# Section 12 - Études de cas

FRANÇOISE BANAT-BERGER
CLAUDE HUC

version 1

22 novembre 2011

# Table des matières

# Chapitre 1. Objet de la section

## A. Introduction

Dans cette partie seront présentés divers exemples de mises en œuvre de plateformes d'archivage électronique dans plusieurs secteurs : scientifique, patrimonial, institutionnel (archives, bibliothèque...).

Dans le deuxième chapitre, une brève incursion dans les projets mondiaux nous donnera un aperçu de l'engouement général pour la mise en place d'un archivage électronique. Nous constaterons qu'en Europe les projets dépassent de beaucoup le cadre de la Francophonie. De toutes ces réalisations, nous n'en mentionnons que quelques-unes et sans entrer dans les détails, en donnant toutefois la possibilité d'y accéder. Nous voulons illustrer par là le bouillonnement d'activités autour de l'archivage numérique et souligner la nécessité d'une saine coopération au sein de la mondialisation de la communication, encore plus indispensable que dans une gestion traditionnelle des échanges professionnels (voir module 14).

Le troisième chapitre traite de retours d'expériences qui ont eu lieu en France ; ils sont de deux ordres :

- d'une part, l'analyse des multiples difficultés rencontrées dans le passé dans la gestion, la pérennisation et la mise à disposition d'informations numériques,
- d'autre part l'examen des premiers enseignements résultant de la mise en place d'organisations et de plates-formes d'archivage numérique dans différents domaines, de la mise en application du modèle de référence OAIS et des autres normes applicables au domaine de l'archivage numérique.

**Cette partie devrait être enrichie par des retours d'expériences complémentaires francophones**.

### Remarque

Pour les apprenants débutants

Vous, les apprenants qui étiez débutants en début de parcours dans le domaine des archives numériques, vous êtes désormais supposés aptes à suivre les développements des expériences relatées ici.

C'est pourquoi dans cette section nous avons peu chargé le texte de caractères gras vous aidant jusqu'ici à sa lecture, ni surligné des termes du glossaire à ce stade désormais assimilés, ni proposé de thèmes en « complément ».

Nous vous supposons maintenant au niveau de ceux qui, avec plus de pratique professionnelle, ont abordé ce module.

La lecture des chapitres ci-dessous peut vous servir de test. Si vous avez de grosses difficultés de compréhension, alors n'hésitez pas à reprendre l'étude des sections précédentes, cette fois-ci en abordant les « compléments ».

# Chapitre 2. Contexte mondial

Au plan international, les initiatives, programmes d'études et de recherche, réalisations, projets émanant des institutions publiques sont nombreux.

## A. 2.1. International

Ils sont nombreux, notamment aux États-Unis, au Canada et en Australie.

Nous n'en citerons que quelques-uns.

- Bien que datant un peu, le rapport de synthèse « Digital preservation and permanent access to scientific information: the state of the practice » publié en 2004 sous l'égide du Conseil International pour l'Information Scientifique et Technique propose une vue de synthèse très intéressante des multiples activités entreprises dans ce domaine.
- Le SDSC (San Diego Supercomputer Center) (*http://www.sdsc.edu/*) est très actif dans le domaine. Il a conduit un certain nombre de projets en coopération avec d'autres institutions fédérales aux États-Unis : ICAP (Incorporating Change Mangement into Archival Processes), ou encore PAT (Persistent Archives TestBed) qui vise à expérimenter les technologies de grilles informatiques pour l'archivage. Le laboratoire SALT (Sustainable Archives and Library Technologies) du SDSC joue un rôle essentiel dans ces projets.
- Fin 2007, un groupe américain intitulé « Blue Ribbon Task Force on Sustainable Digital Preservation and Access », *(http://brtf.sdsc.edu/*[1]) financé par la National Science Fondation et la fondation Andrew W Mellon, a été créé dans le but d'élaborer un modèle économique de l'archivage numérique.
- Le secteur des entreprises privées n'est pas absent de ce mouvement. IBM et SUN sont très actifs sur le domaine. D'abord spécialisés dans le domaine du stockage de données (la société StorageTek avait été rachetée par SUN il y a quelques années), ces industriels ont compris les grands enjeux de l'archivage et l'intérêt qu'ils pouvaient avoir à être présents sur ce secteur d'activité.

1 - http://brtf.sdsc.edu/

# B. 2.2. Europe

En Europe les réalisations aussi sont nombreuses, notamment au niveau des institutions européennes, et les projets foisonnent un peu partout.

## 2.2.1. Commission européenne

Les actions entreprises et menées par la Commission européenne, au regard de l'archivage numérique, se situent sur plusieurs plans.

- **Un ensemble de directives** ayant un impact indirect important sur l'archivage numérique sont et seront applicables à l'ensemble des pays de l'Union. Citons à ce titre la Directive 1999/93/EC du 13 décembre 1999, sur un cadre communautaire pour les signatures électroniques (voir la section 10 Intégrité, authenticité, preuve), ou encore la Directive 2007/2/CE du 14 mars 2007 établissant une infrastructure d'information géographique dans la Communauté européenne (INSPIRE) (voir la section 9 sur les métadonnées).
- **Une action plus volontariste et directement ciblée** sur les applications d'archivage électronique, notamment dans le domaine de la gestion de l'archivage (records management), avec l'organisation et le financement de l'élaboration du MoReq2 (Exigences-types pour la maîtrise de l'archivage électronique. Mise à jour et extension - 2008. Spécifications MoReq2, traduction sur le site de la direction des Archives de France : http://www.archivesdefrance.culture.gouv.fr/static/2094) préparé sous la conduite scientifique du DLM Forum et publié sous l'égide de la Commission (http://www.moreq2.eu/).
- **Un ensemble d'actions retenues au sein du Programme Cadre** de Recherche et Développement (PCRD) de l'Union européenne. Un budget de plusieurs dizaines de millions d'euros a ainsi été consacré à l'analyse, à la recherche et à l'expérimentation de solutions globales adaptées à tel ou tel contexte d'archivage, au développement de synergies entre les acteurs du problème, à l'éducation, l'enseignement et la prise de conscience au sein de communautés qui étaient souvent éloignées de ces préoccupations.

Les actions de recherche qui mettent à contribution un grand nombre d'acteurs géographiquement dispersés au sein d'un même projet ont un rapport résultat/investissement qui est généralement modéré. Par contre, ces actions jouent un rôle réel dans la mise en relation de tous les acteurs de la chaîne, ainsi que des différents secteurs d'activités confrontés au problème, dans l'élargissement de la prise de conscience, dans la mise à disposition d'informations, de résultats d'études et d'analyses.

Les projets décidés et financés dans le cadre du PCRD sont limités à quelques années. La poursuite de l'activité entreprise par les partenaires du projet au-delà de la période de financement est souvent incertaine. En outre, ces projets résultent de l'initiative de groupes et d'institutions très divers. Il s'ensuit que la cohérence et la complémentarité des différents projets sont difficiles à assurer.C'est pour pallier ces insuffisances qu'un certain nombre d'organismes majeurs dans le domaine de l'archivage numérique se sont regroupés au sein d'une entité pérenne qui porte le nom de « Alliance for permanent access » (http://www.alliancepermanentaccess.eu/). Ces organismes représentent aussi bien le domaine de l'archivage patrimonial des documents et des publications (British Library, bibliothèques nationales des Pays-Bas, d'Allemagne, archives nationales de Suède...) que le domaine de l'archivage des données (Agence Spatiale Européenne, Centre Européen de Recherche Nucléaire...). L'orientation d'ensemble est limitée au domaine scientifique au sens large, couvrant tout aussi bien la physique, la biologie, la chimie, que les sciences de l'environnement ou les sciences humaines.

L'Alliance a pour ambition d'être le lieu naturel de développement des collaborations et des relations privilégiées entre les Archives et d'être un puissant porteur de l'expression de leurs besoins vis-à-vis des instances nationales et européennes. Elle vise à élargir, au niveau international, les efforts de mutualisation déjà entrepris au niveau national. Cette approche

se situe dans une vision globale d'infrastructures à l'instar d'autres grands programmes de recherche relatifs aux infrastructures réseau européennes (le réseau haut-débit GEANT) ou aux infrastructures de grilles informatiques. De même que l'infrastructure réseau européenne s'appuie sur un ensemble d'infrastructures nationales organisées en cohérence les unes par rapport aux autres, l'infrastructure de l'information numérique européenne devrait s'appuyer, dans cette perspective, sur un ensemble de moyens nationaux coordonnés. Le maintien de la cohérence d'ensemble implique l'existence d'une organisation européenne pérenne.

### 2.2.2. Autres principaux projets en cours ou achevés

Certains de ces projets sont des projets de recherche qui illustrent le bouillonnement d'activités autour de l'archivage numérique, nous ne considérons pas pour autant à ce stade que les questions difficiles auxquelles ils s'intéressent sont résolues.

Voici quelques-uns de ces projets

- **CASPAR** (*Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval*) (*http://www.casparpreserves.eu/*)

est un projet intégré du 6ème PRCD (programme cadre de recherche et développement ) d'une durée de 3 ans et qui a commencé en avril 2006. Il centre ses objectifs sur la pérennisation de l'information dans les domaines artistique, scientifique et culturel. Il présente aussi l'intérêt de rechercher, mettre en œuvre et diffuser des solutions innovantes basées sur le modèle OAIS.

