information packages » (DIP) seront mises à disposition des utilisateurs demandeurs[37]. Ce modèle, aujourd'hui largement utilisé, sert de cadre conceptuel à la plupart des systèmes d'archivage électronique français mais également internationaux. Il permet de garantir l'interopérabilité entre les différentes institutions productrices et conservatrices d'informations. Il sert également de base fonctionnelle à d'autres normes ou standards tels que le Référentiel général d'interopérabilité (RGI) ou bien le Standard d'échange de données pour l'archivage (SEDA). Les bases de données, en tant qu'objets numériques, n'échappent donc pas à ce cadre conceptuel.

Après avoir redéfini quelques notions générales nécessaires pour envisager l'archivage d'un document, qu'il soit papier ou numérique, concentrons-nous désormais plus spécifiquement sur l'archivage d'une base de données.

2. 2 : L'évaluation d'une base de données

Le processus d'archivage ne débute pas seulement lorsqu'une base de données entre dans l'institution vouée à la conserver mais commence bien avant, lors de la collecte. Cette étape recouvre l'ensemble des opérations intellectuelles et matérielles conduisant au transfert d'une ou plusieurs archives d'un service producteur vers un service d'archives[38]. Elle inclut l'évaluation, la sélection puis la mise en œuvre du sort final. L'évaluation est l'opération archivistique visant à déterminer l'intérêt public, administratif ou historique du document ou de l'ensemble des documents considérés[39]. Comme les archives papier ou bureautiques, les bases de données doivent aussi être évaluées. On va pour cela procéder en plusieurs étapes.

Tout d'abord, il convient d'identifier le cycle de vie des données contenues dans la base. En effet, chaque donnée, quel que soit son support, suit, à partir de sa création, un cycle de vie en fonction de l'usage pour lequel elle a été créé et parfois également, d'exigences juridiques. Traditionnellement, ce cycle est séquencé en trois moments consécutifs[40]. Dans un premier temps, lorsque le document est fréquemment utilisé pour les besoins pour lesquels il a été crée, il est appelé « archive courante ». Puis, dès lors que son utilisation est plus ponctuelle, qu'on le conserve à titre de référence ou pour sa valeur probante, on considère qu'il devient une « archive intermédiaire ». Enfin, à partir du moment où le document n'est plus utilisé pour les besoins pour lesquels il a été créé, qu'il n'a plus de valeur probante mais revêt un intérêt à être conservé à titre

---

[37] Définition des acteurs et de leurs fonctions empruntée à la page dédiée au modèle de référence OAIS sur le site WordPress : https://archivengines.wordpress.com/2011/12/07/le-modele-de-reference-oais, consulté le 26 juillet 2023.

[38] Définition de « collecte » disponible sur le site du PIAF : https://www.piaf-archives.org/sites/default/files/bulk_media/m02s1/co/02section1_38.html, consulté le 26 juillet 2023.

[39] *Cadre méthodologique pour l'évaluation, la sélection et l'échantillonnage des archives publiques*, Délégation interministérielle aux Archives de France, Juillet 2014, https://francearchives.gouv.fr/file/5f01f41db3790b5201ff6c29413c16521a57ccf6/static_7742.pdf, consulté le 26 juillet 2023.

[40] *Guide méthodologique pour l'archivage des bases de données* élaboré par le Centre d'informatique national de l'Enseignement Supérieur (CINES), https://www.cines.fr/wp-content/uploads/2022/05/GM_archivage_BDD-v1.1.pdf (consulté le 8 avril 2023).

historique, c'est une « archive définitive ». L'identification du cycle de vie est faite en étroite collaboration entre l'archiviste et le service producteur. Il est également nécessaire, pour chaque type de données, de déterminer bien en amont de l'archivage, la 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[41] ». La durée d'utilité administrative couvre donc surtout les âges courant et intermédiaire. À partir de là, plusieurs questions peuvent être posées : le cycle de vie est-il le même pour toutes les données contenues dans la base ? Toutes les données ont-elles la même durée d'utilité administrative (DUA) ?

Une fois le cycle de vie des données identifié, il convient de caractériser l'usage fait de la base de données, étape essentielle pour déterminer la meilleure stratégie d'archivage. L'archiviste peut se demander si les données contenues dans la base qui l'intéresse sont figées dès leur création ou bien, si, au contraire, elles sont couramment modifiées. En effet, on ne conservera pas de la même manière une base de données toujours alimentée, d'une base de données, qui, elle ne subit plus de modifications. On distinguera alors une base de données « figée » qui ne subit plus aucun ajout, modification ou effacement, d'une base de données « vivante » dont les données sont encore modifiées[42]. Parmi elles, il convient également de dissocier celles dites « consultées » sur lesquelles un grand nombre de consultations est fait, de celles appelées « cumulatives » sur lesquelles on ne fait qu'ajouter de nouveaux éléments sans en modifier ou en effacer, de celles « dynamiques » pour lesquelles l'ajout et la modification sont l'un et l'autre encore en vigueur[43].

Après avoir identifié le cycle de vie des données et caractérisé l'usage de la base, il s'agit ensuite d'évaluer la confidentialité des données contenues. En effet, il est particulièrement fréquent qu'une base de données contienne des informations confidentielles. Cela implique à la fois les données à caractère personnel, c'est à dire « toute information se rapportant à une personne physique identifiée ou identifiable[44] » ou encore les identifiants et mots de passe utilisés pour accéder à cette base. L'identification de ces différents types d'informations permettra de prendre les mesures nécessaires. Les données à caractère personnel sont encadrées par le Règlement Général sur la Protection des Données (RGPD). Il convient donc d'avertir la Commission nationale de l'informatique et des libertés (CNIL) afin de savoir si les données peuvent être conservées au-delà de leur durée d'utilité administrative (DUA). Dans le cas des identifiants et mots de passe, il suffit de paramétrer les exports de la base de sorte à les rendre invisibles lors de l'archivage. Enfin, une fois les différents enjeux identifiés, il convient de dresser un état des lieux de l'existant, c'est à dire de lister toutes les informations connues et inconnues sur la base de données. Parmi les éléments à

---

41 Association des archivistes français, *Abrégé d'archivistique, Principes et pratiques du métier d'archiviste*, Paris, 2020, « Durée d'utilité administrative (DUA) » p. 331.

42 *Guide méthodologique pour l'archivage des bases de données* élaboré par le Centre d'informatique national de l'Enseignement Supérieur (CINES), https://www.cines.fr/wp-content/uploads/2022/05/GM_archivage_BDD-v1.1.pdf (consulté le 8 avril 2023).

43 *Loc. cit.*

44 Définition de « donnée personnelle » d'après le site de la CNIL https://www.cnil.fr/fr/rgpd-de-quoi-parle-t-on, consulté le 25 juillet 2023.

renseigner, on notera le nom et la taille de la base, l'application l'utilisant, l'identification et la localisation des SGBD sur lesquels elle est hébergée, la liste des utilisateurs, les dates d'entrée et de fin de service, les objectifs répondant à sa création, ses différentes fonctionnalités, un descriptif général du type d'informations contenues ainsi que la fréquence des mises à jour.
Pour que l'archivage soit le plus complet possible, il est important d'accompagner la base d'un maximum d'informations, c'est pourquoi une attention particulière doit par ailleurs être portée aux métadonnées qui permettent de décrire la base, l'origine de sa création, son utilité et ses destinataires. Elles sont classiquement divisées en trois types : les métadonnées dites « de gestion » renseignant les modalités d'accès à la base, mais également les métadonnées de description, qui permettent de comprendre le contenu de la base, et enfin, les métadonnées de préservation qui sont là pour garantir la pérennité de l'accès à cette base[45]. Si au terme de tout ce processus, des informations venaient à manquer, l'archiviste devrait effectuer un travail de rétro-ingénierie, sujet sur lequel nous reviendrons ultérieurement.

L'évaluation, entendue comme le moyen de décrire l'existant et de prendre conscience des données contenues dans la base de données à archiver est une étape essentielle qui permettra finalement à son issue, de décider quel sort appliquer à celle-ci.
En fonction des résultats précédents, plusieurs possibilités s'offrent à l'archiviste. Dans un premier cas, les données à conserver sont versées au service d'archives. Parmi elles, celles dont la durée d'utilité administrative n'est pas échue seront traitées comme des archives intermédiaires. Les données n'ayant pas vocation à être conservées sont, quant à elles, éliminées[46].
Une fois ce travail préalable et indispensable effectué, vient le moment d'archiver la base de données. Concentrons-nous donc désormais sur les différentes pratiques actuelles en termes d'archivage de bases de données.

---

45 *Guide méthodologique pour l'archivage des bases de données* élaboré par le Centre d'informatique national de l'Enseignement Supérieur (CINES), https://www.cines.fr/wp-content/uploads/2022/05/GM_archivage_BDD-v1.1.pdf (consulté le 8 avril 2023).

46 *Loc. cit.*

Chapitre 3 : Archiver une base de données : focus sur les pratiques actuelles

Comme nous l'avons vu précédemment, l'évaluation est une étape essentielle permettant à l'archiviste de choisir la stratégie d'archivage la plus adaptée à la base de données qu'il doit traiter. Ainsi, la volumétrie, la structure, le type de liens entre les données, la volatilité, le contexte de création et d'utilisation de cette base ainsi que la potentielle confidentialité des données sont autant d'éléments à prendre en compte en amont de l'archivage. C'est pourquoi ce choix doit être pensé à la lumière des caractéristiques de la base, des fonctionnalités à préserver et des types d'utilisations futures à garantir.

3. 1 : L'archivage à plat

L'archivage dit « à plat » est l'une des différentes méthodes existantes pour migrer (ou exporter) des données. La migration consiste en l'export des données sous une forme indépendante du système de gestion de base de données ou sous une forme permettant leur import dans une autre application dotée d'un système de gestion de base de données différent[47]. Il existe deux types d'exports : l'export partiel qui consiste à extraire une sélection d'informations et l'export total, qui, lui, permet de préserver l'intégralité des informations contenues dans la base. Au-delà du type d'export, il est également possible de choisir le format dans lequel on souhaite migrer les données. Ce choix réside couramment entre le format XML ou le format CSV qui sont les plus fréquemment utilisés. Si le premier est très exploité pour les échanges de données, notamment dans le cadre du standard d'échange de données pour l'archivage (SEDA), c'est traditionnellement le second que l'on utilise en France.

Nous l'avons vu, l'archivage à plat est la méthode utilisée par les Archives Nationales pour la conservation des données structurées depuis le début des années 1980 et le lancement de Constance. En effet, à cette époque avait été fait le choix d'adopter une méthode de stockage des données qui soit la plus indépendante possible d'outils et de logiciels. L'archivage à plat consiste à exporter chaque table d'une base de données dans un fichier « plat » de type .csv ou .txt. Les données contenues dans chaque fichier sont encodées à l'aide d'un ensemble défini de caractères, généralement ASCII[48] ou UTF-8[49]. Cette technique permet de conserver uniquement les données de la base en les rendant indépendantes de leur structure originelle. Les valeurs, auparavant contenues dans des cellules, sont désormais séparées les unes des autres par un séparateur, c'est à dire un caractère spécial, généralement un point-virgule, une tabulation ou une barre verticale. Tout l'enjeu autour du séparateur réside dans le fait de choisir un caractère typographique qui ne soit pas contenu dans les données. Les fichiers, une fois exportés en .csv peuvent ensuite être ouverts à l'aide d'un éditeur de texte ou d'un tableur. La documentation ainsi que les informations de représentation, essentielles pour comprendre les données, sont stockées parallèlement, dans d'autres fichiers. Si ce n'est pas

---

47 *Loc. cit.*

48 ASCII : American Standard Code for Information Interchange

49 UTF-8 Universal Character Set Transformation Format - 8 bits

le cas, il appartiendra à l'archiviste de tenter de la reconstituer de sorte à rendre les données conservées exploitables. Documenter la structure de la base de données, la façon dont les différentes tables sont interconnectées et la clé étrangère nécessaire pour faire les liens entre elles est essentiel. Aussi, l'archivage à plat exige une documentation de la structure de chaque table. Ainsi, le libellé de chaque colonne, la définition de chaque valeur, leur caractère obligatoire ou non ainsi que le type de la valeur doivent être explicités. Enfin, il est important de documenter d'une part, les processus métiers opérés à partir de la base de données, c'est à dire décrire à quoi sert la base et comment elle s'utilise et d'autre part, les choix d'export[50].

Pour accéder à une base de données archivée à plat, l'utilisateur aura donc autant besoin des fichiers CSV correspondant à chacune des tables que de la documentation stockée à côté. L'accès peut se faire par deux biais : sur un tableur (type Excel) ou en important les tables dans une base de données en utilisant un système de gestion de base de données SQL par exemple, ce qui permettra ensuite à l'utilisateur d'effectuer des requêtes.

Dans le but de connaître les pratiques des services d'archives départementaux en termes d'archivage de bases de données locales, nous avons établi un questionnaire sur le logiciel Framaforms avec l'aide de Violette Lévy, Dominique Naud et Louis Vignaud et l'avons diffusé à l'ensemble du réseau le 21 juin 2023[51]. Afin de conserver du temps pour traiter les résultats de cette enquête, nous avons fixé une date limite de réponse au 7 juillet. Sur les 99 services concernés, dix-huit ont répondu. Bien que pouvant sembler très faible, nous avons considéré que cet échantillon serait exploitable pour nous faire une idée des pratiques locales. Si plusieurs sujets y sont traités, nous nous concentrerons exclusivement ici sur les modalités d'archivage des bases de données. Après avoir renseigné les informations permettant d'identifier le répondant, à savoir son nom, son prénom, sa fonction au sein du service, son courriel et le service d'archives auquel il appartient, le répondant était amené à répondre à la question suivante : « Disposez-vous d'un service d'archivage électronique? ». Sur les dix-huit répondants, dix-sept ont répondu « Oui » et un « Non ». A la question suivante : « Des opérations d'archivage de bases de données sont-elles réalisées dans votre service? », onze services ont répondu « Oui » contre sept « Non ». Si cette question ne traite pas directement des modalités de collecte des bases de données, elle n'en est pas moins intéressante car elle nous permet de voir que les bases de données ne concernent pas seulement les grands services d'archives centraux.

Après cela, il semblait important de s'intéresser à la fréquence de ces collectes. Trois possibilités étaient proposées : tous les x ans, zéro à une fois par an ou plusieurs fois par an. Si pour la plupart, la collecte de bases de données relève de l'exceptionnel[52], pour deux services d'archives, c'est une opération qui se répète plusieurs fois par an. En ce qui concerne le mode d'archivage utilisé, on note que dix services recourent à l'export au format CSV, dont les deux services réalisant plusieurs collectes de bases de données par an. L'archivage à plat est donc utilisé non seulement pour archiver les bases conséquentes statistiques au niveau central mais également pour l'archivage de bases locales.

---

50 Propos extraits du cours sur les bases de données d'Edouard Vasseur à Abu Dhabi.

51 Message accompagnant l'envoi du questionnaire au réseau. Voir annexe n°1.

52 Sur les dix-huit services répondants, dix ont répondu « zéro à une fois par an » et six ont répondu « tous les x ans ».

Ainsi, méthode indépendante des outils et logiciels, peu coûteuse en termes de place et dont on sait qu'elle fonctionne car elle permet actuellement d'ouvrir des données structurées archivées depuis les années 80, l'archivage à plat est un moyen sûr pour conserver les données brutes. Adaptable tant aux bases de données en production qu'à celles décommissionnées, aux grandes bases statistiques comme aux bases locales contenant des données individuelles, il offre plusieurs moyens d'accéder aux données en fonction des besoins utilisateurs. Enfin, très connu des chercheurs, le CSV est un format de fichier largement utilisé dans des communautés allant même bien au-delà de celle des archivistes et historiens.
Néanmoins, c'est une méthode qui comporte également quelques inconvénients. Parmi eux, nous pourrions citer le fait que les données et métadonnées soient stockées dans des fichiers séparés ou encore le fait que la base de données originale ne soit pas immédiatement accessible et consultable. C'est pour cela que l'archivage à plat n'est pas l'unique solution d'archivage de bases de données : d'autres possibilités, qu'il nous faut désormais exposer ont été mises au point.

3. 2 : Le format SIARD

L'archivage à plat n'est pas la seule méthode existante pour l'archivage des bases de données. En effet, comme nous l'avons vu plus haut, les années 2000 sont marquées par un accroissement massif du nombre de bases de données et une accélération de la réflexion sur les modalités d'archivage de ce type de données. C'est donc dans ce contexte qu'en 2008, les Archives fédérales suisses (AFS) mettent au point le format SIARD et la suite logicielle associée, SIARDSuite. SIARD, de l'anglais Software Independent Archival of Relational Databases, est un format ouvert et indépendant permettant de récupérer de manière automatisée tant le contenu de la base de données que sa structure mais aussi les relations entre les tables et les informations de gestion. Compatible avec le modèle OAIS, le format SIARD est également basé sur plusieurs normes ISO telles qu'UNICODE[53], XML[54], SQL:2008, URI[55] et ZIP[56]. En s'appuyant sur des normes internationales, le format garantit donc une grande interopérabilité entre les services.
En 2008, le projet européen PLANETS œuvrant pour l'archivage des bases de données relationnelles fait de SIARD son format d'archivage officiel. En 2013, à l'échelon suisse, l'association eCH fait de SIARD un standard : eCH-0165[57]. Le 31 août 2021, le DILCIS Board (Digital Information Life Cycle Interoperability Standards Board), les Archives fédérales suisses et le projet européen E-ARK mettent conjointement au point

---

53 UNICODE est la version courte de « Universal Character Encoding ». Il s'agit d'une norme standardisée pour le codage des caractères en représentation binaire. https://www.ionos.fr/digitalguide/sites-internet/creation-de-sites-internet/unicode/

54 XML (Extensible markup language) est une norme qui permet de structurer l'information de manière hiérarchique en imbriquant des éléments. XML permet de définir des formats que devront respecter les documents qui seront créés. (Définition issue de l'abrégé d'archivistique, p. 334)

55 Uniform Resource Identifier (URI) permet d'identifier les ressource abstraites ou physiques sur internet. A partir d'un URI, un système peut lire l'information, connaître son emplacement et savoir par quel moyen elle est accessible. https://www.ionos.fr/digitalguide/sites-internet/developpement-web/le-uniform-resource-identifier/

56 Le ZIP est un format de fichier permettant l'archivage et la compression de données sans perte de qualité. (Définition prise sur Wikipedia, rajouter le lien)

57 Site de l'association eCH, https://www.ech.ch/fr/ech/ech-0165/1.0 (consulté le 7 août 2023).

la version 2.0 du format. SIARD est aujourd'hui utilisé dans plus de 54 pays et disponible en quatre langues : anglais, allemand, français et italien[58].

Pour mettre en oeuvre le format SIARD, la suite logicielle SIARDSuite est requise. Également créée par les Archives fédérales suisses, elle a deux fonctions principales : extraire le contenu d'une base de données relationnelle et le sauvegarder au format SIARD mais aussi ré-importer les informations d'une archive au format SIARD dans un système de gestion de bases de données compatible. En effet, SIARDSuite peut extraire et remonter des bases de données seulement dans les systèmes de gestion de base de données suivants : MS Access, DB/2, MySQL, Oracle, PostgreSQL et SQL Server. Si la suite logicielle couvre donc les SGBD les plus couramment utilisés, elle ne peut pas traiter de bases de données provenant d'autres logiciels de gestion.

Afin de conserver les relations entre les données des différentes tables, SIARD considère une base de données relationnelle comme une entité unique et indivisible. Toute la base à archiver est donc stockée dans un conteneur ZIP 64. À l'intérieur figurent deux dossiers : un premier nommé « header » qui stocke toutes les métadonnées décrivant la structure de la base et un second appelé « content » qui, lui, contient les données tabulaires. Dans chaque dossier figurent deux types de fichiers : un fichier .xml et un autre .xsd. Dans le header, le fichier metadata.xml contient les métadonnées tandis que metadata.xsd correspond au schéma de données du fichier xml. Pour le content, la logique est la même. À chaque table de la base de données correspondent deux fichiers : un .xml et un .xsd. Afin d'illustrer cela, reprenons un exemple donné par Baptiste Nichele dans son mémoire : une base de données possédant 50 tables exportée au format SIARD aura, dans son dossier content, 50 fichiers tables .xml et .xsd numérotés chacun depuis 0 jusqu'à 49[59].
SIARD est donc un format d'archivage de bases de données relationnelles ouvert, basé sur des normes internationales, qui, contrairement à l'archivage à plat, propose de stocker données et métadonnées dans deux dossiers séparés mais dans un conteneur unique. La structuration explicite des données et métadonnées confère au format une capacité d'automatisation de traitement et permet de décrire aisément les relations entre les tables ainsi que les informations de gestion. Par le biais de SIARDSuite, il est alors possible de remonter l'intégralité de la base de données archivée et de naviguer dans son arborescence, d'en importer et/ou d'en exporter les données et de les remonter dans un autre système de gestion de bases de données, rendant de facto la structure indépendante du SGBD d'origine.

Tout cela nous conduit donc à nous interroger : pourquoi ce format, très utilisé à l'étranger n'a fait l'objet que de rares expérimentations en France ? En 2010, Baptiste Nichele pointe dans son mémoire quelques limites du format. Il explique notamment que SIARD, bien que supportant les éditeurs les plus couramment rencontrés, n'est malgré tout pas exhaustif et se prive donc de plusieurs réalités différentes. Il relève aussi que le format est entièrement dépendant de SIARDSuite, seule suite logicielle permettant de le

---

58 Spécification de la version 2.1.1 du format SIARD, https://siard.dilcis.eu/SIARD 2.2/SIARD 2.2.pdf, (consulté le 7 août 2023).

59 NICHELE Baptiste, *Interopérabilité et pérennisation des archives électroniques. L'exemple du SEDA/EAD et du SIARD* (étude de cas), mémoire de master Technologies numériques appliquées à l'histoire, École nationale des chartes, 2010.

manipuler[60]. D'autres travaux, plus récents ainsi que des retours d'expériences mettent en exergue d'autres limites[61]. Parmi elles, la nécessité d'une collaboration entre service producteur et service d'archives compliquant l'archivage de bases de données déjà décommissionnées, la difficulté d'effectuer une sélection au sein d'une base de données puisque SIARD l'archive dans son intégralité ou encore la volumétrie d'une archive SIARD par rapport à celle d'une base de données archivée à plat dans un fichier CSV. En effet, nous avons pu constater que pour une même table conservée au format CSV et au format SIARD, la volumétrie augmentait drastiquement[62]. Par ailleurs, les chercheurs français n'étant pas familiarisés avec ce format, la communauté d'utilisateurs est, pour l'instant, restreinte aux archivistes.

Néanmoins, SIARD étant au même titre que l'archivage à plat, l'une des méthodes d'archivage des bases de données, il semblait important de contacter directement les Archives fédérales suisses afin d'avoir davantage d'informations. Stefan Kwasnitza, directeur suppléant des Archives fédérales suisses et Audun Lund, très actif dans la gestion de la spécification SIARD et dans l'archivage des bases de données, ont donc eu la gentillesse de nous faire une présentation actualisée du format. On en retient qu'au niveau international, il s'agit du format le plus utilisé pour l'archivage des données structurées. Non plus seulement dirigé par les Archives fédérales suisses, il est aujourd'hui piloté par un consortium international, ce qui est représentatif de l'intérêt commun porté pour ce format.

3. 3 : L'émulation

Avec le passage à l'ère numérique, les archivistes sont confrontés à deux enjeux majeurs : la fragilité des supports numériques et l'obsolescence rapide des logiciels informatiques. Ils doivent donc intervenir avant que des données soient devenues illisibles ou soient perdues. Si la migration, à savoir l'action de transférer des documents d'un système à un autre en préservant leur authenticité, leur intégrité, leur fiabilité et leur exploitabilité[63] est traditionnellement utilisée, il existe une alternative : l'émulation[64]. L'émulation consiste à reproduire au moyen d'un émulateur[65] un environnement matériel et logiciel qu'on appelle

---

60 *Loc. cit.*

61 JACOBSON Michel, « Retour d'expérience sur l'utilisation du format SIARD pour l'archivage des bases de données relationnelles », 6 novembre 2014.

62 Lors des expérimentations réalisées sur la base de données sur les sportifs de haut niveau, nous avons exporté une même table d'une part au format CSV et d'autre part au format SIARD. La table validationPole pèse, au format CSV 58 octets et passe à 28 Ko une fois transposée en SIARD.

63 Définition issue du petit glossaire des termes archivistes disponible sur le site de l'Association des Archivistes français : https://www.archivistes.org/petit-glossaire-de-termes#:~:text=- Migration :,leur fiabilité et leur exploitabilité, consulté le 7 août 2023.

64 Dans la note d'orientation de la Digital Preservation Coalition intitulée « Préservation des bases de données. Collection de notes d'orientation sur les types de données » et publiée en juillet 2021, l'émulation est définie comme étant un « moyen de surmonter l'obsolescence technologique du matériel et des logiciels en développant des techniques permettant d'imiter des systèmes obsolètes sur les générations d'ordinateurs futures. », https://www.dpconline.org/docs/dpc-technology-watch-publications/technology-watch-guidance-notes/translations-2/2789-dpc-guidance-note-preserving-databases-fr/file, consulté le 29 août 2023.

65 Un émulateur est un logiciel dont la fonction est de recréer les fonctionnalités d'un ancien environnement technologique sur une plateforme moderne.

« invité » généralement ancien sur un environnement matériel et logiciel « hôte » plus actuel[66]. Cela permet de maintenir l'objet dans sa forme originelle mais d'y accéder en utilisant l'environnement technologique actuel. L'émulation peut être réalisée à plusieurs niveaux : logiciel, système d'exploitation ou plateforme matérielle et il est même possible d'empiler les émulateurs.
L'émulation est théorisée par Jeff Rothenberg à la fin des années 90. Il oppose alors cette technique de préservation à la conversion de formats qu'il considère trop coûteuse et risquée. En effet, selon lui, la migration comporte un grand risque d'atteinte à l'authenticité des documents[67]. Bien que l'approche de Jeff Rothenberg ait été vivement critiquée, l'émulation connaît un regain d'intérêt depuis le début des années 2010 grâce notamment à trois projets : l'émulation de l'ordinateur personnel de Salman Rushdie, le projet Rhizome de la BnF[68] ainsi que le projet européen KEEP[69].

Quelle que soit la technique choisie, le décision dépend selon Laurent Duplouy, chef du service multimédias au département audiovisuel de la BnF, de trois facteurs : l'objet à archiver, l'évaluation des risques et l'intention de préservation. En effet, tandis que la migration vise à préserver l'objet lui-même, l'émulation se concentre davantage sur l'environnement de l'objet numérique. C'est donc l'intention de préservation, c'est-à-dire, ce que l'on cherche à conserver qui va conditionner la méthode la plus appropriée. L'usage le plus répandu de l'émulation se trouve dans le domaine des jeux vidéos. La conservation de cet objet est particulièrement complexe car, comme le souligne Nicolas Esposito dans son article « Émulation et conservation du patrimoine culturel lié aux jeux vidéos[70] », il y a dans le jeu vidéo « un aspect ludique qui ne peut être pleinement appréhendé que par le fait de jouer ». Ainsi, conserver un jeu vidéo implique selon lui de « conserver la possibilité d'y jouer[71] ». Effectivement, comme il en fait mention, Frédéric Le Diberder identifie dans son ouvrage *L'univers des jeux vidéo* paru en 1998, cinq plaisirs indissociables du jeu : la compétition, l'accomplissement, la maîtrise d'un système, le plaisir du récit et le spectacle. Or, il n'est pas possible de goûter à ces notions en conservant seulement l'objet jeu vidéo. Néanmoins, il n'est pas rare que les consoles ou les jeux eux-mêmes soient devenus obsolètes et ne permettent donc plus leur utilisation. C'est en cela que l'émulation semble une technique utile puisqu'elle permet « de reproduire le fonctionnement de la machine d'origine sur une machine plus récente pour y lancer le jeu qui aura été stocké dans un fichier[72]. »

---

66 Chantier de préservation VITAM, compte rendu de la réunion du 17 mai 2019.

67 Emulation as a Digital Preservation Strategy, *D-Lib Magazine*, Volume 6, Number 10, Stewart Granger, UK Project Co-ordinator of the CAMiLEON Project, University of Leeds.

68 Page du projet Rhizome : https://productionsrhizome.org/numerique/preservation-de-lart-numerique-a-la-bibliotheque-nationale-de-france, consultée le 7 août 2023.

69 Page officielle du projet européen KEEP : https://actions-recherche.bnf.fr/bnf/anirw3.nsf/IX01/A2012000845_keep-keeping-emulation-environments-portable, consultée le 30 août 2023.

70 ESPOSITO Nicolas, *Emulation et conservation du patrimoine culturel lié aux jeux vidéo*, Université de technologie de Compiègne, 2004. https://www.researchgate.net/profile/Nicolas-Esposito/publication/215673465_Emulation_et_conservation_du_patrimoine_culturel_lie_aux_jeux_video/links/54e0aab00cf29666378d40a6/Emulation-et-conservation-du-patrimoine-culturel-lie-aux-jeux-video.pdf (consulté le 11 juin 2023).

71 *Loc. cit.*

72 *Loc. cit.*

Bien que traditionnellement utilisée pour la préservation des jeux vidéo ou plus largement, de contenus multimédias interactifs, l'émulation est aujourd'hui l'une des méthodes suggérées pour la conservation des bases de données. C'est notamment la méthode utilisée par le département multimédias de la BnF pour la conservation des bases de données. En effet, comme l'explique Laurent Duplouy lors d'un échange survenu le 6 juin, ce service ne compte pas moins de 3953 bases de données issues de collections diverses[73]. Pour chaque base de données, la stratégie adoptée est la même : il s'agit d'émuler le comportement de la base de données, c'est à dire permettre à l'usager de visualiser les données dans les mêmes conditions qu'à l'époque où elles ont été générées[74]. Ce qui justifie cette approche, c'est l'intention de préservation de l'institution qui, non pas tant axée sur la garantie de réexploitation des données se focalise plutôt sur la promesse d'accéder aux données telles qu'elles ont été diffusées à l'origine : c'est le principe du dépôt légal.

Appliquée aux archives, l'émulation suppose d'opérer une distinction entre deux types de bases de données, d'un côté, celles remplies directement par l'utilisateur via un tableur par exemple, et de l'autre, celles produites et gérées via une application et de fait, cachées derrière une interface graphique. C'est pour ces dernières que l'émulation pourrait être intéressante puisque l'archivage à plat, s'il permet de conserver de manière pérenne l'information générée, suppose malgré tout la perte de l'expérience utilisateur. L'aspiration première de l'archiviste étant de conserver de manière pérenne et de garantir l'accès au contenu de l'information, l'émulation pourrait être envisagée, non pas comme une technique de préservation unique mais davantage comme une méthode complémentaire. Cette possibilité fut notamment envisagée pour l'archivage de l'application Diplomatie du ministère de l'Europe et des Affaires étrangères. En effet, il fut décidé de récupérer d'un côté, la donnée brute et de l'autre, le système de Diplomatie afin de permettre à l'usager de naviguer dans l'application et donc, d'avoir accès à la fois au contenu de l'information mais également à l'expérience utilisateur et ainsi, pouvoir visualiser la base de données de la même manière que l'usager de l'époque. Ce scénario fut néanmoins abandonné car l'émulation aurait nécessité de maintenir à la fois le logiciel et l'émulateur, ce qui a été considéré trop coûteux[75].

L'émulation pose aussi la question des limites. En effet, où s'arrête l'émulation ? Jean-Philippe Humblot l'expose dans son article « Préservation de l'information numérique à la Bibliothèque nationale de France[76] » : dans tous les cas, il ne sera pas possible de reconstituer l'intégralité de l'expérience utilisateur puisqu'un écran cathodique sera remplacé par un écran plat, ce qui altère l'impression visuelle. De plus, les accessoires périphériques, à savoir le clavier, la souris ou bien la manette seront différents de ceux d'origine, ce qui, là aussi entraîne une modification des sensations. Appliquée aux archives, la problématique est la

---

[73] Parmi elles, nous pourrions mentionner à titre d'exemple la base de données de la banque des données des maladies rares ou bien la base de données numérique des sceaux conservés en France.

[74] Propos issus d'un échange avec Laurent Duplouy survenu le 6 juin 2023.

[75] Propos issus d'un échange avec Erwann Ramondenc survenu le 26 mai 2023.

[76] HUMBLOT Jean-Philippe, *Préservation de l'art numérique à la Bibliothèque nationale de France*, https://productionsrhizome.org/numerique/preservation-de-lart-numerique-a-la-bibliotheque-nationale-de-france (consulté le 7 août 2023).

même : faut-il anticiper tous les usages futurs possibles et proposer systématiquement un double archivage ou bien doit-on considérer le travail de conservation effectué dès lors que la donnée brute est préservée et rendue accessible ?

Il semble finalement que les approches soient plurielles, contextuelles, dépendantes des moyens, de l'évaluation des risques, de l'intention de préservation et des usages tant passés que futurs. Ainsi, l'émulation, loin d'être une méthode dominante pour l'archivage des données structurées, semble plutôt être appréhendée comme un moyen de sauvegarde complémentaire, indissociable d'une conservation brute des données.