La synthèse des contributions d'organismes aussi différents que l'agence spatiale européenne, l'UNESCO, l'Institut National de l'Audiovisuel (INA) ou encore l'Institut de recherche et de coordination acoustique/musique (IRCAM) peut être fructueuse.

Les objectifs essentiels de CASPAR consistent d'une part à élaborer une méthodologie applicable à une grande diversité de domaines et de situations, d'autre part à étudier, développer, valider et intégrer des composants constituant les blocs de base logiciels d'un cadre général et enfin créer ce cadre global et l'expérimenter.

De ce point de vue, les concepts d'identifiants pérennes associés à chaque objet et pointant sur l'Information de représentation correspondante et de réseau « Representation Information registry/repository network » sont intéressants.

Le concept de virtualisation mis en avant par CASPAR doit pouvoir être utilisé dans une série de domaines. Ce concept contribue à l'indépendance de l'approche par rapport aux technologies et aux logiciels existants. Il est mis en pratique depuis longtemps pour les fonctions de stockage mais son extension à d'autres fonctions peut ouvrir des voies nouvelles qu'il convient d'explorer au maximum.

- **DPE** (*Digital Preservation Europe*) (*http://www.digitalpreservationeurope.eu/*)

DPE est un programme de coordination relevant du 6ème PCRD (programme cadre de recherche et développement). Il est le successeur d'ERPANET qui avait un rôle similaire. L'institut leader, le « Humanities Advanced Technology and Information Institute (HATII) », (University of Glasgow) est d'ailleurs le même. DPE vise à contribuer au regroupement et à faciliter les interactions entre les différentes expertises existant en Europe dans le domaine de la recherche, dans le domaine culturel, dans l'administration publique et dans l'industrie sur la question de la pérennisation de l'information numérique.

Plus précisément, DPE a défini trois objectifs principaux :

• mettre en place une plate-forme de coopération, de collaboration, d'échange et de diffusion des résultats des recherches et des expériences dans le domaine de la pérennisation des objets numériques ;

• contribuer au développement de services d'archivage numérique viables : apporter un support à un développement européen des standards en matière d'audit et de certification des archives, ce développement étant une étape essentielle pour la création de services de gestion de contenus numériques et de fédérations d'Archives numériques ; inciter les

industriels dans le domaine de l'information et de la communication à prendre en compte les questions de conservation de l'information dans les futures générations de logiciels relève également de ce second objectif ;

• développer la prise de conscience, les compétences et les ressources disponibles.

L'analyse du projet DPE s'intéresse aussi aux aspects culturels du document numérique, c'est-à-dire à la perception de la nature de cette information au sein de la société.

- **PLANETS** (*Preservation and Long-term Access through Networked Services*) (*http://www.planets-project.eu/*)

PLANETS est un programme de recherche du 6ème PCRD (programme cadre de recherche et développement) qui a commencé en juin 2006 pour une durée de quatre ans. Son objectif est de développer un réseau de services et des outils d'aide à la conservation de l'information numérique dans les domaines de la culture et de la science. En pratique, il est fortement centré sur les besoins des bibliothèques et des archives institutionnelles.

Les bibliothèques et les archives nationales ont la responsabilité légale de la pérennisation de l'information numérique, elles doivent offrir un accès fiable et constant à la connaissance dans les domaines culturel et scientifique mais elles ont souvent des moyens réduits pour assurer aujourd'hui cette conservation de l'information pour les générations futures. PLANETS a pour hypothèse que le problème à résoudre exige des moyens et des compétences qui dépassent largement les capacités propres de chaque institution prise individuellement.

Dans ce contexte, PLANETS vise donc à accroître les compétences et le savoir faire de l'Europe au travers d'un certain nombre d'actions :

• sensibiliser les décideurs aux problèmes posés par la pérennisation du patrimoine scientifique et culturel,

• maîtriser les coûts de la pérennisation en facilitant l'automatisation des tâches et en développant des infrastructures adaptables à la volumétrie,

• aider au développement d'un consensus des communautés utilisatrices,

• définir le périmètre du marché commercial des services et des outils dans ce domaine,

• construire des bancs de test mettant en œuvre des solutions concrètes en intégrant l'expertise et les outils existants.

- **Le projet PROTAGE** (*http://www.protage.eu/*)

s'intéresse à l'automatisation des différentes tâches de transfert et d'archivage des objets numériques en vue de réduire les coûts et à l'intégration de ces automatismes au sein des infrastructures existantes.

- **PrestoSpace** (*Preservation towards storage and access - Standardised Practices for Audiovisual Contents in Europe*) (*http://www.prestospace.org/*)

vise à élaborer et à intégrer des solutions techniques en vue de l'archivage des diverses formes de collections audiovisuelles numériques. Enfin, très récemment, le projet SHAMAN (http://cordis.europa.eu/fp7/ict/telearn-digicult/digicult-projects-shaman_en.html) analyse et définit un ensemble de règles de gestion des objets et des collections en vue de leur pérennisation. Le projet s'appuie sur un système de gestion de données à base de règles pour les grilles de données. Ces règles constituent des ensembles d'assertions permettant à chaque détenteur d'objets et de collections numériques de caractériser ces objets et ces collections en terme d'état et de besoin de pérennisation. Les règles sont alors interprétées par un moteur d'application des règles qui décidera des choix de stockage des données sur un ensemble de nœuds de la grille et des conditions d'accès à ces données.

Au plan des nombreuses initiatives nationales, citons encore le réseau NESTOR en Allemagne et le Digital Curation Center en Grande-Bretagne.

- **Le réseau allemand NESTOR** (*Network of Expertise in Long-Term Storage of Online resources*) (*http://www.langzeitarchivierung.de/index.php?newlang=eng*)

NESTOR a développé une compétence dans le domaine de la certification des archives.

- Le **DCC** (*Digital Curation Center*) (*http://www.langzeitarchivierung.de/index.php?*

*newlang=eng*)

en Grande Bretagne sur les besoins des bibliothèques et des archives institutionnelles.

# Chapitre 3 : Expériences en France

**Remarque**

D’autres chapitres pourront concerner d’autres expériences de pays francophones.

Plusieurs retours d’expériences vont être analysées, la première concerne celle d’un organisme scientifique, les suivantes sont des retours d’archives institutionnelles nationales ou locales dont les expériences nous intéressent au premier chef de par les professions que la plupart d’entre nous, dans notre communauté d’internautes PIAF, exerçons.

## A. 3.1. L'expérience d'un grand organisme scientifique : le Centre national d'études spatiales (CNES)3.

Le retour d'expériences du CNES mérite d'être analysé à plusieurs titres.

En premier lieu, il témoigne d'une expérience de longue date sur le sujet, expérience qui a conduit à des réalisations qui ont valeur d'exemple et qui ont fortement contribué à l'émergence d'une méthodologie aujourd'hui largement reconnue.

En second lieu, elle fait aussi apparaître les décalages qui peuvent exister entre d'une part les ingénieurs et les chercheurs sur le terrain qui ont une conscience aiguë des problèmes posés et des risques encourus, et d'autre part les directeurs et les présidents des établissements publics qui ont une vision plus floue de ces questions dont ils considèrent parfois qu'il s'agit d'un sujet essentiellement technique.

### 3.1.1. Données scientifiques spatiales

Les missions spatiales peuvent être classées en différentes catégories en fonction de leurs objectifs : les unes ont une vocation fondamentalement scientifique, les autres sont tournées vers des applications de diverses natures : télécommunication, télévision, localisation et collecte de données, météorologie, observation de la Terre, applications qui peuvent être civiles ou militaires. Certaines missions, en particulier toutes celles qui observent et étudient la Terre et son environnement, visent à répondre en même temps à des besoins scientifiques à long terme et à des questions économiques à court et moyen terme comme le suivi de la végétation ou l'étude du cycle de l'eau.

Les missions scientifiques couvrent un large spectre de domaines. L'étude de l'Univers y tient une large place, avec en premier lieu les différentes disciplines de l'astronomie (astrophysique, astrochimie, astrométrie, ...) et l'exploration du système solaire et de ses

différentes planètes encore fort mal connues. L'observation du soleil et des phénomènes qui s'y produisent, l'étude des effets de ces phénomènes sur notre planète et son environnement ionisé en font également partie. Un autre volet très important est celui de l'étude et de l'observation de la Terre, de sa forme et de sa surface (géodésie), de ses océans, de son évolution à moyen et long terme dans tous les domaines. D'autres domaines importants de la science donnent lieu à des expériences de chimie, de biologie, de médecine en apesanteur. Certaines enfin, et sans pour autant que la liste soit close, visent à vérifier à l'échelle des distances interplanétaires, des théories fondamentales de la physique comme celle de la relativité.

Toutes les observations sont réalisées à l'aide d'instruments complexes qui ont été conçus à cet effet, et qui fournissent des mesures sous forme de données numériques. Une caractéristique particulière d'un nombre important de missions est la capacité de ces instruments à effectuer des observations systématiques de longue durée et à fournir des données de manière continue pendant un certain nombre d'années. C'est ce qui explique les très grands volumes de données générées par cette activité. Ces données sont transmises au sol par l'intermédiaire d'un signal électromagnétique codé qu'on appelle la télémesure. Depuis le début des années 1980, les données produites par les instruments sont mises sous forme numérique à bord même du véhicule spatial puis transmises au sol.