DEUXIÈME PARTIE - CAS D'USAGE CONCRET : PROPOSER UNE SOLUTION POUR L'ARCHIVAGE D'UNE BASE DE DONNÉES SUR LES SPORTIFS DE HAUT NIVEAU

Nous l'avons vu, il existe aujourd'hui plusieurs solutions pour archiver des bases de données. L'archivage à plat, habituellement utilisé en France permet de conserver de manière pérenne les données brutes. Néanmoins, l'information étant détachée de sa structure originelle, cela suppose une reconstitution manuelle de la structuration de l'information. Marion Ville explique en effet dans son mémoire que « la conservation à plat nécessite de reconstituer manuellement la base pour pouvoir interpréter les données[77]. » Le format SIARD conçu pour répondre aux besoins d'archivage des bases de données relationnelles, est basé sur des normes internationales permettant une grande interopérabilité entre les services. Stockant données et métadonnées dans un même dossier, il promet la visualisation rapide d'une base de données structurée dans le système de gestion de base de données choisi. Très utilisé à l'étranger, il n'est employé que marginalement en France. L'émulation quant à elle, surtout employée pour la conservation de jeux vidéo, semble être davantage envisagée en ce qui concerne les bases de données, comme un moyen de conservation complémentaire. L'enjeu, pas tant au niveau de l'émulation que des deux autres solutions d'archivage, consisterait à savoir si le format SIARD possède réellement un intérêt par rapport à l'archivage à plat. Pourrait-il être appliqué à toutes les bases de données relationnelles, quels que soient leur taille et le cycle de vie des données concernées ? Pourrait-il se substituer à l'archivage à plat ? Devrait-il, lui aussi être considéré comme un moyen de conservation complémentaire ?

Afin de répondre à ces questions, nous nous appuierons désormais sur un véritable cas d'étude : l'archivage d'une base de données sur les sportifs de haut niveau. La description séquentielle des étapes de rétro-ingénierie et des expérimentations, que nous avons trouvé intéressante, vise non pas à raconter notre stage mais résulte plutôt d'une volonté d'inscrire notre démarche dans une temporalité précise à titre d'exemple.

Chapitre 4 : De la présentation du projet au travail de rétro-ingénierie

1. 1 : La base de données sur les sportifs de haut niveau

Dans le but de rendre concrète la réflexion sur l'archivage des bases de données, le service interministériel des Archives de France a choisi d'inclure dans ce stage l'étude d'un cas d'usage concret de base de données à archiver. Le projet est le suivant : il s'agit de réaliser tout le travail préalable à l'archivage et d'expérimenter plusieurs méthodes de conservation, à savoir ici, l'archivage à plat et le format SIARD. Le cas d'étude est une base de données sur les sportifs de haut niveau confiée par la Direction du numérique (DNUM) des ministères sociaux à la mission affaires sociales. Entrée en service en 2011, cette base de données est décommissionnée en 2021 par la Direction du numérique elle-même suite à l'observation d'une inactivité de la base depuis plusieurs années. Comme le spécifie l'*Abrégé d'archivistique*, le décommissionnement ou « retrait de service » est une opération consistant à arrêter les anciennes applications et les infrastructures utilisant des données structurées qui n'évoluent plus ou ne sont plus

[77] VILLE Marion, *La matrice cadastrale : archiver et exploiter une base de données*, mémoire de master Technologies numériques appliquées à l'histoire, École nationale des chartes, 2012.

utilisées par les métiers[78]. Après avoir constaté l'intérêt des données contenues dans cette base, la mission des archives auprès des ministères sociaux décide de la préserver intégralement pour la verser aux Archives Nationales. Conservée sous forme d'une extraction complète Excel, cette base de données contient 110 tables dont certaines comptent parfois plus de 100 000 entrées[79]. Un seul document accompagne cet export ; il s'agit d'un schéma de la base de données également confié par la Direction du numérique. Néanmoins, sur les 110 tables, seules 69 y sont référencées, il est donc incomplet. Deux raisons peuvent expliquer cela : il est possible qu'il corresponde à un état intermédiaire de la base ou bien qu'il s'agisse d'une tentative de rétro-ingénierie incomplète.

Dans la mesure où cette base de données concerne les sportifs de haut niveau, il semblait indispensable avant même de réfléchir à la procédure d'archivage, de se renseigner sur ce statut particulier afin d'en comprendre les enjeux. Pour cela, il fallu donc dépouiller la documentation législative mise à disposition par la mission affaires sociales. On y apprend que selon la loi, la qualité de sportif de haut niveau (SHN) résulte de l'inscription sur une liste arrêtée par le ministère des Sports, selon des modalités définies par décret en Conseil d'État (article L. 221-2 du Code du Sport). Pour être inscrit sur liste ministérielle, le sportif doit répondre à plusieurs critères. Il doit pratiquer une activité reconnue de haut niveau, justifier d'un niveau sportif suffisant selon des conditions précises, avoir au moins douze ans et avoir fait l'objet d'une proposition en ce sens par sa fédération[80].

La liste établie par le ministère des Sports distingue quatre catégories : jeune (sportif ayant été sélectionné dans une équipe de France pour préparer les échéances internationales de sa catégorie d'âge), élite (sportif ayant réalisé une performance significative dans l'une des compétitions de référence ou dans des compétitions dont la liste est fixée par la commission nationale du sport de haut niveau (CNSHN)), senior (sportif ayant été sélectionné dans une équipe de France pour préparer les compétitions de référence) et reconversion (sportif ayant mis un terme à sa carrière sportive, présentant un projet d'insertion professionnelle et qui a été inscrit sur la liste des SHN dans la catégorie Elite ou sur une autre liste des SHN pendant au moins quatre ans dont au moins trois en catégorie Senior).

Au delà de cette liste principale, il existe deux listes complémentaires : celle des espoirs qui reconnaît les sportifs pour lesquels a été détecté un potentiel mais qui ne remplissent pas encore les conditions requises pour figurer sur la liste des sportifs de haut niveau et celle des partenaires d'entraînement, qui, elle, concerne les sportifs participant à la préparation des équipes de France. Le nombre de sportifs ainsi que les critères de mise en liste sont proposés par les directeurs techniques nationaux (DTN[81]) et arrêtés par le ministre chargé des sports après avis de la commission nationale du sport de haut niveau. Les critères et les quotas fixés pour

---

78 Association des archivistes français, *Abrégé d'archivistique, Principes et pratiques du métier d'archiviste*, Paris, 2020, p. 330.

79 On note en effet que la table ListeHistorique ne compte pas moins de 371 019 lignes.

80 Rapport d'information de l'Assemblée Nationale sur l'évaluation de la loi n°2015-1541 du 27 novembre 2015 visant à protéger les sportifs de haut niveau et professionnels et à sécuriser leur situation juridique et sociale, présenté par Maxime Minot et Bertrand Sorre.

81 Ministère de la ville, de la jeunesse et des sports - Inspection générale de la jeunesse et des sports, Evaluation du programme des aides personnalisées allouées aux sportifs de haut niveau, mars 2017.

chaque fédération sont valables quatre ans et modifiables au cours de l'olympiade. L'inscription des sportifs sur liste ministérielle est valable un an, deux ans pour ceux inscrits en catégorie Élite[82].

Dans le cadre de la politique du « double projet[83] » instaurée par le ministère des Sports en 2009, le sportif de haut niveau a accès à plusieurs aides : un soutien financier tout d'abord mais également des aides à la formation ainsi qu'à l'insertion professionnelle. En effet, un sportif inscrit sur liste ministérielle peut percevoir des aides sociales, des aides visant à soutenir leurs projets sportifs et de formation, des aides tendant à couvrir le manque à gagner sportif et/ou professionnel ainsi que des primes à la performance.
De plus, afin d'exploiter son potentiel sportif sans pour autant délaisser sa formation professionnelle, le SHN peut conserver dans la limite de cinq sessions des notes du baccalauréat général et technologique. Il peut également participer à la session de remplacement de septembre du même examen à condition que son absence à celle de juin soit justifiée par son directeur technique national. Enfin, au delà de l'accompagnement qui leur est proposé dans l'orientation et la recherche d'emploi, la convention d'aménagement d'emploi (CAE) dans le secteur public et la convention d'insertion professionnelle (CIP) dans le secteur privé permettent à un sportif de haut niveau titulaire d'un contrat de travail d'être mis à disposition de sa fédération afin de mener à bien ses projets sportifs tout en conservant la totalité de sa rémunération[84].

Plusieurs acteurs interviennent dans la formation et le suivi-socioprofessionnel du sportif de haut niveau. Ils sont divisibles en deux catégories : nationale et territoriale. Il nous a semblé important de les évoquer ici car ils sont susceptibles de donner des clefs de compréhension supplémentaires de la base de données qui nous a été confiée.
Tout d'abord, au niveau national, nous pouvons mentionner la Direction des Sports dont la mission première est de définir la politique de l'État en matière de sport. Elle possède surtout un rôle de pilotage puisqu'elle est en effet chargée d'organiser les dispositifs de suivi socio-professionnel des sportifs de haut niveau, d'établir les listes ministérielles des arbitres et juges sportifs de haut niveau, des sportifs Espoir et des Partenaires d'entraînement mais aussi d'instruire les demandes de reconnaissance du caractère de haut niveau des disciplines présentées par les fédérations sportives auprès du Comité national olympique et sportif français (CNOSF).
Le Comité national olympique et sportif français (CNOSF) représente le Comité international olympique en France. Il a pour mission d'incarner le sport français auprès des pouvoirs publics et des organismes officiels,

---

82 Rapport d'information de l'Assemblée Nationale sur l'évaluation de la loi n°2015-1541 du 27 novembre 2015 visant à protéger les sportifs de haut niveau et professionnels et à sécuriser leur situation juridique et sociale, présenté par Maxime Minot et Bertrand Sorre.

83 La politique du double projet vise à permettre à des SHN de concilier la recherche de la performance sportive et la réussite scolaire, universitaire et professionnelle. Considérant qu'une carrière sportive finit toujours par s'arrêter (retraite, blessure…), il s'agit de permettre au sportif de poursuivre sa scolarité tout en bénéficiant d'aménagements afin de lui permettre de s'entraîner. Le double projet s'appuie sur trois piliers : l'optimisation de la préparation sportive, l'encouragement à la formation et à l'insertion professionnelle, la préservation de leur santé et le respect de l'éthique.

84 Evaluation des dispositifs mis en place par les ministères chargés des sports et de l'éducation nationale visant à la formation des sportifs de talent, mai 2012.

de collaborer à la préparation et à la sélection des sportifs français et d'assurer leur participation aux jeux olympiques mais également de favoriser la promotion des sportifs sur le plan social[85].
Enfin, l'Institut national du sport, de l'expertise et de la performance (INSEP) a davantage une mission d'animation puisqu'il est chargé d'organiser des formations à l'intention des cadres du sport de haut niveau, de définir, en lien avec la Direction des sports, les axes de travail prioritaires dans l'ensemble des domaines touchant au sport de haut niveau, mais également d'animer le réseau[86].

Au niveau territorial, nous pouvons mentionner le rôle central des fédérations qui ont la responsabilité d'organiser les compétitions sportives à l'issue desquelles sont délivrés les titres internationaux, nationaux, régionaux et départementaux. Elles procèdent aux sélections correspondantes et proposent l'inscription sur les listes de sportifs, d'arbitres et juges sportifs de haut niveau, sur la liste Espoirs et sur celle des Partenaires d'entraînement. Elles veillent également à la bonne mise en oeuvre du double projet des sportifs et assurent leur surveillance médicale, le respect de l'éthique ainsi que la prévention du dopage[87].
Le directeur technique national, rattaché et rémunéré par le ministère des Sports est également placé sous l'autorité fonctionnelle du président de la fédération à laquelle il est affilié. Personnage central dans le suivi socio-professionnel du SHN, il est chargé de veiller à la bonne mise en oeuvre de la politique sportive de l'État au sein de sa fédération, s'occupe de la formation et du perfectionnement des cadres sportifs, techniques, entraîneurs fédéraux et animateurs et coordonne les actions entre sa fédération et les fédérations sportives associées. Il est surtout le moyen d'assurer la communication entre le Ministère des Sports et les fédérations[88].
Enfin, les pôles font partie des structures pouvant être retenues pour accueillir les sportifs de haut niveau. Ils appartiennent au parcours de l'excellence sportive à côté des équipes nationales, des groupes élites, des centres de formation des clubs professionnels ou encore des clubs…

Le suivi socio-professionnel du sportif de haut niveau est donc soumis au contrôle d'acteurs divers et nombreux, tant au niveau central qu'à l'échelon des collectivités territoriales[89]. Nous pouvons désormais nous demander quelle est l'utilité de cette base de données sur le sport de haut niveau. Placée sous la responsabilité du ministère des Sports, elle semble centraliser les informations relatives aux sportifs inscrits sur listes ministérielles ou l'ayant été. Outil privilégié d'échanges entre les différents membres du réseau, elle permettrait de coordonner les diverses actions mises en place à leur profit. Les renseignements contenus dans la documentation législative sur cet objet étant très succincts, un travail de rétro-ingénierie s'imposa.

---

85 Présentation du Comité national olympique et sportif français (CNOSF), site institutionnel de France Olympique, https://cnosf.franceolympique.com/cnosf/, (consulté le 24 juillet 2023).

86 Ministère des Sports, Instruction n°DS/DSA2/2011/341 du 22 août 2011 relative au réseau national du sport de haut niveau.

87 *Loc. cit.*

88 *Loc. cit.*

89 Evaluation des dispositifs mis en place par les ministères chargés des sports et de l'éducation nationale visant à la formation des sportifs de talent, mai 2012

1. 2 : Rétro-ingénierie de la base de données

La rétro-ingénierie désigne l'ensemble des opérations d'analyse d'un produit, d'un logiciel ou d'un matériel destinées à retrouver le processus de sa conception et de sa fabrication ainsi que les modalités de son fonctionnement. Lorsqu'un document papier ou numérique est livré à un service d'archives sans documentation associée, il appartient alors à l'archiviste d'essayer de la reconstituer. En effet, bien qu'un document puisse être archivé malgré l'absence de documentation, connaître le contexte et les motifs de sa création, l'entité créatrice, son fonctionnement ainsi que la communauté utilisatrice est très important pour permettre la meilleure exploitation possible des données contenues à l'intérieur. Or, comme nous l'avons expliqué précédemment, la base de données sur les sportifs de haut niveau fut seulement livrée avec un schéma de surcroît incomplet. Ce travail fut donc nécessaire. Nous pouvons le scinder en trois étapes distinctes : une première phase d'analyse de la structure des données, une deuxième de dépouillement d'archives papier et une troisième d'identification et de contact de personnes ressources.

Le 26 avril, avant même d'avoir pris connaissance de la base de données sur les sportifs de haut niveau, Anne Lambert, cheffe de la Mission affaires sociales nous fait parvenir une liste de documents papier. L'objectif est de sélectionner ceux qui, selon leur descriptif, pourraient nous en apprendre davantage sur la base. Nous sélectionnons plus d'une dizaine de cartons à consulter[90]. Le 2 mai, la base de données ainsi que le schéma associé sont copiés sur l'ordinateur mis à disposition par le SIAF pour toute la durée du stage. Après avoir observé le fichier Excel sur lequel elle figure, ses 110 tables, leurs intitulés ainsi que le type de données contenues, nous décidons rapidement de créer un fichier texte explicitant la structure de cette base. Sous forme d'un tableau, nous répertorions chaque table, explicitons en une phrase son contenu, son nombre de lignes, définissons chaque sigle et mettons en exergue ceux restés incompris. Chaque table dont le contenu ne semble pas évident est surlignée en vert. Ce document, très utile car centralisant au même endroit des informations importantes sur la structure de la base, s'est étoffé tout au long de la rétro-ingénierie. Document de travail, il n'a cependant pas vocation à accompagner la base.

Le 5 mai est programmée la première séance de dépouillement des archives papier. Nous nous focalisons sur un premier carton dont la cote est la suivante : B013027. La plupart des documents figurant dans ce carton sont des courriers ou des notes provenant du ministère des Sports et plus particulièrement de la sous-direction de la vie fédérale, du bureau de la vie de l'athlète ou bien du ministère de la Jeunesse et des Sports (Direction de l'administration générale ou Direction des Sports). Ils sont généralement destinés soit

90 L'envoi de ces documents le 26 avril s'explique par les délais de réception des cartons. La prise de connaissance de la base étant prévue pour le 2 mai, nous avons pensé qu'il pourrait être utile d'avoir déjà à disposition les cartons commandés.

au sein même de l'administration centrale[91] soit au réseau directement[92]. On y apprend notamment que dès le début des années 90, il est question de créer une base de données qui répertorierait les sportifs de haut niveau. En 1995, trois bases de données coexistent : celle de la Direction des sports qui gère la population des sportifs de haut niveau pour 54 sports, celle du Comité national olympique et sportif français qui centralise les résultats des sportifs aux grandes épreuves européennes et mondiales pour les 87 sportifs affiliés et enfin, celle du Sport d'Élite et de préparation olympique qui répertorie les résultats et le suivi des athlètes de haut niveau dans les 27 sports olympiques. A cette époque, l'objectif est donc de créer une base de données commune qui éviterait ainsi les saisies multiples.
En 2002, une base de données relative au sport de haut niveau en cours de test est mise en ligne sur l'intranet du ministère de la Jeunesse et des Sports. L'intention est double : il s'agit de donner accès au public aux données concernant les sportifs de haut niveau et surtout, de permettre aux différents partenaires du haut niveau de l'alimenter directement. Si ces éléments ne nous renseignent pas réellement sur la base de données qui nous intéresse, cela nous permet néanmoins de comprendre qu'elle semble résulter non pas d'une création réalisée à partir de rien mais plutôt de la reprise d'une ou plusieurs bases de données précédentes.

Nous nous concentrons ensuite sur les cartons suivants. La plupart des dossiers s'y trouvant sont en fait antérieurs aux années 90, aucune information intéressante pour la rétro-ingénierie n'y fut donc relevée. Bien que le premier carton nous ait été très utile, nous ne disposons pas de plus d'informations sur la base de données à proprement dite. Il fallut donc trouver une alternative aux archives papier.
C'est là que commence la troisième partie du travail de rétro-ingénierie : l'identification et la prise de contact avec des personnes ressources. Tout d'abord, nous contactons le 31 mai le chef de la mission des archives auprès du ministère des Sports mais cette piste ne fut pas concluante. Nous décidons alors de procéder autrement. Dans les archives, nous identifions une rédactrice régulière des courriers envoyés par le bureau de la vie de l'athlète du ministère des Sports. Bien que son nom soit mentionné dans des courriers datant pour les plus récents de 2003, nous tentons de contacter cette personne sur son adresse du ministère des Sports le 14 juin en lui expliquant notre problématique. Là encore, nous ne recevons pas de réponse. Plusieurs raisons peuvent l'expliquer : premièrement, les archives dans lesquelles elle est mentionnée étant datées de plus de vingt ans, sa situation a pu changer. De plus, un agent travaille sur de nombreux dossiers au cours de sa carrière, or, bien que pour nous, ce dossier ait une importance particulière, il n'en est peut-être qu'un parmi d'autres pour elle. Enfin, même si nous n'avons pas reçu de retour de mail automatique, il est également possible qu'elle ne consulte plus cette adresse ou qu'elle ait simplement manqué de temps.
Si les pistes non concluantes pourraient être considérées comme des échecs, nous avons pris le parti de les appréhender comme des moments d'avancement significatifs puisqu'elles permettent de refermer des portes et ainsi de restreindre le champ de recherche.

---

91 Courrier du Ministère de la Jeunesse et des Sports (Direction des sports - Sous-direction du sport de haut niveau et de la vie fédérale - Bureau de la vie de l'athlète), destiné au directeur du groupement d'intérêt public sport d'élite et préparation olympique (GIP-SEPO), 21 novembre 1994. Objet : Compte rendu de la réunion du 8 novembre 1994 relative au système d'informatisation et de suivi des sportifs de haut niveau.

92 Courrier du Ministère de la Jeunesse et des Sports (Sous-direction de la vie fédérale - Bureau de la vie de l'athlète) destiné aux directeurs techniques nationaux et aux présidents des fédérations sportives, non daté. Objet : Enquête sur la situation socio-professionnelle des sportifs de haut niveau.

Le 26 juin, nous organisons un échange avec l'historien Patrick Clastres afin de réfléchir aux recherches potentielles qui pourraient être menées sur la base de données des sportifs de haut niveau. Lors de cette discussion, ce dernier nous conseille de prendre contact avec Martine Gustin-Fall, présidente du Comité d'histoire des ministères chargés de la jeunesse et des sports. Nous la contactons et échangeons avec elle le 7 juillet. Bien qu'ayant connaissance de l'existence de cette base, elle nous oriente vers plusieurs personnes ressources susceptibles de nous renseigner précisément. Parmi les personnes conseillées, se trouvent deux anciens sportifs de haut niveau dont l'un travaille désormais au ministère des Sports et le second dans un CREPS ainsi qu'une ancienne directrice générale adjointe à la politique sport de l'INSEP. Contactés le même jour, nous sommes, une fois de plus, restés sans réponse.
Alors que nous pensions être arrivés au terme du travail de rétro-ingénierie, le 4 juillet, Chloé Moser, adjointe à la cheffe de la Mission affaires sociales retrouve à point nommé un dossier numérique collaboratif faisant mention de la base de données sur les sportifs de haut niveau. Y sont regroupés plusieurs documents : une fiche décrivant l'installation de l'application, le dossier d'architecture technique (DAT), un guide d'exploitation ainsi qu'un fichier expliquant la procédure de décommissionnement. Cela donne une dernière impulsion très bienvenue à l'étude. On apprend d'abord dans le dossier d'architecture technique datant du 29 avril 2016 qu'il y aurait trois moyens d'accéder à la base de données sur les sportifs de haut niveau : premièrement, l'application « Shn » qui permet d'accéder au portail sur les sportifs de haut niveau et comporte trois modules : ShnSportifs, ShnListe et ShnExploit. Elle est accessible via l'intranet du ministère des Sports. Puis, l'application « ShnPublic » qui, elle, est accessible sur internet via le site du même ministère et vise à fournir au grand public la liste des sportifs en liste. Enfin, le « Web Service ShnWS » permet aux fédérations, principalement par le biais des directeurs techniques nationaux, d'y entrer des données.
Les autres documents, très techniques ne nous ont malheureusement pas permis d'en apprendre davantage sur la base.

Bien que n'ayant pas permis de reconstituer une documentation complète permettant de retracer clairement et fidèlement le cycle de vie de cette base de données, ce travail a malgré tout porté ses fruits puisque la connaissance de la base a notablement progressé. Il nous a par ailleurs conduit à nous poser plusieurs questions : à partir de quand l'archiviste peut-il et/ou doit-il considérer la rétro-ingénierie comme étant terminée ? Peut-il se contenter d'une rétro-ingénierie incomplète ? L'absence d'informations ne peut-elle et/ou ne doit-elle pas être considérée comme une information en soit ?

Un travail de rétro-ingénierie est en réalité soumis à diverses contraintes : le temps et les réponses. En effet, la rétro-ingénierie d'un objet n'est souvent qu'une étape d'un travail de plus grande envergure. Dans le cas de la base de données sur les sportifs de haut niveau, nous avons décidé de borner la rétro-documentation à la durée du stage, à savoir quatre mois. De plus, il ne s'agissait pas là du coeur du stage puisque ce travail s'insérait dans une étude plus globale sur l'archivage des bases de données. Par ailleurs, la complétude de la rétro-ingénierie dépend des informations trouvées par le biais de la consultation de documents papier ou numériques mais également d'échanges humains. Or, il n'est pas possible de prévoir quel sera le taux de réussite de ces étapes déterminantes. On considèrera alors une fois la limite de temps imparti dépassée et tous les interlocuteurs potentiels identifiés contactés, que la rétro-ingénierie sera achevée.

Néanmoins, l'objet même sur lequel nous travaillons dans ce cas d'étude, c'est à dire une base de données, permet de pouvoir y revenir ultérieurement. En effet, la rétro-ingénierie peut se dérouler sur un temps long et être reprise à plusieurs moments en fonction des contraintes et avancées de l'archiviste.

Chapitre 5 : Expérimenter diverses solutions d'archivage

Nous l'avons vu, l'objectif principal de ce stage est de faire avancer le chantier sur l'archivage des bases de données. Afin de donner une dimension concrète à cette réflexion, nous avons décidé d'expérimenter deux méthodes d'archivage sur notre cas d'étude : l'archivage à plat et le format SIARD. L'expérimentation est une « méthode scientifique reposant sur l'expérience et l'observation contrôlée pour vérifier des hypothèses[93]. » Bien qu'étant traditionnellement associée aux mathématiques ou autres sciences « dures », elle peut également être appliquée à la sphère archivistique. Les visées sont multiples puisqu'il s'agit de voir, au moyen de tests réels, quels sont les coûts générés par chaque méthode en termes de temps et de complexité mais c'est également l'occasion d'expérimenter le format SIARD. Aussi, il s'agit d'identifier, pour ce cas précis, quelle méthode d'archivage semble la plus adaptée. Toutes les étapes mentionnées dans cette étude ne seraient pas nécessairement réalisées dans le cadre d'un réel travail d'archivage de base de données et sont à replacer dans une démarche d'essais.

Avant d'initier chacune des deux expérimentations, nous commençons par sacraliser l'extraction Excel fournie par la Mission affaires sociales nommée « BASE DE DONNÉES SHN ». En effet, il est préférable de préserver ce fichier et de travailler sur une copie que nous avons appelé « COPIE BASE DE DONNÉES SHN ».

2. 1 : L'archivage à plat

Nous pouvons distinguer deux étapes cruciales dans l'archivage à plat : l'export de la base de données au format CSV et le choix du séparateur.

Bien que la base de données nous ait été confiée sur un tableur Excel, nous avons préféré travailler avec LibreOffice, qui, contrairement à Excel, permet de personnaliser les critères de génération d'un fichier CSV à savoir, les champs englobants et le séparateur. En effet, sur Excel, les champs englobants sont automatiquement les guillemets. De même, le séparateur, à savoir « la séquence d'un ou plusieurs caractères servant à délimiter la frontière entre différentes régions de texte ou autres flux de données[94] » est toujours le point-virgule. Si la tabulation est également un séparateur proposé, le fichier final généré ne sera alors plus un .csv mais un .tsv.

Contrairement à Excel, Libre Office offre également la possibilité de modifier l'encodage du document (UTF-8, UTF-16 ou ASCII). Si la conversion d'une table en fichier CSV ne requiert pas de manipulations très techniques, une grande attention doit être portée au choix du séparateur. En effet, dans la mesure où il est assimilé à une colonne et doit permettre de remonter facilement la structure originelle de la base de données, le caractère choisi ne doit nullement figurer dans les données, auquel cas, le fichier CSV généré serait erroné. Il est donc indispensable de vérifier, en amont, que le caractère spécial choisi comme séparateur ne figure nul part dans les données. Pour cela, il fallut donc ouvrir chaque table, sélectionner toutes les données (CtrlA) et y rechercher (CtrlF) le caractère spécial envisagé comme séparateur. Lorsque toutes les données sont

---

93 Définition d'expérimentation, page Larousse, https://www.larousse.fr/dictionnaires/francais/expérimentation/32242, consultée le 15 août 2023.

94 Définition de séparateur, page Wikipédia, https://fr.wikipedia.org/wiki/Séparateur, consultée le 24 août 2023.

structurées de la même manière, trouver un séparateur peut être très rapide. Cependant, lorsque certaines tables contiennent des données entrées en saisie libre, cela peut s'avérer plus difficile, or, ce fut le cas pour la base de données sur les sportifs de haut niveau. Lors de cette étape de vérification, nous nous sommes rendus compte que la virgule, le point-virgule ainsi que la tabulation figuraient déjà dans les données. Nous avons donc testé d'autres caractères tels que le point d'interrogation, le point d'exclamation, le dollar et le pipe mais tous étaient déjà utilisés. Nous avons donc essayé un autre caractère spécial, le beta de l'alphabet grec (Alt 225) mais Libre Office l'interprétait comme la lettre « B », ce fut donc un nouvel échec.
À force d'éliminations, nous finissons par tester le symbole diviser (Alt 0247) qui s'avère n'être présent dans aucune table de la base de données. Cependant, ce caractère étant peu conventionnel pour un travail d'archivage, nous décidons de procéder autrement. En effet, bien qu'Excel propose automatiquement le point-virgule en guise de séparateur, il est également d'usage d'utiliser le pipe qui est rarement présent dans les données. Bien qu'il soit utilisé dans notre cas, nous décidons d'isoler les trois tables dans lesquelles nous l'avons trouvé, à savoir EnquetePole, ListeHisto et SuiviSocioProjSport. Nous exportons donc chacune des 107 tables ne contenant pas de pipe en CSV en indiquant bien le séparateur choisi et renommons chaque fichier du nom de la table exportée. Nous créons un dossier sur le bureau nommé « BDD SHN CSV » à l'intérieur duquel nous insérons deux dossiers : un premier nommé « CSV PIPE » dans lequel sont déposés tous les fichiers .csv générés avec un séparateur pipe et un second intitulé « CSV SEPARATEUR À TROUVER » où sont rassemblées les trois tables contenant au moins un pipe dans leurs données.
Pour ces dernières, deux cas de figure sont envisagés : choisir un séparateur différent pour les trois tables ou bien choisir un même séparateur pour chacune d'elles. Nous avons préféré cette dernière solution et donc opté pour faire du symbole diviser le séparateur de ces trois tables. Nous avions également envisagé de remplacer chaque pipe contenu dans les tables par le symbole diviser et ainsi, pouvoir faire du pipe leur séparateur en documentant précisément les modifications effectuées. Néanmoins, considérant que cette méthode aurait pu altérer les principes mêmes d'authenticité et d'intégrité fondamentaux pour l'archiviste, cette possibilité fut abandonnée.

Une fois les fichiers CSV créés, il est préférable de les ouvrir sur un éditeur de texte tel que Notepad ++ ou Atom afin de vérifier qu'il n'y ait pas eu d'erreur lors de l'export. Après avoir procédé aux vérifications nécessaires, l'expérimentation de l'archivage à plat peut être considérée comme achevée.
Il est important de compter environ cinq minutes pour l'export de chaque table. Pour les 110 tables de la base de données sur les sportifs de haut niveau, l'usage de cette méthode a nécessité deux jours de travail. Bien que peu coûteux en termes de temps et de technicité, l'archivage à plat nécessite néanmoins de porter une grande attention au séparateur puisque l'utilisation d'un caractère présent dans les données pourrait conduire à la sauvegarde d'un document erroné.

2. 2 : L'archivage au format SIARD

Lors des différents échanges organisés avec le programme Vitam et avec les Archives Nationales, nous avons pu évoquer le format SIARD. Ainsi, si pour Marion Ville qui a pu l'expérimenter sur la matrice cadastrale, ce format constitue un pivot pour l'archivage des bases de données relationnelles, les Archives Nationales ne s'y sont pas engagées. En effet, selon l'institution, la volumétrie des données transposées en

SIARD[95] ainsi que la dépendance du format au logiciel SIARDSuite sont des freins majeurs à l'utilisation de cette méthode dans une visée de pérennisation. Par ailleurs, pour les Archives Nationales, en cas de bug informatique, le travail de rétro-conversion depuis un fichier SIARD serait trop technique et coûteux.
Le 21 juin, nous définissons la stratégie à suivre pour les expérimentations du format SIARD. Dès le début, un constat est effectué : bien que la base de données sur laquelle nous travaillons, collectée sous forme d'extraction, ne permette pas un usage habituel de SIARD, l'intérêt est ici de manipuler les outils et d'en décrire les mécanismes. Pour cela, deux scénarios sont envisagés : convertir toutes les tables de la base de données sur les sportifs de haut niveau comme nous l'avons fait pour tester l'archivage à plat ou bien n'exporter qu'une seule table et voir quels en sont les résultats. Même si l'expérimentation de SIARD s'annonce plus technique que la précédente, nous convenons que l'export d'une seule table reviendrait à exploiter le format uniquement sous l'angle de la consultation. Or, comme nous l'avons évoqué plus haut, il subsiste en France une véritable interrogation quant à l'intérêt de préférer le format SIARD à une autre méthode d'archivage. Nous avons donc opté pour la première option : convertir chacune des 110 tables.

La base de données sur les sportifs de haut niveau nous ayant été fournie sous forme d'une extraction Excel, nous ne pouvions pas utiliser la connexion des SGBD permise par SIARD. Dans la mesure où les conditions préalables d'utilisation du format n'étaient pas remplies, nous nous sommes demandé comment le tester. S'il n'est pas possible de convertir un fichier Excel en SIARD, il est en revanche faisable de le faire depuis un document .csv. Deux outils développés par le KOST-CECO permettent de réaliser cette opération : csv2siard, logiciel libre et FullConvert qui est, quant à lui, un logiciel propriétaire. Nous choisissons donc d'utiliser csv2siard. La documentation est consultable sur le site du KOST-CECO mais n'est disponible qu'en allemand[96]. Nous avons donc dû procéder à une traduction en français. Par ailleurs, les captures d'écran illustrant les instructions sont relativement datées et ne sont plus en concordance avec les interfaces actuelles. Le logiciel est téléchargeable sur le même site[97]. Parallèlement, nous décidons d'installer SiardSuite, disponible gratuitement sur le site des Archives fédérales suisses.