Le besoin de conserver la majeure partie de ces données à long terme répond à deux impératifs : un impératif scientifique et un impératif patrimonial

### 3.1.2. Question du stockage et des formats d'enregistrement

Au début des années 1970, les données ont d'abordété enregistrées sur des bandes magnétiques 7 pistes (6 pistes pour les données et une piste dite de parité pour le contrôle des erreurs de bit) avec des densités d'enregistrement de 200, 556 puis 800 bpi (bits par pouce). Ces bandes ont disparu très vite et ont été remplacées par des bandes 9 pistes (8 de données plus une pour la parité) en relation avec le développement des jeux de caractères à huit bits (comme par exemple l'EBCDIC d'IBM). Les densités des bandes 9 pistes ont évolué de 800 bpi à 1600 bpi puis 6250 bpi jusque vers la fin des années 1980. Au-delà de cette période, cette technologie a cessé d'évoluer et a pratiquement disparu à la fin des années 1990. Les capacités d'enregistrement étaient faibles au regard de nos support actuels : de 15 Mo pour la bande 800 bpi à 150 Mo pour la bande 6250 bpi.

En 1990, toutes les données issues des missions scientifiques du CNES étaient stockées sur plusieurs dizaines de milliers de bandes entreposées dans les locaux du centre de calcul du CNES à Toulouse. Presque toutes ces bandes et les fichiers de données qu'elles contenaient avaient la structure propriétaire liée aux puissants (pour l'époque !) ordinateurs Control Data du centre. Chaque collection de bandes magnétiques était gérée, contrôlée, entretenue par les équipes projets propriétaires des données contenues dans ces collections.

C'est à cette époque que la situation et les perspectives à moyen terme ont changé de manière significative en raison de trois facteurs : le premier et le plus contraignant a été la disparition annoncée des technologies de stockage sur bandes magnétiques 6250 bpi. En second lieu, en relation avec l'évolution générale de l'informatique en milieu scientifique, le CNES a planifié l'arrêt des machines Control Data du centre informatique, basées sur le système d'exploitation NOS/VE et leur remplacement par des machines basées sur le système UNIX. Enfin, il y avait la volonté de rendre les données scientifiques accessibles et utilisables par la communauté la plus large.

### 3.1.3. Mise en place du STAF et migration des données

Face à cette situation, le CNES a fait le choix, dès 1992, de la mise en place d'un service central de stockage apportant une véritable garantie de conservation à long terme des bits, quelle que soit la technologie de stockage utilisée. Ce service spécialisé en charge de pérenniser les fichiers est le STAF. Il est opérationnel depuis 1994 et se présente comme une entité indépendante des projets ou des services d'archive. Ces derniers sont les clients

du STAF et s'adressent à lui au moyen d'un ensemble de commandes de base permettant notamment de demander le stockage ou la restitution d'un fichier ou d'un ensemble de fichiers. Ces communications passent par le réseau interne du CNES. Le STAF a donc une mission très simple :

- recevoir des fichiers sans avoir à connaître leur format ni leur contenu informationnel,
- assurer la conservation à long terme de ces fichiers,
- garantir leur intégrité,
- garantir leur confidentialité,
- et les restituer à la demande.

La migration immédiate et indispensable des données sur bandes vers le STAF a donc été décidée et entreprise dès 1994. Elle a duré cinq ans, porté sur plus de 60 000 bandes magnétiques, soit environ 500 000 fichiers, et a impliqué une mobilisation momentanée d'une série d'acteurs qui ont pris véritablement conscience de la grande vulnérabilité des données et des multiples causes pouvant conduire à leur perte. Le nombre de fichiers perdus à cette occasion pour des raisons liées au stockage a été très faible mais non nul en raison de l'existence de quelques bandes ayant une densité obsolète (800 bpi) pour lesquelles le CNES n'avait plus d'équipement de lecture, et de rares bandes illisibles à cause de leur dégradation physique.

La migration a immédiatement révélé que la plupart des données présentaient des structures logiques et des encodages propres aux systèmes d'exploitation qui avaient été utilisés pour créer ces données, systèmes d'exploitation eux-mêmes en voie de disparition. En conséquence, les fichiers n'étaient pas portables et donc non lisibles sur un autre système. Une première tâche a donc consisté à débarrasser les fichiers de toutes les informations supplémentaires propres au système d'exploitation. Cela a pu être réalisé à l'aide de logiciels utilitaires disponibles dans le système d'exploitation lui-même.

Une seconde opération beaucoup plus délicate a dû être entreprise. Dans la plupart des cas, les fichiers contenaient des résultats de traitement scientifique sous forme de suites de nombres entiers et réels codés en binaire. La représentation binaire des nombres réels de très haute précision, d'une taille de 128 bits était une représentation propriétaire. Le CNES a fait le choix d'utiliser une représentation standard des nombres (notamment IEEE pour les nombres réels), ce qui a nécessité une transformation des données. Il s'agit donc ici de transformations ou migrations de format. Ces transformations présentaient plusieurs difficultés :

- elles ne pouvaient être exécutées de manière automatique et impliquaient un développement logiciel spécifique pour chaque collection de fichiers,
- elles n'étaient pas réversibles en raison des inévitables erreurs d'arrondi portant sur les derniers chiffres significatifs.

Elles exigeaient donc un effort de validation considérable. Pour plus de sécurité, les fichiers d'origine ont été conservés quelques années puis détruits.

En outre, les éléments descriptifs de ces données, permettant d'en connaître la signification, étaient parfois inexacts ou incomplets, voire pas toujours disponibles, Il a donc été nécessaire, au cours d'une même opération qualifiée de « réhabilitation des données », de reformater les données pour les doter de structures indépendantes des systèmes d'exploitation, de reconstituer les métadonnées pour autant que cela était encore possible et de migrer ces données vers le STAF. Plusieurs ingénieurs ont consacré plusieurs années à cette opération, ils ont dû avoir recours à des experts scientifiques encore disponibles pour reconstituer les métadonnées, ils se sont appuyés sur des sociétés de service pour les développements des logiciels de transformation de formats, ils ont consommé des ressources machines considérables. L'essentiel des données a été sauvé mais cette opération a permis de mesurer à quel point le fait de prendre des dispositions tardives pouvait coûter cher. Compte tenu du rythme actuel des évolutions des technologies numériques, un tel sauvetage ne serait probablement plus possible aujourd'hui.

D'autres signes de l'accélération de l'obsolescence des technologies ont également marqué cette période. L'exemple des documents textuels issus du domaine de la bureautique est

éloquent. Le CNES a utilisé un premier système bureautique disponible sur le marché dans la seconde partie des années 1980. Il s'agissait d'un système propriétaire constituant l'avatar électronique de la machine à écrire traditionnelle et permettant la saisie, la mise en page et l'impression de documents texte en bénéficiant des possibilités de l'informatique. Des masses importantes de documents ont été saisies à l'aide de ce système. Au début des années 1990, avec le développement de la micro-informatique et les débuts du monopole de Microsoft sur la bureautique, la plupart des autres systèmes propriétaires existants ont disparu du marché, ... mais les documents sont restés. Sans la moindre possibilité technique d'opérer une migration des documents vers le progiciel Word pour DOS qui constituait la première version de Word utilisée au CNES, les documents ont été saisis une nouvelle fois. Six ans plus tard, le CNES a fait le constat que les documents enregistrés sous Word pour DOS n'étaient que partiellement compatibles avec Word 97 pour Windows. Pour l'ensemble des documents qui devaient être conservés, le texte a pu être récupéré mais la mise en page de milliers de tableaux complexes a été entièrement reprise.

Aujourd'hui, les inévitables migrations de support sont réalisées de façon continue par le STAF et ne sont pas visibles des clients du service. En quinze ans d'existence et avec une volumétrie qui s'approche à grands pas du pétaoctet, le STAF n'a pas perdu une seule donnée, il a démontré l'intérêt et l'efficacité des principes sur lesquels il a été construit, à savoir une totale indépendance de la fonction de stockage par rapport aux autres entités fonctionnelles d'un service d'archivage long terme, il met en pratique ce qu'on appelle aujourd'hui la virtualisation du stockage.

### 3.1.4. Des systèmes d'archivage générique pour réduire les coûts

Il est possible, dans une certain mesure, de définir des formats de données et de métadonnées totalement indépendants des systèmes d'exploitation et des technologies. Ce choix limite la vulnérabilité de ces données et métadonnées par rapport aux changements de ces technologies. Par contre, il n'est pas possible de construire un système d'archivage numérique, composé de matériels et de logiciels qui soient indépendants de la technologie. Par ailleurs, sachant que le système d'archivage numérique est le moyen par lequel nous allons pouvoir recevoir les données à archiver, les stocker, les gérer, les rendre accessibles, ce système doit être pérenne. Soumis à tous les aléas des obsolescences technologiques, il conviendra d'assurer la maintenance de ce système pour qu'il reste en fonctionnement permanent. Périodiquement, en raison de la disparition de telle ou telle technologie utilisée, ce ne sont plus des travaux de maintenance mais ce sont des travaux de reconstruction partielle du système qu'il faudra entreprendre et donc financer.

La question de la limitation et si possible de la réduction des coûts de maintien en fonctionnement permanent du système d'archivage est un point critique auquel le CNES a tenté de répondre par deux choix.