Afin de réutiliser les fichiers CSV créés sans compromettre le travail réalisé lors de l'expérimentation de l'archivage à plat, nous réalisons une copie du dossier « BDD SHN CSV » que nous renommons « BDD SHN CSV SIARD ». Nous conservons la structure du dossier créée pour l'expérimentation de l'archivage à plat afin de garder isolées les tables ayant un autre séparateur que le pipe. Pour convertir un fichier CSV en SIARD, le déroulé est le suivant : il s'agit d'ouvrir le dossier csv2siard et de placer le ou les fichiers à convertir dans le dossier intitulé « datatype ». Ensuite, il est nécessaire de se rendre dans le fichier « preferences », d'indiquer le séparateur ainsi que l'encodage choisis et de préciser la

---

95 Le format SIARD s'appuie sur le langage de balisage XM, or, les problèmes de volumétrie inhérents au XML ont été soulignés par le Référentiel Général d'Interopérabilité (RGI) qui suggère donc d'éviter d'y recourir pour l'archivage de bases de données volumineuses. https://www.numerique.gouv.fr/uploads/Referentiel_General_Interoperabilite_V2.pdf, consulté le 20 août 2023.

96 Documentation csv2siard, https://kost-ceco.ch/cms/dl/5021cd25f704f4a646e88050de880e28/Anwendungshandbuch_v1.8.pdf?target=1, consulté le 14 août 2023.

97 Page SiardSuite sur le site des Archives fédérales suisses, https://www.bar.admin.ch/bar/fr/home/archivage/outils-et-instruments/siard-suite.html, consulté le 14 août 2023.
Lien Github d'installation de SiardSuite : https://github.com/sfa-siard/SiardGui/releases, consulté le 14 août 2023.

présence ou non de champs englobants. Pour commencer, nous préférons déposer une table unique dans « datatype » et nous indiquons dans les préférences les mêmes critères que ceux établis pour l'archivage à plat : le séparateur est un pipe, l'encodage est en UTF-8 et il n'y a pas de champs englobants. Il suffit ensuite d'ouvrir le logiciel que l'on trouve dans le même dossier zippé sous le nom de « c2sGUI.exe » et de convertir le fichier. Seulement, là, nous rencontrons une première erreur : csv2siard indique que le fichier .csv n'est pas encodé en UTF-8 contrairement à ce qui a été exprimé dans les préférences. Nous décidons donc de générer un nouveau fichier .csv de la table concernée, de vérifier l'encodage et de le replacer immédiatement dans csv2siard. Après avoir eu le même message d'erreur sur toutes les autres tables testées, nous transformons de nouveau chaque table en fichier .csv en prêtant grande attention à l'encodage choisi. Une fois cela fait, il suffit de charger la table dans SiardSuite et la structure est remontée.
Sur les 110 tables, 82 fonctionnent et 28 signalent une erreur. Nous pouvons distinguer deux types d'erreurs : d'abord, si csv2siard semble accepter facilement le séparateur pipe, il n'en est pas de même pour le dollar et le symbole diviser, les trois tables concernées ne sont donc pas converties en fichiers SIARD. Aussi, le logiciel de conversion rencontre des difficultés d'interprétation de certaines colonnes vides.

Le 7 juillet, soit trois semaines avant la fin du stage, nous décidons de reprendre contact avec l'institution la plus à même de nous aider à régler les difficultés rencontrées ; les Archives fédérales suisses. Un échange avec Audun Lund est organisé le 20 juillet. Celui-ci répond à un double objectif : s'il s'agit d'abord de lui exposer les problèmes rencontrés pour la conversion de certains fichiers CSV dans le logiciel csv2siard, c'est également l'occasion de lui faire part de nos premières constatations quant à l'utilisation de ce logiciel de conversion. En effet, le fait que la documentation et l'interface de cet outil soient en allemand est un réel frein à une utilisation efficace. L'existence, la mise en ligne ou la réalisation d'une traduction en anglais serait selon nous, une suggestion pertinente.
De cet échange d'une heure résultent les conclusions suivantes : les manipulations effectuées tendant à exporter chaque table originellement issue d'un document Excel en fichier CSV pour les convertir en fichier SIARD via csv2siard n'est, selon Audun Lund pas « la bonne » manière de procéder. En réalité, une base des données livrée sous forme de tableur Excel n'est pas transposable en SIARD puisque par définition, des données structurées mises à plat dans un fichier CSV perdent les liens qu'elles entretiennent entre elles. Or, comme nous l'avons exposé précédemment, l'intérêt du format suisse est de remonter une base de données en conservant les relations qui s'y trouvent.
Passer du CSV au SIARD est donc réalisable comme nous l'avons expérimenté, néanmoins, cela n'a pas de valeur ajoutée puisque cela perd tous les intérêts du format suisse. Mais alors, comment expliquer l'existence d'outils tels que csv2siard et FullConvert ? Selon l'archiviste suisse, csv2siard est un outil ancien, créé de manière expérimentale lors des réflexions sur le format mais inutilisable en réalité. L'inexistence d'une traduction anglaise et la désuétude des captures d'écran utilisées pour illustrer les instructions s'expliquent donc par l'obsolescence même de l'outil qui doit être prochainement retiré du site du KOST-CECO.

Bien que notre démarche se soit donc avérée être une impasse, nous nous interrogions également sur le temps nécessaire pour l'export des tables au format SIARD qui nous semblait particulièrement conséquent. Or, après avoir assisté à une démonstration réalisée par Audun Lund lors de la réunion, nous comprenons que si les conditions d'export en SIARD sont réunies, c'est-à-dire que si la base de données à

archiver est stockée sur l'un des systèmes de gestion de bases de données compatibles avec SIARD, alors il ne suffit que de quelques minutes pour réaliser la conversion. Cet échange nous a permis de comprendre que finalement, l'intérêt de SIARD résidait dans le fait de pouvoir connecter une base de données issue d'un SGBD particulier à un autre. Cette connexion nécessite des informations telles que le nom du serveur sur lequel est hébergée la base de données, le nom de la base de données, le nom d'utilisateur du service producteur et son mot de passe de connexion, exclusivement récupérables auprès du service producteur. L'utilisation du format nécessite donc un travail commun entre service producteur et service d'archives réalisé en amont du décommissionnement de la base concernée.

Finalement, cela nous a permis de conclure définitivement les expérimentations sur le format SIARD et surtout, d'en comprendre concrètement tant les atouts que le fonctionnement. Cela nous permet également de répondre à la problématique récurrente de la non utilisation de SIARD en France. En effet, les pays européens recourant à ce format travaillent sur des bases de données hébergées originellement sur des systèmes de bases de données compatibles avec SIARD. Or, étant donné qu'en France, la méthodologie la plus courante consiste à extraire une base de données pour ensuite l'archiver sous forme de fichiers CSV, le recours à SIARD est invalidé. La base de données sur les sportifs de haut niveau nous ayant été confiée sous forme d'extraction Excel, les conditions nécessaires pour la mise en oeuvre du format SIARD ne sont donc pas réunies.

Chapitre 6 : Conclusions des expérimentations

3. 1 : Mettre en perspective le travail du questionnaire et les expérimentations

Nous l'avons expliqué précédemment, l'objectif de ces expérimentations était double : il s'agissait à la fois de réaliser, comme un archiviste, toutes les étapes nécessaires à l'archivage d'une base de données et ainsi, évaluer quelle méthode semblait la plus adaptée à notre cas d'usage mais également de faire avancer la réflexion sur le format SIARD. Pour répondre à cette double visée, nous avons parallèlement à ce travail, réalisé un questionnaire sur les pratiques en termes d'archivage de bases de données collectées localement[98]. Celui-ci se divise en quatre parties dont seulement trois nous intéresseront ici : la collecte (A quelle fréquence le service d'archives collecte-t-il des bases de données? Quel est le type de bases de données collectées ? Quelle solution d'archivage est mise en place?), le format SIARD (Le service d'archives a-t-il connaissance de l'existence de ce format ? L'utilise-t-il pour la conservation des bases de données[99] ?) et le descriptif de quelques bases de données collectées (nom, type de données, modalités d'archivage). L'intérêt de mener conjointement ces deux études était de savoir s'il était possible d'établir une forme de cohérence entre les résultats des expérimentations et les réponses au questionnaire.

De cette réflexion, nous pouvons tirer les conclusions suivantes : l'archivage à plat est la méthode la plus couramment utilisée pour l'archivage de bases de données en France, tant par les institutions centrales que territoriales. En effet, sur les 18 services d'archives ayant répondu au questionnaire sur les pratiques en termes d'archivage de bases de données collectées localement, 11 ont expliqué avoir systématiquement recours à l'archivage à plat pour conserver des données structurées. Bien que la collecte de bases de données relève de l'exceptionnel pour la plupart des services répondants, cette réponse est significative des pratiques territoriales et montre que c'est une méthode dont l'usage est répandu et dont on sait qu'elle fonctionne. Sur les 11 services recourant à l'archivage à plat, 7 ont évoqué des exemples de bases de données collectées. Celles-ci, au nombre de 19, sont très intéressantes puisqu'elles nous permettent de mener une étude comparative.

D'abord, si l'archivage à plat a originellement été associé à la collecte de données structurées statistiques, on remarque que sur les 19, seules trois contiennent ce types de données. Les autres renferment des données dites « individuelles », c'est à dire concernant des personnes ou bien « d'autres données ». C'est donc une méthode qui semble se prêter à tous types de données structurées.

Par ailleurs, on constate que toutes les bases de données mentionnées pèsent entre 0 et 1 Go et peuvent contenir un nombre de tables très hétérogène allant de 2 à 2665. La base de données sur les sportifs de haut niveau entre donc complètement dans ce cadre puisqu'elle est composée de 110 tables et renferme des données tant statistiques qu'individuelles. Si exporter 110 tables une à une au format CSV nous a semblé réalisable lors de l'expérimentation, nous pourrions émettre des doutes quant à la pertinence de réaliser le

---

98 Voir annexe n° 2 : Questionnaire sur les pratiques des services d'archives en terme de collecte, d'archivage et de mise à disposition des bases de données locales.
Voir annexe n° 3 : Résultats du questionnaire.

99 Hors matrice cadastrale.

même travail pour plus de 2000 tables. Néanmoins, il importe de mentionner qu'une méthode existe afin d'effectuer l'export d'une base de données en une seule fois mais nous n'avons pas eu suffisamment de temps pour étudier davantage cette possibilité.
Aussi, il semble que la présence ou non de documentation pour accompagner la base ne gêne pas l'archivage à plat. Effectivement, sur les 19 bases de données mentionnées, seules 11 ont été collectées avec de la documentation. Celle-ci peut être exclusivement papier ou numérique, c'est le cas pour 9 d'entre elles mais peut également être conservée sur les deux supports, comme nous pouvons le constater pour l'une des bases de données. Là encore, le fait que la base de données sur les sportifs de haut niveau n'ait pas été accompagnée d'une documentation aboutie ne semble donc pas être un frein à sa conservation.
Enfin, l'archivage à plat, en plus d'être reconnu par la communauté archivistique semble l'être des chercheurs. En effet, en 2018, les Archives Nationales organisent un atelier intitulé « Conserver et accéder aux données structurées » avec des archivistes, des statisticiens et des historiens[100]. Au cours de cet échange, Claire Lemercier, historienne, directrice de recherche au CNRS et spécialiste d'histoire quantitative évoque que le CSV est un format acceptable pour les usagers non experts. Etienne Ollion, sociologue chargé de recherches au CNRS met quant à lui en exergue l'importance pour le chercheur de travailler sur des données qui soient les plus brutes possibles, c'est à dire les moins retravaillées. En cela, le CSV est un bon format. Néanmoins, il met également en garde sur le fait que ce n'est pas une méthodologie adaptée à l'archivage de bases de données très lourdes, comme le XML et le SIARD. Enfin, Bo Yun Park, doctorante en sociologie politique comparée de l'université de Harvard, met en évidence le fait qu'il est possible de rendre les données brutes plus lisibles en ayant recours à des logiciels tels que R, Stata ou encore MathLab qui permettent, à partir d'un fichier type CSV, de remonter la structure d'une base de données.

L'analyse des réponses au questionnaire fut très intéressante pour notre réflexion sur l'archivage à plat puisque cela a permis de confirmer les constats que nous avions pu tirer lors des diverses lectures effectuées et des expérimentations. Au delà de l'archivage à plat, l'analyse des résultats du questionnaire nous a également apporté divers éléments de réponse relatifs au format SIARD.
Deux questions étaient posées sur le sujet : Le service d'archives connaît-il le format SIARD ? L'a-t-il déjà utilisé pour l'archivage d'une base de données[101] ? À la lecture des résultats, nous apprenons tout d'abord que sur les 18 services répondants, seulement un ne connaît pas le format suisse. C'est un élément intéressant qui confirme que SIARD n'est pas une solution d'archivage de bases de données marginale et peu répandue mais bien une alternative majeure connue des archivistes issus d'institutions tant centrales que territoriales. Sur les 17 services ayant connaissance du format, 7 affirment l'avoir déjà utilisé pour archiver des bases de données. Bien que nous ayons signalé dans le questionnaire que nous souhaitions exclure de la réflexion l'archivage de la matrice cadastrale dont on sait qu'il est effectué au format SIARD, il semble que les réponses recueillies n'en aient pas pris compte. En effet, sur les 7 services ayant répondu « oui », trois ont mentionné des exemples de bases de données archivées au format SIARD mais il s'agit à chaque fois de la matrice cadastrale. Il n'est donc mentionné aucun exemple d'autre base de données conservée dans ce

100 Atelier « Conserver et accéder aux données structurées », dont le compte rendu fut rédigé par Martine Sin Blima-Barru le 31 août 2018 et diffusé le 21 septembre 2018.

101 Hors matrice cadastrale.

format. Il pourrait être pertinent de se demander pourquoi cette méthode, connue de la plupart des services n'est pas utilisée : est-ce par manque de connaissances sur le sujet ? Est-ce parce qu'il est considéré comme moins pérenne que l'archivage à plat ou bien est-ce parce que les bases de données concernées ne répondent pas aux critères nécessaires à la mise en application du format ? Il aurait pu être intéressant de demander au réseau d'archives si une formation sur ce format les aurait attirés.
Quoi qu'il en soit, les réponses au questionnaire confirment finalement qu'il était nécessaire d'expérimenter le format SIARD dans une perspective d'apport de connaissances. Mettre en avant les pré-requis indispensables à la mise en pratique du format, les difficultés rencontrées et les intérêts constatés sont autant de clefs permettant de comprendre pourquoi, bien que connu des services d'archives, il n'est à ce jour que très peu utilisé par les archivistes français[102].

3. 2 : Choix d'archivage final

Bien que l'archivage à plat soit en France la méthode la plus utilisée pour la conservation des données structurées, l'objectif en partant d'un cas d'étude précis était de confronter cette méthode à une autre largement pratiquée internationalement et voir laquelle semblait ici être la plus adaptée. En effet, tandis qu'au niveau international, on collecte les bases de données à la source, en France, l'extraction est une pratique systématique.
La situation initiale était la suivante : nous disposions d'une base de données composée de 110 tables reliées les unes aux autres et livrée sous forme d'une extraction Excel. Décommissionnée en 2021 par le service producteur, elle n'était accompagnée que d'une documentation très partielle se résumant à un schéma incomplet de la structure de ladite base. Renfermant des données précieuses et lisibles sur la condition des sportifs de haut niveau telles que leur formation professionnelle, leur palmarès, leur situation personnelle ou encore leur parcours scolaire, son intérêt archivistique était indéniable. Pour la conserver, plusieurs possibilités : archiver les données brutes au format CSV, l'archiver au format SIARD ou réaliser un double archivage incluant les deux méthodes.
Nous pouvons tirer de l'expérimentation de l'archivage à plat les conclusions suivantes : c'est une méthode peu coûteuse en termes de temps dont on sait qu'elle fonctionnera à long terme. Elle est par ailleurs appréciée des chercheurs puisqu'elle ne nécessite pas d'intervention directe sur les données. Bien que tendant à rompre les liens entre les tables, la documentation jointe au fichier CSV permet de reconstituer aisément la structure de la base dans un tableur Excel par exemple. Un point d'attention est à souligner néanmoins : le choix du séparateur. Quel qu'il soit, ce choix doit être documenté afin de permettre une conservation et une réutilisation pérennes. Dans notre cas d'usage, le choix du séparateur fut la seule problématique rencontrée mais fut résolue aisément.