- Un choix d'architecture consistant à structurer les systèmes en blocs fonctionnels indépendants les uns des autres de façon à ce que tout changement technologique majeur entraînant des modifications profondes sur un bloc soit sans impact sur les autres. Le CNES avait rencontré dans le passé le cas de systèmes monolithiques dans lesquels un changement limité sur un domaine induisait une propagation en chaîne de modifications sur l'ensemble du système, avec des conséquences importantes en termes de coût de modification et de validation,
- Un choix de généricité visant à construire des systèmes réutilisables par plusieurs applications au sein de l'établissement et par plusieurs organismes ayant des activités d'archivage de données scientifiques. Ce choix vise un partage des coûts de développement, puis de maintenance et d'évolution par les différents sites utilisateurs du système. Jusqu'en 1995, chaque mission scientifique spatiale conduisait au développement d'un système dédié permettant la réception, le traitement, la diffusion et l'archivage des données de cette mission. Le CNES a rapidement fait le constat qu'il serait dans le futur impossible de conserver en état de fonctionnement autant de systèmes que de missions spatiales passées.

Il était donc impératif de construire un système de gestion et d'accès aux données capable d'offrir des fonctions d' accès aux données de toutes les missions d'une même discipline scientifique, voire de plusieurs disciplines scientifiques distinctes. Un premier système générique de ce type, le SIPAD (Système d'Information, de Préservation et d'Accès aux Données), a été développé pour assurer la gestion et la mise à disposition des données du Centre de Données de la Physique des Plasmas (http://cdpp.cesr.fr). Sa première mise en service date de 1999. Ce système était notamment basé sur un produit commercial de gestion des données techniques nommé Métaphase. Après quelques années de fonctionnement seulement, il est apparu nécessaire de résoudre un ensemble de questions techniques liées au SIPAD : maîtriser les performances d'accès à la base de données, performances qui se dégradent en même temps que le nombre de jeux de données augmente, éliminer la dépendance du SIPAD par rapport au produit Métaphase dont la pérennité n'était plus garantie, introduire des fonctions nouvelles dans la perspective d'interrogations automatisées, disposer de possibilités de spécialisation avancée de l'interface homme-machine. L'ampleur des besoins d'évolutions a conduit au développement d'un système entièrement nouveau, le SIPAD-NG. La mise en service en 2006 de ce nouveau système a donc impliqué au préalable de revoir le schéma de la base de données. Les métadonnées ont été globalement extraites du SIPAD, puis transformées et enrichies conformément aux nouvelles spécifications de métadonnées, puis validées et ingérées dans le SIPAD-NG. Le système SIPAD-NG dispose aujourd'hui d'une solide assise : utilisé au sein du CNES par plusieurs entités distinctes en charge de l'archivage de données, utilisé également au sein de plusieurs autres organismes de recherche comme le CNRS ou l'IFREMER (Institut français de recherche pour l'exploitation de la mer), il commence réellement à répondre à ce besoin de disposer d'une assise de sites utilisateurs du système, entre lesquels les coûts de maintenance et d'évolution sont partagés.

### 3.1.5. Une méthodologie qui se consolide

Les travaux de sauvetage et de « réhabilitation des données » ont été accompagnés d'une analyse méthodologique sur ce qu'il convenait de faire et de ne pas faire dans le futur. Un premier document de spécification de l'archivage long terme des données spatiales a été rédigé et diffusé en juin 1993. Il souligne l'importance du patrimoine de données scientifiques et technologiques conservé et maintenu depuis le début des années 1970 et observe trois évolutions essentielles :

- l'augmentation constante des volumes de données produites,
- l'accroissement important des durées de conservation minimales requises (plusieurs dizaines d'années),
- les besoins d'une accessibilité de plus en plus large à ces données par la communauté scientifique.

Cette première spécification contient déjà toutes les exigences essentielles appliquées aujourd'hui aux données scientifiques :

- l'identification de l'ensemble des informations qu'il convient d'associer aux données cibles de l'archivage : description syntaxique et sémantique des fichiers, paramètres d'étalonnage, ... (Informations de représentation), métadonnées au format DIF (Directory Interchange Format) (Information de description), base documentaire descriptive de la mission, de l'expérience, de l'instrument (Information de provenance et de contexte),
- l'exigence de l'intégrité physique de l'ensemble des informations numériques à conserver,
- l'exigence de modes de codage normalisés et de structures de fichiers indépendantes des systèmes d'exploitation,
- l'accessibilité des données aux utilisateurs autorisés.

C'est sur la base de ses expériences pratiques et de ses premières réflexions méthodologiques que le CNES a pu participer de façon fructueuse à la rédaction du modèle OAIS et qu'il a pris ensuite la responsabilité directe de la rédaction de la norme PAIMAS et de la future norme PAIS. Le modèle OAIS permet d'analyser les systèmes développés au

CNES avec un point de vue et un vocabulaire nouveau. Ce sont ces itérations entre l'approche pragmatique née de l'expérience de terrain et la réflexion méthodologique qui confèrent peu à peu la robustesse et la fiabilité nécessaires aux systèmes d'archivage numériques.

C'est également sur cette base qu'une nouvelle branche intitulée « Ingénierie des données » a été ouverte au sein du Référentiel Normatif du CNES qui définit l'organigramme des normes qui sont applicables à ses projets et à ses structures. Outre une vision de synthèse des besoins en matière de pérennisation et d'accès aux données, cette branche du Référentiel comporte un certain nombre de règles et de recommandations applicables aux projets producteurs et données ainsi qu'aux services en charge d'archiver ces données.

### 3.1.6. Distorsion possible entre les besoins et les décisions

Un plan stratégique du CNES pour la période 2001-2005 a été élaboré courant 2000. Ce plan prenait en compte, dans ses grandes lignes, la problématique de gestion, diffusion et valorisation des données issues des expériences spatiales. Ce plan a donné lieu à un travail de déclinaison de ses grandes orientations en actions concrètes.

La première action proposée dans ce cadre était plus que symbolique : « Préparer une décision à la signature du Président du CNES pour affirmer les responsabilités et les objectifs du CNES pour la valorisation, l'archivage et la mise à disposition des produits (de données) ».

Une telle décision n'a pas encore vu le jour et ce plan stratégique n'est plus tout à fait d'actualité. Cette situation illustre, si besoin était, l'importance qu'il y a à convaincre les dirigeants et décideurs de l'urgence du problème.

### 3.1.7. Conclusions sur le retour d'expérience au CNES

Pour ce qui concerne la mise en œuvre de l'archivage numérique des données scientifiques spatiales, nous pouvons proposer les conclusions provisoires suivantes :

• en matière de stockage, nous considérons que jusqu'au niveau du pétaoctet, les besoins de préservation physique des fichiers sont résolus avec un niveau de fiabilité satisfaisant,

• la description sémantique des données et l'élaboration des métadonnées descriptives constituent les éléments clés de la réutilisabilité des données dans le futur. Ces métadonnées sont encore dépendantes d'ontologies et de terminologies en forte évolution. Il s'agit ici d'une vraie difficulté qu'il convient de ne pas ignorer,

• la pérennité des systèmes informatiques développés pour le versement des données, leur gestion et leur diffusion pose des problèmes d'une autre nature. Le défi ici est de maîtriser et de minimiser les coûts de développement, de maintenance et d'évolution de ces systèmes. Nous avons vu comment cela pouvait être envisagé,

• en ce qui concerne les fichiers de données, on peut observer que dans les disciplines scientifiques où un format standard des fichiers de données a pu émerger, les outils libres de traitement, d'analyse, de visualisation et les services à valeur ajoutée se sont rapidement développés. À l'inverse, les disciplines pour lesquelles aucun format n'a réellement émergé sont fortement pénalisées. D'où la nécessité pour ces disciplines d'entreprendre ou d'accélérer le travail dans ce sens.

Au plan politique et plus précisément des décisions de mise en œuvre, la situation reste incertaine et l'archivage des données n'est que partiellement couvert. Les orientations du CNES en matière d'archivage long terme de toutes les données spatiales pour lesquelles ce besoin existe, restent à expliciter et à officialiser.

Le CNES ne dispose pas de référentiel central de l'ensemble de ses données et ne sait donc pas rendre compte de manière complète de son patrimoine. En outre, rien ne permet de penser que toutes les données issues des projets du CNES sont effectivement archivées. Certaines données ne sont sous la responsabilité d'aucun centre d'archivage identifié, ce qui

n'augure évidemment pas d'une quelconque garantie de pérennité. D'autres encore sont définitivement perdues comme cela a été le cas pour certaines missions du passé. Même si un projet de constitution d'un référentiel global commence à voir le jour, plusieurs années seront nécessaires avant de parvenir à une situation gérée et maîtrisée.

De nombreuses missions spatiales sont en pratique organisées dans un cadre de coopérations internationales. Il n'est alors pas possible à une agence spatiale d'imposer ses standards à toutes les autres. Ceci renforce – si besoin était – la nécessité d'une coopération aussi étroite que possible entre les agences sur l'Archivage des données, d'autant que la communauté des utilisateurs est elle aussi internationale.

# B. 3.2. L'expérience des services publics d'archives français

Les Archives nationales, site de Fontainebleau puisque, dans le cadre du programme Constance, elles reçoivent, contrôlent et conservent et communiquent des archives numériques depuis le début des années 1970, suivant une méthodologie éprouvée, mais surtout adaptée à la réception de fichiers statistiques structurés . A l'heure actuelle, l'archivage numérique devient une priorité avec le développement de l'administration électronique, la production d'originaux numériques avec utilisation de la signature électronique, qui rendent indispensables le développement de nouveaux outils intégrant des automatismes plus nombreux, l'utilisation de langages XML, ainsi que des fonctionnalités permettant d'assurer l'intégrité des fichiers et leur pérennité (identification, contrôle et conversion si nécessaire des formats, duplication des archives sur des sites distants, garantie d'une traçabilité forte...).