L'expérimentation du format SIARD a quant à elle posé davantage de problématiques. En effet, la mise en application de cette méthode d'archivage sous-entend les conditions suivantes : la base de données doit être contenue sur un système de gestion de bases de données accepté par la suite logicielle SIARDSuite

[102] Nous pouvons bien sûr mentionner l'archivage de la matrice cadastrale au format SIARD, néanmoins, dans ce cas précis, les fichiers SIARD sont conçus avant les versements. Cette stratégie est donc différente de celle consistant à réaliser le fichier SIARD soi-même.

dont le rôle est de faire fonctionner le format. Aussi, un travail entre service producteur et service d'archives en amont de son décommissionnement est nécessaire afin de recueillir des informations indispensables à la connexion entre les différents systèmes de gestion de bases de données. Bien que très utile pour la conservation et surtout la réutilisation d'une base de données relationnelle comme celle sur les sportifs de haut niveau, la mise en oeuvre du format n'a pas abouti puisque dans la mesure où celle-ci était conservée sous forme d'export Excel. Si nous pourrions considérer cela comme un échec, nous avons décidé de le voir différemment. En effet, l'intérêt de ce travail était de faire progresser la réflexion sur le sujet. Or, initialement, subsistait une réelle incompréhension quant au paradoxe suscité par la large utilisation du format à l'étranger et le retrait de la France, pourtant très impliquée dans travaux sur la conservation de ce type de données. À l'issue des expérimentations et des différents échanges menés avec les Archives fédérales suisses, nous pouvons conclure que si le format SIARD n'est actuellement pas utilisé en France, c'est notamment parce que les conditions initiales de stockage des bases de données ne sont pas les mêmes en France que dans les pays utilisant SIARD. L'archivage dans ce format se prépare bien en amont d'un décommissionnement puisqu'il nécessite de remplir des conditions particulières.

En l'état du contexte français, si nous devions réellement archiver la base de données sur les sportifs de haut niveau, la méthode la plus appropriée serait donc la première, à savoir l'archivage à plat au format CSV.

TROISIÈME PARTIE - RÉFLEXIONS AUTOUR DE LA BASE DE DONNÉES EN TANT QU'OBJET PATRIMONIAL

Cette ultime partie, sans doute moins factuelle et plus réflexive que les deux précédentes résulte des nombreux échanges organisés tout au long de ce stage. En effet, il était initialement prévu de se concentrer uniquement sur l'état de l'art actualisé de l'archivage des bases de données et le cas d'étude sur la base de données sur les sportifs de haut niveau. Seulement, les premiers recueils d'expériences d'archivistes ont fait émerger diverses notions situées au coeur de leur métier telles que l'usage, l'accessibilité ou encore l'intention de préservation. Or, si nous nous étions jusqu'à présent concentrés sur les modalités de conservation car elles représentent le coeur du métier d'archiviste, nous n'avions pas traité la mise à disposition de la source pour l'usager. C'est donc davantage sur cet aspect que s'articule cette dernière partie, synthèse d'échanges entre deux métiers distincts mais corrélés : archivistes et chercheurs.

Chapitre 7 : Réflexions autour de la notion de source

1. 1 : Appréhender les bases de données sous l'angle de la source

A l'heure où les données croissent de manière exponentielle, où existe l'idée qu'il suffit de cliquer pour obtenir l'information souhaitée, évidemment pertinente car organisée dans ces puissants réservoirs de connaissances que sont les bases de données, quelle place peut-il y avoir pour une réflexion sur les sources ? A quels enjeux singuliers l'archiviste doit-il répondre ? Une base de données, entendue comme un ensemble de données structurées regroupées dans un conteneur logiciel est un objet complexe. Si cette structuration permet de représenter lisiblement des informations relatives à une activité humaine, elle constitue un véritable défi pour l'archiviste devant conserver des données originellement conçues pour être maintenues dans une structure à la fois particulière et unique.

1. 2 : Distinguer l'objet intellectuel de l'objet technique

Nous l'avons relevé précédemment, il est relativement courant de considérer l'archivage d'objets nativement numériques tels que les photographies, les documents bureautiques ou les bases de données comme un ensemble. Cependant, chaque catégorie pose en réalité des contraintes de conservation et des enjeux d'accessibilité différents. De plus, derrière l'expression « base de données », se cachent différents types d'objets. En effet, il est possible de distinguer au sein de la classification traditionnelle (hiérarchique, relationnelle, graphe), deux types des bases de données. Parmi elles, certaines, de forme élémentaires et gérées en direct par un ou plusieurs utilisateurs sont directement appréhendables. D'autres, plus techniques, sont quant à elles alimentées par le biais d'une interface, ce qui les rend plus abstraites et opaques. La différence notable entre ces deux types de bases réside dans le fait que dans le premier cas, la base de données a été conçue par les utilisateurs eux-mêmes dans le but d'avoir un objet achevé lisible et réutilisable. Lors du décommissionnement, l'archiviste devra donc centrer sa réflexion autour de cet objet métier produit par l'utilisateur primaire. A contrario, pour les bases de données cachées derrière une interface ou une application et qui ne sont donc pas visibles directement par l'usager, les enjeux diffèrent. Les données, qu'elles soient représentées par le biais de l'interface ou agrégées à une base de données cachée derrière, seront conservées, mais comment ? Faut-il préserver le site qui les hébergent dans le cadre d'une application,

conserver les boîtes de messagerie sous forme de fichiers .pst ou bien faut-il archiver ces réservoirs de données sous forme de fichiers à plats ?

Identifier la nature de la base de données à laquelle il est confronté est une étape très importante du travail de l'archiviste puisque c'est notamment de cette constatation que pourra ensuite découler le choix du type d'archivage le plus adapté. Ce fut d'ailleurs l'une de nos interrogations lors du travail sur la base de données sur les sportifs de haut niveau. Bien que nous ne disposions uniquement d'une extraction Excel, nous nous sommes demandés si la base était visible directement par les usagers ou bien si elle était dissimulée derrière une application. En cela, le travail de rétro-ingénierie fut très instructif. En effet, bien que nous n'ayons récolté que très peu d'informations sur la base de données sur les sportifs de haut niveau (2011-2021), nous avons pu comprendre grâce aux dossiers papiers consultés, quelques bribes de son fonctionnement. Tout d'abord, nous savons que cette base de données était remplie par divers acteurs incluant au moins les directeurs techniques nationaux et le ministère des Sports. Aussi, cette base de données contenant de nombreuses informations confidentielles sur les personnes, il est peu probable que tous les utilisateurs y aient eu accès en totalité. Surtout, nous avons trouvé dans les documents papier consultés, plusieurs copies d'écran révélant l'interface par laquelle la base de données antérieure à celle de 2011 était accessible. Néanmoins, dans la mesure où il s'agit d'agréger en un seul et unique document toutes les informations nécessaires au suivi des sportifs de haut niveau, cette base de données peut être considérée comme un objet intellectuel. Cela nous permet alors d'énoncer le constat suivant : l'identification de la nature de la base et les usages qui en sont faits sont étroitement liés.

Chapitre 8 : Placer l'usage au centre des enjeux de l'archivage : corrélation entre passé et futur

2. 1 : Identifier les usages passés pour anticiper les usages futurs

Afin d'établir la méthode d'archivage la plus adaptée, il est important d'identifier le ou les usages originels faits de l'objet à archiver. Effectivement, pour répondre à la question « Que faut-il conserver ? », il est important de connaître tant les raisons ayant motivé la création de l'objet que son utilité. Le passé renseigne le futur et c'est en cela que lorsqu'une base de données est livrée sans documentation, une rétro-ingénierie s'impose. En ce sens, le thème de la journée d'études doctorales du centre Jean Mabillon de l'École nationale des chartes, diffusé en mai dernier, est venu alimenter notre réflexion. Cette journée, qui s'articulera autour du titre suivant : « Source, poison ou accident : comprendre le document dans les sciences historiques », propose de réfléchir au rapport que l'historien entretient avec ses sources[103]. Comme l'appel à propositions l'évoque, « revenir à la source permet de construire une méthode scientifique à partir du contexte de production du document pour changer le paradigme de compréhension de ce matériau. » Or, c'est justement en cela qu'en l'absence de documentation, la rétro-ingénierie revêt un intérêt considérable : elle permet de comprendre le(s) usage(s) originel(s) fait(s) de l'objet à conserver et ainsi, pouvoir imaginer et préparer les utilisations futures.

Afin d'illustrer ce propos, nous pouvons mentionner l'exemple évoqué par Erwann Ramondenc lors de l'échange organisé le 25 mai. Dans le cadre de ses anciennes fonctions au ministère de l'Europe et des affaires étrangères, ce dernier fut confronté aux problématiques suscitées par l'archivage de l'application Diplomatie. Diplomatie est une application de correspondance derrière laquelle se trouve une base de données dans laquelle s'agrègent les données échangées par le biais de ladite application. S'est donc posée l'interrogation suivante : Que faut-il conserver ? L'une des clefs de réponse est selon Erwann Ramondenc d'identifier la manière dont l'utilisateur primaire use de l'application. Une fois l'usage identifié, c'est-à-dire, dans ce cas précis, pour correspondre avec d'autres usagers, il est possible d'anticiper le type d'utilisateur intéressé par ledit objet ainsi que les utilisations futures et donc les motifs de consultations qui seront évoqués. Dans la mesure où, ici, l'objet métier est la correspondance diplomatique, l'on peut penser que les futurs utilisateurs intéressés seront notamment des historiens ou des particuliers. De plus, l'objet métier étant la correspondance et non la base de données stockant les données brutes, il est plus probable que l'utilisateur futur veuille consulter la correspondance, plus représentative et explicite des usages passés.

Néanmoins, dans une perspective de complétude et d'anticipation d'un maximum de besoins, il aurait également pu être envisageable de conserver la base de données dans la mesure où un statisticien par exemple, ne serait pas en quête du même type d'informations qu'un historien. Selon lui, la finalité du travail de l'archiviste est avant tout de préserver la donnée initialement produite : ici, des correspondances. L'archivage de la base de données, dans la mesure où elle n'a pas été explicitement produite et remplie par

---

103 « Source, poison ou accident : comprendre le document dans les sciences historiques », appel à propositions pour la journée d'études doctorales proposée par le centre Jean Mabillon, École nationale des chartes prévue le 19 octobre 2023.

l'usager, ne doit donc être considéré que comme un travail optionnel et supplémentaire en fonction de sa pertinence, qui, elle aussi, doit être réfléchie au regard de chaque nouveau cas.

Au cours de cet échange, Erwann Ramondenc nous fait part d'un autre exemple, celui de la base de données Registre répertoriant les français détenus à l'étranger. Le fonctionnement de l'application est le suivant : l'utilisateur réalise une requête par le biais d'une interface et obtient la ou les fiches demandées. À la question : « Que conserver ? », il semblait donc assez évident de répondre : les fiches des détenus. Cependant, il était techniquement plus simple de fournir une extraction de la base de données générée derrière l'application en fichiers CSV. Dans ce cas de figure, il aurait été intéressant selon le directeur des Archives de la Nièvre, de conserver à la fois les fiches mais également la base de données. En effet, dans la mesure où les fiches sont représentatives de ce que l'usager primaire voyait et de ses besoins originels et parce que ce format permettrait à l'usager futur d'avoir à disposition seulement les informations dont il a besoin, conserver les fiches semble répondre à la plupart des besoins futurs. Pour un utilisateur futur tel qu'un chercheur en sciences humaines ou un particulier faisant une recherche sur un ou plusieurs individus précis, il n'apparaît pas nécessaire voire même souhaitable d'avoir accès à l'entièreté de la base. En revanche, des statisticiens pourraient être bien davantage intéressés pour travailler sur les données brutes plutôt que sur une série de fichiers PDF. Cet échange pose donc finalement la problématique suivante, au coeur du métier de l'archiviste : comment inclure dans l'archivage la conservation des usages passés et l'anticipation des besoins futurs ? Comment les notions d'usage et d'accessibilité influencent-elles les stratégies d'archivage ?

2. 2 : Enquête autour des usages futurs possibles de la base de données sur les sportifs de haut niveau

La sous-partie suivante propose certes de réfléchir aux utilisations futures potentielles de la base de données sur les sportifs de haut niveau mais s'inscrit également dans une réflexion plus large sur les enjeux liés à l'archivage. Nous l'avons étudié, identifier les usages originels faits de l'objet à archiver est déterminant pour élaborer une stratégie d'archivage adaptée. Mais est-ce le seul critère à prendre en considération ? Le travail de l'archiviste consiste en la conservation et la mise à disposition d'une information produite dans le passé et s'insérant dans un contexte et une visée précis. Ainsi, selon nous, l'enjeu de l'archivage est double : il s'agit de comprendre le(s) usage(s) originel(s) fait(s) de cet objet pour imaginer et préparer les utilisations futures. C'est donc dans cette perspective qu'il nous a semblé important de réfléchir aux futures utilisations et recherches possibles sur la base de données sur les sportifs de haut niveau.

Pour cela, il fallut dans un premier temps identifier le(s) public(s) potentiellement intéressé(s) par cet objet et susceptible(s) de nous renseigner. La base de données contenant des informations à la fois riches et intelligibles relatives aux aides octroyées aux sportifs, à leur palmarès mais également à leur parcours professionnel et personnel, nous émettons l'hypothèse qu'elle puisse intéresser dans un premier temps les historiens du sport. S'est alors posée la question suivante : qui contacter ? Il se trouve que le contexte sportif nous a particulièrement aidé puisque le SIAF était à ce moment-là, au coeur de la Grande Collecte des

archives du Sport[104] visant à valoriser le patrimoine sportif en vue des Jeux Olympiques de Paris 2024 et travaillait pour cela avec le Comité d'histoire des ministères chargés de la jeunesse et des sports (CHMJS). Sur les conseils de Chloé Moser et avec l'aide de Juliette Hayette, chargée de mission pour la Grande Collecte des archives du Sport au SIAF, nous contactons Patrick Clastres le 15 juin. Ce dernier, membre du CHMJS est un historien du culturel et du politique dont les travaux portent notamment sur l'histoire et la géopolitique du sport international[105]. Un échange est programmé avec lui le 26 juin auquel nous souhaitons donner une double orientation. D'abord, nous souhaitons recueillir son témoignage quant à la manière dont il appréhende une base de données lorsqu'il en a besoin pour ses recherches mais aussi plus spécifiquement sur les potentiels usages qu'il pourrait faire de notre base de données.

Après lui avoir exposé notre projet et constaté son enthousiasme, nous décidons de réfléchir à des éléments de documentation que nous pourrions lui fournir afin de lui donner davantage de clefs de réflexion. La base de données sur les sportifs de haut niveau contenant des informations explicites sur les athlètes dont certaines données médicales qui seront donc protégées pendant 75 ans, nous devons prêter une réelle attention aux éléments fournis qui ne devront bien entendu ne comporter aucune donnée à caractère personnel. L'enjeu est de préparer des documents permettant d'expliciter le type d'informations se trouvant dans la base sans pour autant divulguer aucune donnée confidentielle. Nous avons donc choisi de lui faire parvenir dans un premier temps le dictionnaire de données de la base de données non abouti, contenant le nom et le nombre de lignes de chaque table ainsi que l'intitulé de chaque colonne. Puis, nous lui avons donné accès au schéma non exhaustif de la base ainsi qu'à une capture d'écran de la table SuiviSocioPro représentative de la richesse des données en ce qu'on y lit des renseignements sur le secteur professionnel des athlètes mais également leurs projets sportifs sans pour autant contenir d'informations nominatives.

Lors de notre entretien, Patrick Clastres commence par nous expliquer que bien qu'il n'ait pas de réelle expérience d'analyse de base de données, il projette d'en construire une. Il mène en effet depuis 2018 une étude prosopographique sur les membres du Comité international olympique (CIO) entre 1894 et 1972. Cette enquête vise à reconstituer la trajectoire sociale de ces individus cosmopolites et multi-situés en étudiant leurs engagements associatifs, leur sensibilité politique, leurs lieux et sujets d'études et ainsi rompre la tendance traditionnelle de travail dit « en silos » tendant à isoler l'aspect sportif de l'individu du reste de sa personnalité.

Ce projet permet de faire le lien avec la base de données sur les sportifs de haut niveau puisqu'elle rend compte du fait que dans le sport, les implications sociales sont nombreuses. Selon lui, elle est également intéressante dans la mesure où elle nous renseigne sur l'après-carrière des sportifs ainsi que sur l'évolution de leurs projets professionnels, permettant ainsi de mener des travaux d'anthropologie historique. De ces constats découlent plusieurs sujets de recherches possibles, divisibles en trois catégories. Il est tout d'abord possible de réaliser une histoire des trajectoires individuelles et collectives ou bien une histoire par vagues historiques, par sports ou par genres… Il est également possible de se concentrer sur des problématiques plus factuelles : Quels titres ont été obtenus par tel sportif dans telle discipline ? Quels sont les types de

---

104 Page FranceArchives dédiée à la Grande Collecte des archives du sport, https://francearchives.gouv.fr/fr/article/667843638, consultée le 26 août 2023.

105 Site internet personnel de Patrick Clastres : https://people.unil.ch/patrickclastres/, consulté le 17 juin 2023.

formations proposées ? Enfin, et surtout, cette base de données pourrait donner l'opportunité de mener des travaux centrés autour de la notion de performance. Comment construit-on les doubles carrières ? Comment produit-on de la puissance sportive ? Quel est l'impact du milieu familial sur la performance ? Un sportif de haut niveau a-t-il l'obligation de passer par l'iNSEP pour produire de la performance ? La performance, entendue comme le résultat obtenu par un(e) athlète ou une équipe dans une épreuve sportive est produite par différents acteurs. Son origine réside, selon Patrick Clastres, dans le milieu familial puisque c'est généralement l'un des parents qui emmène son enfant vers une discipline qu'il a parfois lui-même déjà pratiquée. Le club dont il est membre lui permettra possiblement de progresser jusqu'à atteindre une fédération, qui elle-même pourra aller jusqu'à lui faire intégrer l'INSEP.

On pourrait donc se poser la question suivante : est-ce qu'il y aurait dans cette base, des données permettant d'évaluer la capacité des fédérations à produire de la performance sportive française ? En effet, selon l'historien, au delà de son intérêt historique et archivistique, cette base de données est monnayable. Elle permettrait, grâce à une analyse fine des données, d'identifier les fédérations ou entraîneurs produisant moins de performance que d'autres. Un dernier entretien est réalisé avec deux sociologues : Manuel Schotté et Sébastien Fleuriel. Manuel Schotté, professeur de sociologie à l'Université de Lille travaille sur la fabrique sociale de la grandeur individuelle sous l'angle de l'objet sportif. Sébastien Fleuriel, sociologue enseignant-chercheur à l'Université de Nantes s'intéresse quant à lui principalement à la fabrique des élites sportives. Comme Patrick Clastres, les deux sociologues pointent l'intérêt de cette base en ce qu'elle permet de repositionner le sportif dans l'espace social. Ayant déjà connaissance de son existence et après avoir consulté les mêmes documents que ceux fournis à l'historien, Manuel Schotté et Sébastien Fleuriel énumèrent divers sujets de recherches parmi lesquels nous pouvons mentionner l'étude des liens entre performance sportive et scolarité, la condition du sportif de haut niveau en France ou encore la construction des parcours sportifs. Cependant, s'ils confirment l'usage managérial qu'elle permet, ils considèrent qu'aujourd'hui, les modèles à succès et ceux, qui, au contraire ne marchent pas sont déjà connus des spécialistes de la haute performance. Ainsi, selon eux, cette base de données, offrant des sujets de recherches divers et variés, est avant tout une source permettant de réaliser des travaux socio-historiques ou socio-anthropologiques tendant à rompre la logique traditionnelle de travail « en silos ».

2. 3 : Les limites

L'étude des possibles usages futurs de la base de données sur les sportifs de haut niveau nous conduit finalement à réfléchir à la question des limites. Jusqu'où l'archiviste doit-il aller dans l'anticipation des besoins futurs ? Qu'est ce qu'un archivage abouti ? Où s'arrête le travail de l'archiviste ? Où commence celui de l'historien ? Ces questions sont issues des différents échanges cités précédemment auxquels nous devons rajouter un entretien organisé avec Emmanuelle Bermès au mois de mai. Cette réflexion n'a d'autre ambition que de synthétiser les problématiques soulevées lors de ce travail.

Selon nous, l'archiviste précède l'historien, l'enjeu étant de lui donner les clefs d'accès à des sources produites dans le passé. Mais, appliqué aux bases de données, cela ne s'arrête-t-il pas à la garantie d'accéder à un format lisible dans le temps ? Entendu comme cela, nous pourrions ainsi considérer que la conservation

des données brutes sous forme de fichiers à plat par exemple CSV est suffisante. En effet, bien que cette méthode ne permette pas l'appréhension directe de la structure des données, nous savons qu'elle fonctionne. Initialement, nous pensions que l'un des grands points forts de SIARD était justement de répondre à ce besoin de perception de l'organisation des données. En effet, nous établissions une distinction entre les différents utilisateurs futurs possibles d'un tel objet : les statisticiens d'un côté et les chercheurs en sciences humaines de l'autre. Alors qu'il nous semblait cohérent qu'un statisticien préfère travailler sur des données brutes, nous pensions que le chercheur en sciences humaines, lui, trouverait plus aisé de réfléchir à partir d'une base de données dont la structure a été au préalable remontée. Néanmoins, les échanges avec Patrick Clastres, Manuel Schotté et Sébastien Fleuriel ont permis de relativiser ce constat : effectivement, tous préfèrent travailler sur la donnée la plus brute possible et en cela, le format CSV leur convient. C'est par ailleurs un format de fichier qu'ils ont l'habitude de manipuler.

Mais alors, pour anticiper les différents usages possibles, est ce qu'un double archivage peut-être envisagé ? Cette possibilité du double archivage s'est notamment posée pour l'application Diplomatie du ministère de l'Europe et des affaires étrangères. Effectivement, les archivistes avaient alors envisagé de conserver d'un côté la donnée brute et de l'autre, d'émuler le système de l'application afin de permettre à la fois de consulter la donnée brute archivée mais également de pouvoir, comme l'utilisateur originel, naviguer directement dans l'application. L'hypothèse de l'émulation, jugée trop coûteuse, a finalement été abandonnée.

Selon Erwann Ramondenc, le numérique induit finalement une véritable quête du « toujours plus », laquelle n'existait pas du temps où les informations étaient exclusivement produites sur support papier. Avec l'ère numérique, émerge l'idée selon laquelle il faudrait tout conserver afin d'anticiper le maximum d'utilisations futures et perdre le moins d'informations possibles. Mais face à la production croissante de données, l'archiviste ne doit-il pas opérer des choix ? Dans un monde où la problématique de la sobriété énergique se développe de plus en plus, opérer des choix dans la conservation du patrimoine s'impose également. Cet enjeu, bien que visible avec le format SIARD de par la volumétrie qu'il implique, est particulièrement visible avec l'émulation. En effet, cette technique requiert de maintenir dans le temps des émulateurs d'applications ou de systèmes anciens, qu'il faudra eux-aussi émuler dans un temps long. Il semble qu'il soit nécessaire de garder à l'esprit que bien que l'on puisse anticiper certaines utilisations, nous nous inscrivons dans un temps présent qui ne possède pas les mêmes problématiques que celles qui émergeront dans trente, cinquante ou cent ans.

La question des limites est enfin transposable à la rétro-ingénierie. La rétro-ingénierie produit-elle toujours des résultats satisfaisants? À quel moment l'archiviste doit-il considérer son enquête comme achevée ? Le numérique n'induit-il pas, encore une fois, une sorte de quête inassouvie de réponses? Se résigner au « silence des sources » invite finalement à se demander si l'absence d'informations n'est pas elle-même significative. Ce travail nous a conduit à considérer que si la découverte d'informations nous donnait, de facto des éléments de réponses, il en était de même pour l'absence d'éléments. Cette dernière traduit effectivement quelque chose de l'objet à archiver. Bien que l'archiviste et l'historien occupent des fonctions différentes, celles-ci sont pourtant corrélées puisque toutes deux visent à conserver, valoriser et donner à voir aux générations futures les patrimoines matériel et immatériel. Le dialogue et la mise à la place de l'autre

sont donc essentiels. L'archiviste, animé par le respect de l'authenticité et de l'intégrité du document dont il s'occupe, réfléchit au cours de son travail aux enjeux posés par la conservation de l'objet concerné et aux utilisations futures qui pourraient en être faites. De son côté, l'historien ne peut faire son travail sans connaître l'origine, les modalités de conservation, les agrégations ou modifications faites par l'archiviste et les raisons les ayant motivées. Bien que centrée sur ces deux métiers, cette réflexion s'inscrit finalement dans un cadre plus large : celui des enjeux nouveaux posés par l'ère numérique.

Conclusion

Ce stage de quatre mois au service interministériel des Archives de France, aussi riche que passionnant s'est structuré en trois grandes phases dont nous avons tenté de rendre compte dans le présent travail. La première étape qui fut également la plus longue consista en un véritable travail de documentation sur l'archivage des bases de données. Il fallut, dans un premier temps réfléchir à ce qu'était une base de données pour ensuite comprendre les enjeux liés à son archivage. Au delà des multiples articles, travaux menés sur le sujet et retours d'expériences consultés, cet état de l'art s'est appuyé sur divers échanges. Marion Ville, cheffe de projet fonctionnel au sein du programme interministériel VITAM nous exposa les différentes méthodes utilisées aujourd'hui pour l'archivage des bases de données et nous présenta le résultat de sa réflexion menée en 2011 sur l'archivage de la matrice cadastrale. Les Archives Nationales, contactées au titre de leur rôle central, nous firent un historique de l'utilisation de l'archivage à plat au sein de l'institution. Puis, les Archives fédérales suisses répondirent, par le biais d'une présentation claire, aux interrogations posées par le SIAF sur le format d'archivage de bases de données relationnelles SIARD. Enfin, Laurent Duplouy, chef du service multimédias au département audiovisuel à la direction des collections de la Bibliothèque nationale de France, nous exposa la solution utilisée par l'institution pour l'archivage des données structurées. De ces échanges, ressortirent plusieurs techniques : l'archivage à plat, le format SIARD et l'émulation.

Le choix de l'une d'elles résulte de plusieurs critères que nous avons identifiés : les usages originels faits de cette base, les modalités de collecte ainsi que les utilisations futures que l'on envisage. Aussi, si l'archivage à plat est la méthode la plus couramment utilisée en France, il ressort de la documentation que d'autres pratiques existent au niveau international. Au delà de SIARD, nous pouvons mentionner le logiciel DBPTK permettant de faire du SIARD en passant par une autre suite logicielle que SIARDSuite ainsi que CHRONOS, solution propriétaire mise au point par les Etats-Unis et encore peu utilisée aujourd'hui. Si la sollicitation d'institutions centrales et porteuses de solutions d'archivage était à notre sens, absolument nécessaire pour réaliser cet état de l'art, la portée scientifique de ce travail nous a également poussé à nous renseigner sur les pratiques des services d'archives déconcentrés. Pour cela, nous avons choisi de recourir à un questionnaire. Or, bien que ces derniers semblent touchés par la problématique de l'archivage des bases de données, le questionnaire ne nous donne qu'une vision partielle de cet aspect puisque seules dix-huit réponses furent comptabilisées. Il pourrait donc être intéressant d'approfondir cette étude, d'autant qu'elle ne nous renseigne pas sur les pratiques des services d'archives municipaux. Cet état de l'art souligne finalement que bien que le format CSV soit la solution d'archivage la plus largement utilisée en France pour les données structurées, elle n'est pas la seule. Le choix d'une méthode dépend donc de l'objet à archiver et des conditions dans lesquelles le service d'archives le reçoit.

La base de données sur les sportifs de haut niveau confiée à la Mission affaires sociales pour archivage fut un véritable cas d'étude stimulant permettant de transposer concrètement les problématiques théoriques évoquées précédemment. Cette base de données dont nous ne possédions qu'une extraction Excel accompagnée d'une documentation lacunaire nécessita une rétro-ingénierie complète. Ne disposant d'aucune méthodologie précise, ce travail fut guidé par les questions suivantes : Dans quel contexte cette base de données fut-elle créée ? Par qui ? Dans quel(s) but(s) ? A quoi servait-elle ? Était-elle reliée à d'autres

sources ? Comment était-elle alimentée ? Depuis le dépouillement d'archives papier jusqu'à la prise de contact de personnes ressources, cette démarche nécessita une constante adaptation aux résultats trouvés. En effet, nous prîmes le parti d'exploiter chaque piste, considérant que chacune d'elle était l'occasion soit, d'une part, d'enrichir notre connaissance de l'objet à archiver, soit, au contraire, de refermer une porte et de se documenter différemment. Si ce travail ne prétend en aucun cas relever d'une méthodologie de rétro-ingénierie, il tend davantage à mettre en exergue l'idée selon laquelle la rétro-documentation, quels que soient les résultats finaux, consiste en une véritable démarche de recherche nécessitant d'aborder le maximum de pistes possibles. Elle s'inscrit par ailleurs parfois dans un temps et des contraintes fixes dont il faut tenir compte.
Les expérimentations réalisées sur la base de données sur les sportifs de haut niveau des deux solutions d'archivage que sont la mise à plat et le format SIARD furent également l'occasion de confronter objectivement les avantages et les inconvénients de chacune d'elles. Au delà de l'identification du format le plus approprié pour ce cas d'étude, cela permit surtout de mettre en exergue les conditions nécessaires à la mise en pratique du format SIARD.

La troisième et dernière phase de ce stage consista davantage en une réflexion sur la base de données en tant qu'objet pour l'historien. Plus brève que les deux précédentes, elle résulte d'échanges menés avec, d'une part, des archivistes, et de l'autre, des chercheurs et permet d'établir plusieurs liens. Tout d'abord, celui entre archiviste et chercheur, puisque, comme l'évoque l'appel à communications de la journée d'étude proposée par le centre Jean Mabillon, « La source occupe une place fondamentale dans la méthode historienne au point de paraître, aux yeux du chercheur, porteuse d'une vérité absolue qui peut se révéler cependant trompeuse[106]. » L'archiviste, en ce qu'il le précède, permet donc de lui donner les clefs pour accéder le plus justement possible à cette source puisqu'en effet, « revenir à la source permet de construire une méthode scientifique à partir du contexte de production du document pour changer le paradigme de compréhension de ce matériau[107]. » Aussi, il permet de faire le lien entre archivage et usages puisque la conservation d'un document s'inscrit dans un contexte présent, tenant compte à la fois des usages passés mais également de ceux futurs.

De ces quatre mois de stage découle finalement l'idée selon laquelle la politique d'archivage numérique ne couvre pas seulement les documents bureautiques mais concerne également les bases de données. La politique des formats est au coeur du sujet puisque comme pour les autres objets numériques, la question de la pérennisation des formats est centrale, l'enjeu étant de trouver une solution d'archivage des bases de données qui soit la plus stable dans le temps. Considérant que les avancées dans le domaine de la conservation résultent de la mutualisation des savoirs et d'une coordination tant nationale qu'internationale, nous pouvons établir les préconisations suivantes. Avant toute chose, il est important d'établir une nette distinction entre l'archivage d'un document bureautique et celui d'une base de données puisque leur nature diffère. La méthodologie à mettre en place doit donc être adaptée à l'objet concerné. Pour cela, il apparaît

---

106 « Source, poison ou accident : comprendre le document dans les sciences historiques », appel à propositions pour la journée d'études doctorales proposée par le centre Jean Mabillon, École nationale des chartes prévue le 19 octobre 2023.

107 *Loc. cit.*

nécessaire de former les services d'archives. En effet, les bases de données étant de plus en plus nombreuses, la problématique de leur archivage se révèlera de plus en plus fréquente. Ce besoin fut notamment exprimé dans le cadre du questionnaire diffusé au réseau d'archives départementales. Sur les 18 services répondants, 15 répondent positivement à la question « Souhaiteriez-vous assister à une formation sur l'archivage des bases de données ? ». C'est d'ailleurs la réponse à ce besoin qui permettra la mise en place d'une stratégie. De plus, si un service d'archives souhaite utiliser la méthode SIARD pour archiver une base de données relationnelle, il est important qu'il s'assure de répondre aux conditions initiales nécessaires à sa mise en application, à savoir l'hébergement sur un système de gestion de bases de données compatible avec le format et la prise en charge de cette dernière avant son décommissionnement. Enfin, il est préférable dans toute démarche d'archivage d'intervenir le plus en amont possible et d'établir une coordination étroite entre services producteurs et archivistes afin d'expliciter les besoins de chacun.

Entre documentation, recueil de témoignages d'instances tant nationales que territoriales et expérimentations techniques, cette réflexion sur l'archivage des bases de données a nécessité une confrontation continue entre pratiques archivistiques et besoins des chercheurs. Fruit d'un travail collaboratif, il témoigne de l'importance de la mutualisation des savoirs et d'une communication étroite, toujours imprégnée des notions d'intégrité et d'authenticité, situées au centre du métier d'archiviste.

ANNEXES

Annexe n°1 : Message accompagnant le questionnaire destiné au réseau d'archives départementales, diffusé le 21 juin 2023.

Répondre Répondre à tous Transférer MI

BEAU Clémence 'liste.resp-archives-etat@culture.gouv.fr'; archivage.numerique.siaf 21/06/2023

[SIAF] Questionnaire sur l'archivage des bases de données locales

Mesdames, Messieurs,

Actuellement étudiante en master 2 Technologies numériques appliquées à l'histoire dispensé par l'Ecole nationale des chartes, je me permets de vous contacter dans le cadre de mon stage au sein du bureau de l'expertise numérique et de la conservation durable du Service interministériel des Archives de France (sous la direction de Violette Lévy et Dominique Naud). Mon stage, qui a débuté le 3 avril et doit se terminer le 28 juillet, porte sur l'archivage des bases de données.

L'objectif du stage est de produire un livrable dressant à la fois un état de l'art du sujet en France et en Europe, mais aussi des préconisations actualisées sur l'archivage des bases de données tout en s'adossant à un cas d'usage concret.