## 3.2.1. Le contexte

Voici dans quel contexte les expériences ont pu être réalisées.

Même si l'environnement législatif concerne autant les archives papier que les archives sur tout type de support depuis la loi du 9 janvier 1979 sur les archives, le numérique n'a longtemps pas été une grande préoccupation des archivistes, à l'exception notable des Archives nationales, site de Fontainebleau qui, depuis une trentaine d'années, archives des données provenant d'institutions scientifiques et notamment des bases de données statistiques de grands organismes scientifiques.

En effet la production massive des organismes publics restait sur support papier pour plusieurs raisons, la principale en étant le contexte juridique qui exigeait que, pour qu'un acte reçoive une valeur probante, il devait faire l'objet d'un écrit revêtu d'une marque de validation (tampon, signature) manuscrite. Par ailleurs, l'interopérabilité entre les différents systèmes d'information développés depuis les années 1970 n'existait pas, ce qui obligeait à une re-matérialisation de l'information, dès lors qu'elle sortait de son environnement de production. Enfin, un grand nombre d'informations provenant de l'extérieur (courriers d'autres administrations, de citoyens) arrivant dans les services sous forme papier, celui-ci restait le « format » commun le plus approprié et le plus simple à généraliser.

Ceci n'empêchait nullement le développement de bases (applications métier) dans les administrations qui, peu à peu, durant ces trente dernières années, ont remplacé les anciens registres papier (registres d'ordre, registres d'enregistrement), par des bases de données de plus en plus riches et complexes, qui permettent de tracer finement le suivi d'une affaire, d'une situation ainsi que d'éditer des courriers ou formulaires les plus fréquemment utilisés dans l'administration concernée. Ces bases de données donnent toujours accès aux dossiers papier correspondants, ces dossiers pouvant contenir, entre autres, des copies d'écran de l'application ainsi bien évidemment que des copies de documents édités à partir de l'application, qui ont reçu des signes de validation manuscrites.

La prise de conscience de la nécessité d'archiver des données extraites de ces applications a été relativement tardive chez les archivistes : en effet, les opérations de versement

concernent généralement des documents de dix ans ou plus, produits dans un environnement exclusivement papier. Par ailleurs, la gestion de ces applications métier sont entre les mains des services informatiques, qui n'étaient pas les interlocuteurs des archivistes qui jusqu'alors, travaillaient exclusivement avec les producteurs des documents.

Une nouvelle prise de conscience est apparue avec le développement de l'administration électronique initiée dans un nouveau contexte juridique puisque, dorénavant, des originaux numériques pouvaient être produits, pour peu que soit dématérialisé un processus métier. Parallèlement progressait, grâce à l'action de la Direction générale de la modernisation de l'État, l'interopérabilité entre les systèmes d'information qui seule, pouvait permettre cette dématérialisation (voir la partie 10 sur Intégrité, authenticité, preuve).

Parallèlement, la numérisation en masse des archives patrimoniales papier a posé la question de la conservation des fichiers constitués dans des formats dits de conservation, non compressés. Le premier type de réponses a été la gravure de ces fichiers sur des CD-R en double exemplaire qui ont été remis aux différents services d'archives, qui ont ensuite rangé ces supports dans les magasins traditionnels utilisés pour les archives papier. Une prise de conscience est venue tardivement sur la nécessité de surveiller ces supports, de les tester grâce à des échantillons représentatifs et si nécessaire de les migrer vers d'autres CD-R, ainsi que sur la nécessité de maîtriser les modes de production de ces CD-R. Des circulaires de la direction des Archives de France ont encadré et précisé ces modes d'intervention, de contrôle, de conservation mais une partie du stock reste à ce jour non surveillée. Un certain nombre de services ont ensuite modifié leur stratégie, notamment lorsque les volumes devenaient très importants en décidant de conserver ces fichiers sur des serveurs de stockage (sur disques, sur bandes type LTO3 ou 4), qui sont mis à leur disposition par les directions informatiques suivant des modalités variées et qui obéissent généralement aux modes opératoires des services informatiques. Un mode d'hébergement externalisé se développe également. Il est évident que ces différentes stratégies demanderaient à être précisées et encadrées, de manière à ce que les exigences en matière de conservation pérenne soient bien respectées.

## 3.2.2. Le projet PILAE de la Direction des Archives de France

Le projet PIL@E a été lancé à la fin de l'année 2006 sous la conduite du département de l'innovation technologique et de la normalisation de la DAF, en bénéficiant d'une assistance à la maîtrise d'ouvrage précieuse à la DGME (Direction générale de la modernisation de l'Etat).

### 3.2.2.1. Contexte

Ce projet a bénéficié pour son lancement et durant toute sa réalisation d'un appui très fort de la part de la directrice des Archives de France, ainsi que de la directrice des Archives nationales.

Ce projet a en effet été considéré comme stratégique en raison des enseignements essentiels qui en seraient tirés pour la communauté archivistique dans son ensemble qui doit dans les années à venir, surmonter le défi de la nouvelle production nativement numérique et adapter en vertu de cela, ses méthodes de travail, ses compétences, sa formation, ... Ce caractère d'innovation rend parallèlement ce projet difficile à mener et à réussir et pour le maître d'ouvrage, et pour le maître d'œuvre et enfin pour les utilisateurs archivistes de l'outil, la conduite du changement étant particulièrement lourde à mener.

Autre particularité spécifique : l'absence de mesures et de connaissances réelles sur les volumétries à venir, tant la dématérialisation complète des processus métier est encore dans la majorité des cas, en cours ou en projet.

Il a été décidé que le pilote serait mis en production au sein du service des archives électroniques des Archives nationales, site de Fontainebleau, afin de recevoir, conserver et communiquer les archives nativement numériques produites par les services centraux de l'État, durant une période transitoire (2009-2013). En effet, les Archives nationales, au travers du projet de nouveau centre des archives à Pierrefitte-sur-Seine, ont entrepris un

très vaste projet de rénovation et refonte de leur système d'information pour la gestion, la communication, la description et la diffusion des archives papier des services centraux de l'État. À terme, une interface devra être trouvée entre ce système et la plate-forme d'archivage électronique qui aura remplacé PIL@E visant à permettre aux internautes de rechercher et de consulter les archives et les instruments de recherche tant papier qu'électroniques.

PIL@E devait par conséquent pouvoir recevoir, conserver et permettre la recherche et la consultation de différentes natures d'objets nativement numériques : données extraites de bases de données métier, documents issus de gestion électronique de documents, d'intranets collaboratifs, messages électroniques issus d'outils de messagerie, flux de données et de documents issus d'une chaîne de dématérialisation complète. PIL@E devait par ailleurs reposer sur le modèle OAIS ainsi que sur les recommandations et prescriptions du référentiel général d'interopérabilité notamment en matière de formats de documents pris en charge, de politique d'archivage et de formats de métadonnées (standard d'échange de données pour l'archivage).

#### 3.2.2.2. Fonctionnalités du coffre-fort électronique

Le maître d'œuvre retenu proposait de faire reposer l'application sur le progiciel qu'il commercialisait, à savoir le coffre-fort électronique communiquant (CFEC) qui assurait un certain nombre de fonctions couvrant certaines des fonctionnalités requises en matière de stockage sécurisé et de traçabilité :

- fonctions de vérifications d'empreintes, de signatures de fichiers, d'horodatage en se basant sur une source de temps externe « sûre », et fonctions permettant des vérifications régulières de ces paramètres dans le temps,
- fonctions assurant la traçabilité du système : édition de journaux d'événements permettant de suivre les différentes opérations de connexion à l'outil, de versements des objets à archiver, de consultation, d'éliminations ; journaux eux-mêmes scellés régulièrement de manière à ce que l'on ne puisse pas modifier des informations sur ces journaux,
- fonctions permettant de piloter les opérations de réplications synchrones des données, à chaque fois qu'un versement est accepté et validé par l'archiviste : écriture sur deux sites distants (réplication inter-sites) doublée d'une réplication intra-sites; soit au total une réplication sur quatre baies de stockage sur quatre serveurs répartis sur deux sites distants. Par ailleurs la sécurité des données est assurée par des systèmes de redondances concernant aussi bien les données elles-mêmes, que les bases de données, les accès, les installations électriques, sans compter un système classique de sauvegarde intégré à la politique générale de sauvegarde assurée par le département des systèmes d'information du ministère.

Pour l'ensemble de ces tâches, le coffre-fort s'interface avec une infrastructure matérielle. Certaines des fonctions généralement assurées dès lors que des grosses volumétries sont en jeu (ce qui n'est pas le cas de PIL@E), ne sont pas proposées : surveillance automatique des supports, système de « Hierarchical storage management ». Un seul type de support a été choisi, à savoir des disques magnétiques en ligne.

Le coffre-fort lui-même repose sur des briques en « open-source », tandis qu'il garantit la réversibilité si nécessaire de l'ensemble des archives, journaux et autres éléments de preuve qui ont été générés, pour restitution et reprise dans un autre système. PIL@E par exemple pourrait à terme fonctionner avec un autre coffre-fort que celui de Cecurity.com.

#### 3.2.2.3. Fonctionnalités développées au-delà du coffre-fort électronique

Au-delà de ces fonctionnalités de base, il a fallu développer l'ensemble des autres fonctionnalités propres à l'archivage sécurisé des données et documents des services producteurs (administrations des services centraux de l'État).

PIL@E a finalement été conçue pour être utilisée par les services d'archives présents dans les ministères (« les archivistes versants ») et par les archivistes des Archives nationales, site de Fontainebleau. L'accès de l'application n'a pas été élargi à l'ensemble des producteurs en raison du caractère novateur de ce nouveau type de versement et de

l'absence de connaissances précises sur les types de versements concernés, la fréquence des versements, leur volumétrie, leur utilisation une fois le versement effectué.

De même, a finalement été écarté l' accès au « grand public » désirant consulter des archives numériques communicables pour une double raison : réflexion non encore assez poussée quant au niveau de sécurité à mettre en œuvre pour ouvrir sur internet une application gérant et permettant l' accès à des données et documents confidentiels ; pertinence de cette ouverture non avérée en raison d'une couverture qui sera très inégale dans un premier temps entre les différents domaines administratifs susceptibles d'être couverts, aggravée en cela par le caractère non encore communicable de certains de ces documents et données.

Le cœur des développements concerne l'implémentation du standard d'échange de données pour l'archivage (SEDA) centré sur les processus « transfert » et « recherche et communication ».

Ont par conséquent été développés des interfaces et messages relatifs aux différents processus mis en œuvre entre ministères et Archives nationales : envoi par les archivistes versants, réception par les archivistes (processus de contrôle/validation/rejet), accusés de réception... Le processus de recherche et consultation a été simplifié et ne requiert pas l'action des archivistes des Archives nationales (recherche sur la base des droits d’accès sur les archives à partir des métadonnées, téléchargement).

Concernant le format des métadonnées défini par le standard d'échange de données pour l'archivage, essentiel pour automatiser la description des grandes catégories d'archives versées périodiquement, deux modes ont été prévus pour PIL@E, soit un mode principal consistant à transférer « manuellement » des versements déjà formatés au format du standard d'échange (voir partie 9 sur les métadonnées). Le versement transféré est alors reçu et contrôlé par le système (conformité par rapport au schéma du standard d'échange, vérification d'empreintes, identification des formats), qui envoie en réponse un message d’acceptation ou un message d’erreur, suivant le type de problèmes rencontré.Une fois accepté, le versement est alors contrôlé par un archiviste des Archives nationales qui peut par exemple modifier, enrichir les métadonnées du bordereau de versement et qui décide alors de valider ou non le versement. Si celui-ci est validé, les données/documents sont stockés comme dit précédemment sur plusieurs serveurs (réplication) et les métadonnées alimentent une base de données afin de permettre ensuite les recherches et les consultations.

Ce mode de fonctionnement (import de fichiers au format du standard d'échange) est évidemment le seul pertinent pour éviter les ruptures de charge et notamment la re-saisie de métadonnées, impossible dès lors qu'on a des grosses volumétries de fichiers à décrire. Ce mode est particulièrement intéressant dès lors qu'on doit effectuer des versements périodiques d'une même catégorie d'archives, le travail de spécifications en amont pour mise au format du standard d'échange ayant été effectué une seule fois.

Un autre mode dégradé est possible dans PIL@E qui consiste à saisir manuellement les métadonnées de fichiers que l'on souhaite transférer grâce à des écrans de saisie reprenant l'ensemble des champs du bordereau de versement défini par le standard d'échange. L'archiviste versant doit déterminer, comme pour des archives papier, quel sera le plan de classement des archives transférées (détermination des différents niveaux de description) et saisir les métadonnées associées à chaque niveau.

On comprend évidemment qu'il s'agit d'un travail fastidieux qui devra être renouvelé à chaque nouveau versement. Bien évidemment la description concernera alors tout cet ensemble et sera donc sommaire. Lorsque cette saisie est terminée et que l'archiviste versant a joint les fichiers de données concernés, le système effectue lui-même la mise au format du standard d'échange. Le processus est ensuite identique.

Ces deux modes coexistent dans l'application car on ignorait le nombre de versements pour lesquels il aurait été possible d'effectuer ce formatage en amont, par rapport à ceux pour lesquels ce formatage préalable ne serait pas possible, comme par exemple les fichiers bureautiques produits à partir des postes de travail des agents hors d'une gestion électronique de documents.

#### 3.2.2.4. Question des conversions de formats

C'est sur la base d'une convention donnée pour la prise en charge d'une catégorie d'archives, que les archivistes versants vont effectuer leur versement. Cette convention précise notamment si le service producteur autorise ou non les conversions de formats qui se feront à l'entrée dans PIL@E, si les formats initiaux ne sont pas conformes au référentiel général d'interopérabilité, concernant leur préservation sur le long terme. PIL@E intègre des outils d'identification fine des formats et selon les règles fixées, procède ou non à la conversion des formats (par exemple des fichiers issus de la suite Microsoft Office vers le format PDF/A). Les traitements s'effectuent en mode interactif ou traitements différés, dès lors que de gros volumes sont concernés.

Les archivistes qui contrôlent seront avertis si des anomalies concernant les conversions sont constatées mais ils peuvent, même en cas d'échec (aucune conversion n'a pu être réalisée) décider tout de même d'accepter le versement. Les formats d'origine sont bien évidemment conservés dans l'AIP tandis que l'utilisateur qui souhaite les commander ultérieurement, pourra choisir le format d'origine ou le format après conversion.

Toutefois, il est évident que des tests poussés devront être conduits, en-dehors de PIL@E, suivant les types de formats, de manière à pouvoir spécifier plus précisément les « risques » que font peser certaines conversions et les stratégies à tenir suivant ces risques. Des impasses ont été constatées pour les documents graphiques par exemple, pour lesquels on s'est aperçu au cours du projet, que les formats cibles préconisés généralement, ne permettaient pas de conserver certaines des fonctionnalités (liens notamment) des formats d'origine malheureusement propriétaires.

PIL@E tente de résoudre cette problématique de conversions automatiques à l'entrée du système car aucune conversion n'est rendue possible ensuite, si ce n'est à régénérer un autre versement qui remplacera le premier versement.

#### 3.2.2.5. Conclusions sur PILAE

PIL@E devrait être achevée et mise en production au début de l'année 2010. Les années de transition qui suivront permettront de tester le produit et de bien préparer le passage vers un système de plus grande ampleur.

Plusieurs questions devront être étudiées plus précisément afin d'anticiper ces évolutions : on a vu qu'il convenait de préciser les questions relatives aux formats d'encodage des documents. Une réflexion sera à mener relative aux identifiants des SIP/AIP/DIP (nécessité d'utiliser des identifiants uniques et pérennes). Les problèmes afférents à la sécurité des données se poseront : besoin ou non d'un coffre-fort pour l'ensemble des données et documents ; besoin ou non de chiffrements de données particulièrement sensibles ; architecture à repenser lorsqu'on ouvrira le système sur internet, en direction du public.

L'infrastructure de stockage devra s'affiner et inclure d'autres fonctionnalités dès lors que les volumes augmenteront (surveillance automatique des supports, réflexion sur les types de média...).

Les formats respectifs des SIP, AIP et DIP sont actuellement à peu de choses près les mêmes : il sera intéressant de confirmer ou d'infirmer ce choix, notamment concernant les métadonnées techniques actuellement très peu intégrées dans PIL@E. Il conviendra probablement d'utiliser un format d'empaquetage normalisé comme METS (voir partie 9 sur les métadonnées).

Enfin, une fois les AIP constitués, il n'est aujourd'hui pas possible, sauf en choisissant des solutions de contournement, de les modifier (ajout de métadonnées, ajouts de fichiers).

Enfin, n'entre pas à proprement parler dans PIL@E, la « filière » des importants fonds d'archives numérisées à partir de fonds patrimoniaux papier.

Il est normal que toutes ces questions se posent lorsqu'il s'agit de mettre en œuvre un pilote, d'autant que le champ est innovant et que les différents acteurs du marché concernés vivent encore dans des environnements cloisonnés : les éditeurs de coffres-forts spécialisés dans la sécurité et la signature électronique, les éditeurs de solutions dans le marché du stockage, les éditeurs de logiciels de gestion électronique de documents versus « Records management » plutôt présents dans le monde de l'entreprise et enfin les éditeurs

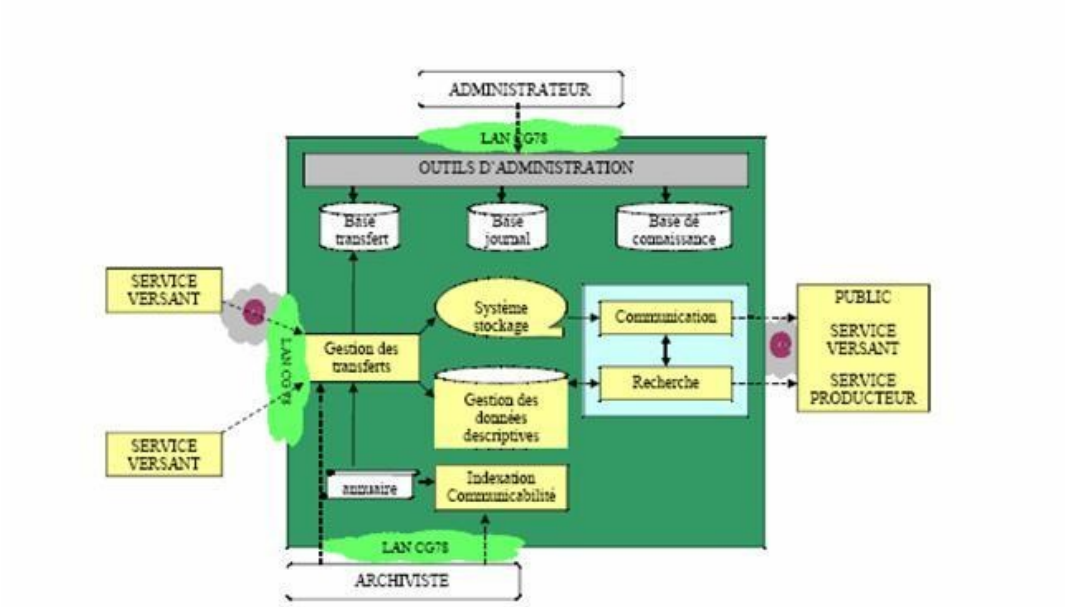
de logiciels dits d'archives, plutôt présents dans le monde patrimonial.