Les Archives nationales ainsi que des institutions étrangères ont été consultées et nous ont renseignés sur leurs méthodologies respectives. Néanmoins, pour que cette étude soit le reflet le plus complet possible des pratiques actuelles, il apparaît indispensable d'affiner les éléments remontés par le biais de l'enquête annuelle sur les archives nativement numériques en recueillant des informations sur vos pratiques en termes d'archivage des bases de données produites et collectées localement.
Pour cela, un questionnaire a été élaboré, accessible en ligne à l'adresse suivante https://framaforms.org/questionnaire-sur-larchivage-des-bases-de-donnees-locales-1687177742.
Il va être également proposé dans l'espace Osmose dédié à l'archivage numérique. Il s'agit de recueillir des éléments sur les typologies de bases de données archivées et sur la stratégie adoptée pour leur conservation.

Je vous serais très reconnaissante de bien vouloir répondre à ce questionnaire avant le 7 juillet. En effet, il est nécessaire de prévoir un temps de traitement des réponses pour que des éléments d'analyse des pratiques concrètes et des besoins des services d'archives en terme de conservation, de pérennisation et de mise à disposition des bases de données locales puissent être inclus dans le mémoire avant la fin du stage.

En vous remerciant par avance pour votre participation, je reste à votre disposition pour toute précision complémentaire.

Bien respectueusement,

Clémence Beau clemence.beau@culture.gouv.fr

Annexe n°2 : Questionnaire diffusé à l'ensemble du réseau d'archives départementales sur les pratiques en termes de collecte, d'archivage et de diffusion des bases de données locales.

Questionnaire sur l'archivage des bases de donnée locales

L'objet de ce document est de recueillir les pratiques des services d'archives départementaux en terme de collecte, de méthodologie d'archivage et de mise à disposition des bases de données locales.

Introduction

Une astérisque signifie que la réponse à la question est obligatoire.

Merci de renseigner les informations ci-dessous :

Nom du répondant *

Prénom du répondant *

Fonction *

Courriel *

Service d'archives *

Pour contacter l'auteur·rice de ce formulaire, cliquez ici
Ne communiquez aucun mot de passe via Framaforms.

Disposez-vous d'un système d'archivage électronique ? *

- ( ) Oui
- ( ) Non

Des opérations d'archivage de bases de données sont-elles réalisées dans votre service ? *

- ( ) Oui
- ( ) Non

Si oui, à quelle fréquence? *

- Sélectionner -

Comment archivez-vous une base de données ? *

- Sélectionner -

L'émulation est une technique consistant à reproduire sur un environnement matériel et logiciel "hôte" et actuel, un environnement matériel "invité" généralement plus ancien au moyen d'un émulateur.

Format SIARD

Connaissez-vous le format SIARD ? *

- ( ) Oui
- ( ) Non

Modalités d'accès

Qu'envisagez-vous pour mettre à disposition les bases de données archivées ? *

- Sélectionner -

Besoins

Avez-vous déjà reçu une formation sur l'archivage des bases de données ?

- ( ) Oui
- ( ) Non

Merci de compléter ces informations pour chaque base de données archivée dans votre service :

Nom de la base

Volumétrie (en Go)

Nombre total de tables dans la base

Nom du producteur

Dates de fonctionnement (exemple : 1997-2001)

Date de décommissionnement (si décommissionnée)

Nature des données de la base

Comment cette base était-elle alimentée ?

- ○ Directement
- ○ Indirectement

Directement : les données sont saisies directement dans l'ensemble des champs de la base de données.

Indirectement : les données sont récupérées (totalement ou partiellement) d'autres bases de données. La récupération des données est automatisée par le biais d'une interface de programmation (API).

Quel type de données cette base contient-elle ?

- ○ Données statistiques
- ○ Données individuelles
- ○ Autres données

A votre connaissance, quelles autres sources d'informations (archives bureautiques, autre(s) base(s) de données, archives d'autres producteurs etc...) seraient utiles à la compréhension des données de la base ou pourraient s'interfacer avec elle ?

Quel fut le facteur de déclenchement du processus d'archivage de la base de données ?

- Aucun(e) -

Comment l'avez-vous archivée ?

- Aucun(e) -

Dans le cas d'un export, qui s'en charge ?

- Aucun(e) -

La base est-elle conservée telle quelle ou bien est-elle transformée ?

- ○ Conservée dans son format d'origine
- ○ Transformée

Documentation

La base de données est-elle accompagnée d'une documentation ?

- ○ Oui
- ○ Non

Pour contacter l'auteur·rice de ce formulaire, cliquez ici
Ne communiquez aucun mot de passe via Framaforms.

Merci de cocher les éléments présents dans la documentation

- ☐ Description générale et historique du projet
- ☐ Documentation technique
- ☐ Modèle de données
- ☐ Liste de codes (référentiels)
- ☐ Liste de description des tables et/ou des champs

Si d'autres types de documents sont présents, merci de les indiquer ci-dessous :

Sous quelle forme cette documentation est-elle délivrée ?

- ○ Papier
- ○ Numérique

Annexe n°3 :

Résultats du questionnaire.

Les résultats de ce questionnaire sont consultables sur Github.

Bibliographie

Pratique archivistique générale :

Ouvrages généraux :

- Association des archivistes français, *Abrégé d'archivistique, Principes et pratiques du métier d'archiviste*, Paris, 2020.

- Direction des Archives de France, *La Pratique archivistique française*, sous la direction de Jean FAVIER assisté de Danièle NEIRINCK, 1993.

Cadre méthodologique :

- *Cadre méthodologique pour l'évaluation, la sélection et l'échantillonnage des archives publiques*, Délégation interministérielle aux Archives de France, Juillet 2014, https://francearchives.gouv.fr/file/5f01f41db3790b5201ff6c29413c16521a57ccf6/static_7742.pdf, consulté le 26 juillet 2023.

Journées d'études :

- « Source, poison ou accident : comprendre le document dans les sciences historiques », appel à propositions pour la journée d'études doctorales proposée par le centre Jean Mabillon, École nationale des chartes prévue le 19 octobre 2023.

Archivage électronique :

Ouvrages généraux :

- Association des archivistes français, *Les archives électroniques*, Paris, 2014 (réed. 2020).

- BANAT-BERGER Françoise, DUPLOUY Laurent, HUC Claude, *L'archivage numérique à long terme. Les débuts de la maturité*, La documentation française, 2009.

Articles :

- Bulletin des Archives de France sur la conservation à long terme des documents électroniques, n°13, novembre 2003, https://francearchives.gouv.fr/file/c29d19e5dc4e4c4cdb22886b3790ccdb040459de/static_1677.pdf, consulté le 29 août 2023.

- CONCHON Michèle, Constance a dix ans : bilan et perspectives de l'archivage des fichiers informatiques aux Archives nationales, *La Gazette des archives*, 1993, n°163, pp. 318-324.

- GUYON Céline, « L'archivage comme dispositif de transformation de la nature intrinsèque des objets nativement numériques », *Balisages*, 2020, https://publications-prairial.fr/balisages/index.php?id=282, consulté le 28 août 2023.

- GUYON Céline, « La pratique archivistique publique en France, entre adaptation et négociation. Expériences et réflexions d'une archiviste. », *Les cahiers du numérique*, n°11, 2015 (p. 77 à 114), https://www.cairn.info/revue-les-cahiers-du-numerique-2015-2-page-77.htm, consulté le 28 août 2023.

- HEROLD Béatrice, LEVY Violette, Exposer, consulter et… réutiliser les ressources : un changement de paradigme, La Gazette des archives, 2019, n°254, pp. 231-245.

- MAGNIEN Agnès (dir.), GUICHARD-SPICA Hélène, LAPERDRIX Marie, LOPEZ Magalie, LEBLANC Marie-Noëlle, GALLET-MENAGER Isabelle, *Vade-mecum de l'archivage des documents électroniques*, janvier 2012.

- SIN BLIMA-BARRU Martine, VAN DE WALLE Thomas, L'archivage numérique aux Archives Nationales : de Constance à ADAMANT, *La Gazette des archives*, 2015, n°240, pp. 73-74.

L'archivage des documents bureautiques :

- POIVRE Joël, *L'archivage des documents bureautiques. Manuel pratique*, Direction des Archives de France, Paris, 2004.

Les bases de données :

Ouvrages généraux :

- BEGON-TAVERA Hélène, *La transformation numérique des administrations*, La documentation française, 2021.
- HAINAUT Jean-Luc, *Bases de données. Concepts, utilisation et développement*, DUNOD, 2022 (5e édition).

La conservation des bases de données :

Guides méthodologiques :

- BÉCHARD Lorène, PRAT Philippe, Guide méthodologique pour l'archivage des bases de données, Centre d'informatique national de l'Enseignement Supérieur (CINES), 16 avril 2013, https://www.cines.fr/wp-content/uploads/2022/05/GM_archivage_BDD-v1.1.pdf, consulté le 8 avril 2023.

Le format SIARD :

Articles :

- GUÉNERAIS Marine, MIGNOT Gaelle, L'archivage d'une base de données : le sommier de l'Établissement public de Saint-Quentin-en-Yvelines, *Gazette des Archives*, 2015, n°240, p. 69-71, https://www.persee.fr/doc/gazar_0016-5522_2015_num_240_4_5279, consulté le 23 août 2023.

Mémoires :

- NICHELE Baptiste, Interopérabilité et pérennisation des archives électroniques. L'exemple du SEDA/EAD et du SIARD (étude de cas), mémoire de master Technologies numériques appliquées à l'histoire, École nationale des chartes, 2010.

- VILLE Marion, La matrice cadastrale : archiver et exploiter une base de données, mémoire de master Technologies numériques appliquées à l'histoire, École nationale des chartes, 2012.

- YILDIRIM Gülsen Delia, « La base de données de l'aide sociale SOSTAT : Analyse, évaluation et proposition d'archivage pour les Archives de la République et Canton du Jura », travail de Bachelor, Haute École de Gestion de Genève (HEG-GE), Porrentruy, 27 juillet 2011.

Webographie

L'archivage électronique :

- Digital Preservation Coalition, *Manuel de préservation numérique*, 2021, https://www.association-aristote.fr/wp-content/uploads/2022/03/pres_format_Handbook_version_fr_2021.pdf, consulté le 30 août 2023.

- GUYON Céline, « L'archivage comme dispositif de transformation de la nature intrinsèque des objets nativement numériques », *Balisages*, 2020, https://publications-prairial.fr/balisages/index.php?id=282, consulté le 30 août 2023.

- International Research in Permanent Authentic Records in Electronic Systems (InterPARES), « Conditions requises pour évaluer et maintenir l'authenticité des documents d'archives électroniques », http://www.interpares.org/display_file.cfm?doc=ip1_authenticity_requirements(french).pdf, consulté le 25 juillet 2023.

Format SIARD :

- Archives fédérales suisses, « Sauvez vos bases de données. SIARD, la solution d'archivage pour les bases de données relationnelles », 2010.
- DILCIS Board, « SIARD-2.2 Format Specification », 31 août 2021, https://siard.dilcis.eu/SIARD 2.2/SIARD 2.2.pdf, consulté le 26 août 2023.

- JACOBSON Michel, Retour d'expérience sur l'utilisation du format SIARD pour l'archivage des bases de données relationnelles, Journée de sensibilisation à la sécurisation et à la pérennisation des données, RBDD, 6 novembre 2014. file:///C:/Users/CLEMEN~1.BEA/AppData/Local/Temp/rbdd_siard.pdf, consulté le 15 avril 2023.

- LEUTHOLD Jérémie, « Solutions pour l'archivage de bases de données relationnelles. Le cas de SIARD. », Groupe PIN, Paris, 7 octobre 2010, http://pin.association-aristote.fr/lib/exe/fetch.php/public/presentations/2010/pin20101007-pres02-siard.pdf, consulté le 5 juin 2023.

- Service interministériel des Archives de France, « Étude du format SIARD - Software Indépendant Archiving of Relationnel Databases », 2010, https://francearchives.gouv.fr/fr/file/d5459ac36adafce153008f6decb3985b1dcf9b46/DGP_SIAF_2010_017_Etude_format.pdf, consulté le 24 juillet 2023.

- Service interministériel des Archives de France, « Note d'information relative au versement de la matrice cadastrale de l'année 2004 aux services départementaux d'archives », 12 janvier 2016,

https://francearchives.gouv.fr/fr/file/2d5cf9bac98dd660e5df54e6f9495093ff5f3c01/DGP_SIAF_2017_004_versement_matrice_2005_2006_VD.pdf, consulté le 23 juillet 2023.

Émulation :

- ESPOSITO Nicolas, *Emulation et conservation du patrimoine culturel lié aux jeux vidéo*, Université de technologie de Compiègne, 2004. https://www.researchgate.net/profile/Nicolas-Esposito/publication/215673465_Emulation_et_conservation_du_patrimoine_culturel_lie_aux_jeux_video/links/54e0aab00cf29666378d40a6/Emulation-et-conservation-du-patrimoine-culturel-lie-aux-jeux-video.pdf, consulté le 11 juin 2023.

- GRANGER Stewart, Emulation as a Digital Preservation Strategy, *D-Lib Magazine*, Volume 6, Number 10, UK Project Co-ordinator of the CAMiLEON Project, University of Leeds.

Autres initiatives internationales en termes d'archivage de bases de données :

CHRONOS :

- BRANDL Stefan, KELLER-MARXER Peter, « Long term Archiving of Relational Databases with Chronos », First International Workshop on Databases Preservation, University of Edinburgh, 23 march 2007, https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=cdfefc849bf379190affafc87d7fdd1c57eb8f4a, consulté le 17 avril 2023.

- LINDLEY Andrew, « Database Preservation Evaluation Report - SIARD vs. CHRONOS. Preserving complex structures as databases through a record centric approach? », Joint Workshop of the Digital Preservation Coalition and UK Data Service, London, 17 march 2016, https://purl.pt/24107/1/iPres2013_PDF/Database_Preservation_Evaluation_Report_-_SIARD_vs._CHRONOS.pdf, consulté le 17 avril 2023.

DBPTK :

- BRÉGIER Frédéric, « Rapport de test de DB-Preservation-Toolkit et Database-Visualization-Toolkit », 16 novembre 2011.

- RAMALHO José Carlos, FARIA Luis, SILVA Hélder, COUTADA Miguel, « Database Preservation Toolkit : a flexible tool to normalize and give access to databases », Biblioteca Nacional de Portugal, 2014, https://purl.pt/26107/1/DLM2014_PDF/15_-_Database_Preservation_Toolkit.pdf, consulté le 17 avril 2023.

Table des matières

Résumé 3

Remerciements 5

Introduction 7

PREMIÈRE PARTIE - ÉTAT DE L'ART : L'ARCHIVAGE DES BASES DE DONNÉES D'HIER À AUJOURD'HUI 11

Chapitre 1 : Besoin de définition et définition des besoins 12

1. 1 : Définitions liminaires 12

1. 2 : Historique 13

1. 3 : Besoins du service interministériel des Archives de France : contextualisation du stage 16

Chapitre 2 : Préalables archivistiques 19

2. 1 : Notions clefs 19

2. 2 : L'évaluation d'une base de données 20

Chapitre 3 : Archiver une base de données : focus sur les pratiques actuelles 23

3. 1 : L'archivage à plat 23

3. 2 : Le format SIARD 25

3. 3 : L'émulation 27

DEUXIÈME PARTIE - CAS D'USAGE CONCRET : PROPOSER UNE SOLUTION POUR L'ARCHIVAGE D'UNE BASE DE DONNÉES SUR LES SPORTIFS DE HAUT NIVEAU 31

Chapitre 4 : De la présentation du projet au travail de rétro-ingénierie 32

1. 1 : La base de données sur les sportifs de haut niveau 32

1. 2 : Rétro-ingénierie de la base de données 36

Chapitre 5 : Expérimenter diverses solutions d'archivage 40

2. 1 : L'archivage à plat 40

2. 2 : L'archivage au format SIARD 41

Chapitre 6 : Conclusions des expérimentations 45

3. 1 : Mettre en perspective le travail du questionnaire et les expérimentations 45

3. 2 : Choix d’archivage final 47

TROISIÈME PARTIE - RÉFLEXIONS AUTOUR DE LA BASE DE DONNÉES EN TANT QU’OBJET PATRIMONIAL 49

Chapitre 7 : Réflexions autour de la notion de source 50

1. 1 : Appréhender les bases de données sous l’angle de la source 50

1. 2 : Distinguer l’objet intellectuel de l’objet technique 50

Chapitre 8 : Placer l’usage au centre des enjeux de l’archivage : corrélation entre passé et futur 52

2. 1 : Identifier les usages passés pour anticiper les usages futurs 52

2. 2 : Enquête autour des usages futurs possibles de la base de données sur les sportifs de haut niveau 53

2. 3 : Les limites 55

Conclusion 58

ANNEXES 62

Annexe n°1 : Message accompagnant le questionnaire destiné au réseau d’archives départementales, diffusé le 21 juin 2023. 63

Annexe n°2 : Questionnaire diffusé à l’ensemble du réseau d’archives départementales sur les pratiques en termes de collecte, d’archivage et de diffusion des bases de données locales. 64

Annexe n°3 : 68

Résultats du questionnaire. 68

Bibliographie 70

Webographie 73

Table des matières 75