Enfin, une grande inconnue subsiste concernant PIL@E qui est le mode d'appropriation qui en sera fait par les archivistes des ministères et des archives nationales, pour lesquels l'archivage électronique dans ses derniers développements (administration électronique) est un domaine à la fois familier (les principes de l'archivistique restent identiques) et tout à fait nouveau (appropriation des modes de travail type assistance à maîtrise d'ouvrage auprès des producteurs et des informaticiens, appropriation des savoirs en matière de formats et de langages XML).

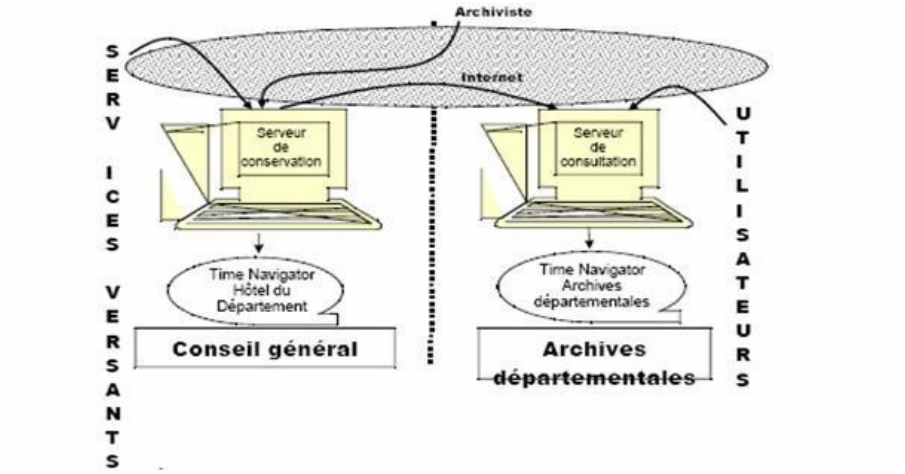
### 3.2.3. La plateforme d'archivage électronique du conseil général des Yvelines

Il s'agit de la première collectivité territoriale en France qui ait mis en place une plateforme pilote, le projet ayant démarré en 2005, dans un contexte de mise en œuvre expérimentale dans ce département, de la transmission dématérialisée des actes soumis au contrôle de légalité (voir partie 10 Intégrité, authenticité, preuve), la plateforme de télétransmission assumant cette transmission devant, une fois la transaction achevée et les voies de recours dépassées, transmettre pour archivage les actes.

Le projet a été mené en interne au conseil général, avec une collaboration très forte entre les services informatiques et les Archives départementales. Le cœur des fonctionnalités de la plateforme est très proche de celles du projet PILAE, le projet des Yvelines s'inspirant à la fois du modèle OAIS, de l'étude sur les coûts des plateformes menée par la direction des Archives de France (voir la partie 11 Organisation et processus) et enfin des spécifications fonctionnelles de PILAE. Les fonctionnalités sont cependant concentrées autour de la réception, contrôle et validation d'archives déjà au format du standard d'échange de données pour l'archivage, sans prise en compte encore des fonctions d'identification, contrôle et conversion éventuelle des formats des documents.




*Les fonctionnalités de plateforme d'archivage électronique des Yvelines*




*Architecture du stockage*

Ce projet est intéressant à plus d'un titre et notamment il a eu pour effet de repositionner les Archives départementales beaucoup plus tôt dans la chaîne du cycle de vie des documents. En effet, la réception des documents se fait dans ce cas très rapidement après la production (dès validation de la transaction) sans attendre l'expiration de la durée d'utilité administrative. Par ailleurs il est projeté qu'une forme de mutualisation soit offerte aux communes qui, ne disposant pas de plateformes d'archivage électronique, désireraient déposer les actes dématérialisés qu'elles produisent.

## 3.2.4. La plateforme d'archivage électronique du conseil général de l'Aube

Il est très intéressant également de constater qu'une autre forme de mutualisation a eu également lieu dans ce cas puisque le conseil général de l'Aube a récupéré l'application développée à partir d'outils libres qui a été développée par le conseil général des Yvelines, avec quelques aménagements et évolutions fonctionnelles.

### 3.2.4.1. Contexte et enjeux

Le contexte est identique à celui des Yvelines, avec la montée de plusieurs projets d'administration électronique et la mise en place par le conseil général d'outils de signature électronique qui sont offerts aux différentes communes et autres collectivités du département. Il s'ensuit que certains documents sont désormais signés électroniquement et constituent de fait des originaux électroniques qu'il convient de conserver. D'où une certaine urgence de développer des outils d'archivage électronique.

Par ailleurs le contexte est très favorable avec la mise en place d'une collaboration très étroite entre les services informatiques et les archives départementales, par le biais d'un comité d'archivage électronique, devant lequel passent tous les projets de dématérialisation dont l'archivage doit impérativement être pris en compte dès le lancement du projet.

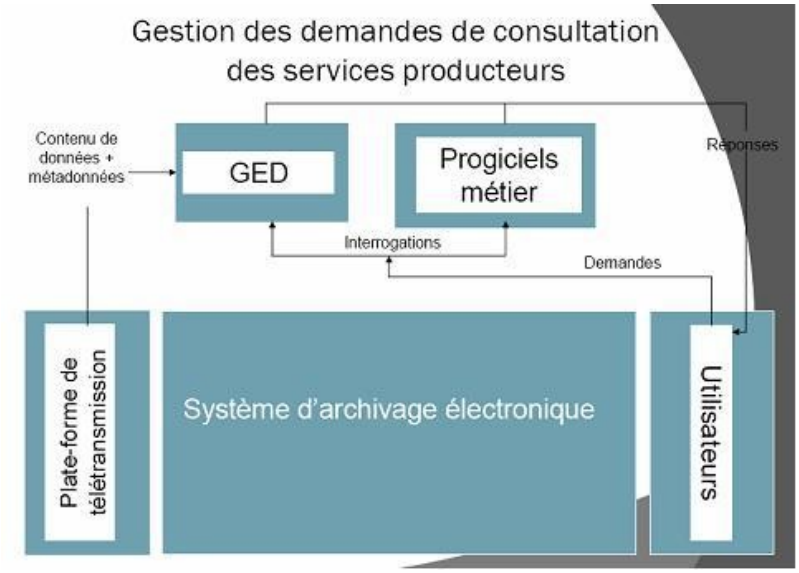
Une réflexion approfondie tant organisationnelle que juridique et technique a été menée.

• Décision pour toute application métier, de déposer les données/documents/flux dans la plateforme d'archivage, dès que les transactions sont terminées, afin de sécuriser l'information, les données étant par ailleurs conservées dans les applications métier durant leur durée d'utilité administrative. Il s'agit d'un changement capital dans l'organisation jusque là en place, puisque seules les archives dites définitives étaient prises en charge par les Archives départementales, une fois la durée d'utilité administrative expirée. Il s'agit pour les Archives d'intervenir très en amont du processus du cycle de vie de l'information et d'avoir ainsi une lisibilité et une responsabilité accrue.

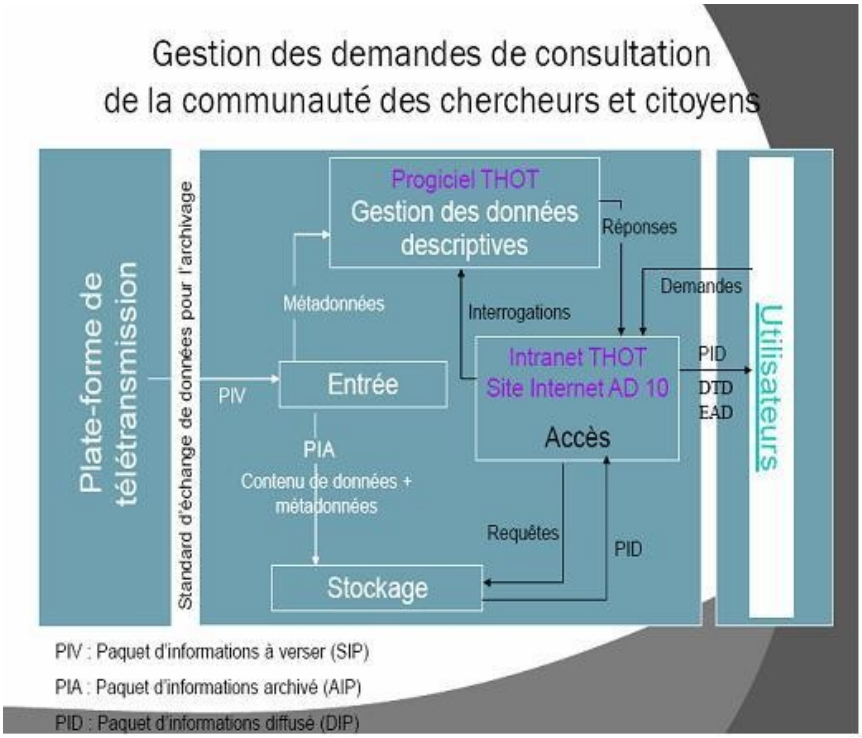
• Mise en place d'une contractualisation systématique par la passation pour chaque nouvelle catégorie documentaire prise en charge, d'une part d'un contrat de service entre le service producteur, le service d'archives et la direction informatique et, d'autre part, d'un protocole de versement cette fois avec le service versant.

• Accès aux archives déposées (à noter : un même accès est mis en place pour les archives quel que soit leur support) :

- pour les services producteurs, via leurs applications métier ou une gestion électronique de documents simple mais adaptée à leurs besoins
- et pour les autres utilisateurs (lecteurs des Archives), via le logiciel de gestion des archives utilisé jusqu'alors pour la recherche et l'accès aux archives papier (le logiciel Thot).




*Gestion des demandes de consultations pour les services producteurs (via leur GED/Gestion électronique des documents métier) et pour les chercheurs et citoyens, via le logiciel de gestion documentaire utilisé par les Archives départementales de l'Aube tant pour les archives papier que pour les archives numériques*




*Gestion des demandes de consultations pour les services producteurs (via leur GED/Gestion électronique des documents métier) et pour les chercheurs et citoyens, via le logiciel de gestion documentaire utilisé par les Archives départementales de l'Aube tant pour les archives papier que pour les archives numériques*

#### 3.2.4.2. La plateforme de confiance

Autre innovation : la mise en œuvre en amont de la plateforme d'archivage, d'une plateforme de confiance qui vise :

- à vérifier la signature des documents transférés qui sont signés,
- à attester de cette vérification,
- à éditer un rapport de vérification qui est lui-même signé,
- à signer le bordereau de versement, une fois le rapport émis.

Sur la plateforme seront également mis en œuvre l'identification et la conversion si nécessaire des formats des documents transférés.

### 3.2.5. Le retour d'expériences d'une grande institution patrimoniale : la Bibliothèque nationale de France (BnF)

La préoccupation liée à la pérennisation de l'information numérique depuis quelques années est aujourd'hui partagée avec les autres grands acteurs patrimoniaux qui ont à prendre en charge des volumétries considérables d'informations numériques, comme l'Institut national de l'audiovisuel (INA) et la bibliothèque de France (BnF).

L'exemple présenté ci-dessous en témoigne largement.

#### 3.2.5.1. Contexte du numérique à la BnF

La diffusion de matériaux numériques est déjà une vieille aventure pour la BnF : plus de 15 ans de programmes de numérisation, activité aujourd'hui en fort accroissement avec la numérisation de masse, la consultation de documents audiovisuels sous forme numérique, l'accès aux premières archives de la toile, etc.

Ces projets ont conduit à la mise à disposition d'un nombre croissant de services : les postes de consultation en interne, la bibliothèque numérique Gallica dès 1997, la vente de reproductions numériques à la demande, la possibilité de collecter les métadonnées via le protocole OAI (Open Archive Initiative) permettant de bâtir des outils de recherche

performants, l'identification pérenne des ressources numériques. Ces différents projets montrent une véritable volonté d'être un acteur majeur de la société d'information, de proposer des outils et des services permettant à ses usagers in situ ou depuis leurs navigateurs Internet de mieux appréhender, comprendre, manipuler, collaborer autour des ressources numériques.

Il faut attendre les années 1990-1991 pour voir apparaître les premières expérimentations de numérisation d'images fixes et d'imprimés. Le support d'enregistrement retenu était la bande magnétique DAT fournie en double exemplaire par les prestataires de numérisation. Ces supports, plutôt destinés à la sauvegarde, avaient une durée de vie de 2 à 3 ans. Avec des bandes produites en 1993, la question de la migration s'est posée avant même la fin du projet.

Plusieurs stratégies ont alors été imaginées. Il convient de noter qu'à cette époque la BnF n'avait pas encore pris conscience de la nature et de l'étendue du problème. L'archivage était uniquement considéré comme un centre de coût et non comme une assurance sur les données.

L'équipe cherchait une solution qui nécessite le moins de compétences possible, le moins d'interventions possible et qui offre la plus grande facilité de lecture sur le long terme. La solution recherchée devait permettre de pouvoir relire les données à l'horizon 10-20 ans, voire plus. Il s'agissait d'une solution d'archive morte.

Les solutions disques magnétiques étaient proscrites pour leur manque de fiabilité. Les solutions bandes nécessitaient trop de cycles de migration de rafraîchissement. Seules les solutions optiques et magnéto-optiques semblaient offrir une opportunité satisfaisante. L'inconvénient du support CD-R (disque compact enregistrable) est son instabilité dans le temps. Néanmoins, une solution basée sur une technologie française permettait de résoudre ce problème.

#### 3.2.5.2. Projet de Century Disk et premières migrations

Il s'agissait du Century Disk développé par la société française Digipress, située à Caen. La technique utilisée s'inspire de celle utilisée pour graver les masters qui servent ensuite de matrice pour presser les CD-ROM en polycarbonate. Le support gravé avec cette technique comporte deux principaux avantages : premièrement il est véritablement gravé comme un CD-ROM et non seulement insolé comme avec le CD-R ; deuxièmement, le support utilisé est du verre trempé à la place du polycarbonate. Ces caractéristiques confèrent aux CD gravés une stabilité et une résistance permettant d'envisager d'accéder aux données sur un siècle voire plus. Le procédé de production est similaire à celui utilisé pour graver les circuits imprimés de très petite taille comme les processeurs et nécessite d'être opéré en salle blanche, salle dénuée de toute poussière. Le verre ayant une très importante inertie physico-chimique permet d'obtenir une gravure pratiquement éternelle. Seule la couche réfléchissante peut poser problème à long terme. Néanmoins, il est en principe possible de « re-métalliser » les supports.

Néanmoins, cette technologie était chère – plus de 207 € le Go – ce qui peut sembler astronomique de prime abord. Néanmoins ce coût pouvant être amorti sur au moins 20 ans, le prix chute à 11 € le Go par an, ce qui plus acceptable même pour les critères d'aujourd'hui. À l'époque, l'équipe avait comparé ce prix au coût à investir en termes de ressources humaines et matérielles à mettre en place pour le contrôle et la régénération des supports CD-R tous les 3 ans sur une période de 20 ans. Les coûts étaient semblables. Étant donné le risque de voir disparaître ce qui pourrait être perçu comme une source potentielle d'économie dans un contexte budgétaire toujours plus strict, ce choix apparaissait être le meilleur : payer beaucoup maintenant, ne pas s'en préoccuper pendant 20 ans voire plus.

Le projet était lancé dès 1997 alors que les premières bandes étaient déjà dans une situation à risque. Les bandes étaient conservées en deux exemplaires dans les magasins du nouveau bâtiment de la BnF, dans des conditions d'hygrométrie et de température contrôlées. L'équipe espérait n'avoir aucune perte.

Les hésitations sur la stratégie à adopter durèrent plus de deux ans. Pendant ce temps, les bandes se dégradaient. Parallèlement, en 1997, la BnF lançait son projet de bibliothèque

numérique : Gallica. La première version fut presque entièrement développée par deux personnes en quelques mois, des ingénieurs de grand talent. Dix ans plus tard la nouvelle version, Gallica 2, vient à peine de la remplacer. Gallica nécessitait de disposer des documents en ligne. Il fallut donc migrer les données à partir des DAT.

La technologie choisie par le service informatique fut celle des librairies de disques optiques numériques (DON). Cette migration dura deux ans de 1998 à 1999 et une perte de 2,5 % des bandes fut constatée. Ce constat permit d'accélérer le projet de transfert sur Century Disk. Les bandes illisibles furent recopiées à partir du deuxième exemplaire. Le procédé de lecture des bandes ne permettait pas d'envisager un transfert direct. Il convenait de passer par un support intermédiaire. C'est le CD-R qui fut retenu. La plupart des disques CD-Century furent gravés pendant l'année 2000. 6 % des bandes furent perdues pendant l'opération.

L'accroissement régulier du fonds et l'augmentation rapide de la fréquentation du site Gallica obligea à changer une nouvelle fois l'architecture de stockage. En effet, les librairies de disques DON n'était pas conçues pour soutenir une sollicitation aussi forte. De plus, aucun algorithme de cache disque au niveau des serveurs de consultation ne permettait de ralentir la sollicitation des librairies de DON du fait de la trop grande disparité des demandes de consultation. Aussi, en 2001, une nouvelle migration dut être envisagée. Les données furent alors copiées vers un système de stockage entièrement basé sur du disque. Aucune donnée ne fut perdue lors cette opération.

#### 3.2.5.3. Projet SPAR

*Une nécessité*

En matière de numérisation, de nouveaux projets ont vu le jour, notamment celui de la numérisation de la presse du XIXème. À cette occasion, une réflexion sur les formats de numérisation a été menée. Les fascicules de la presse étant en grand péril – la mauvaise qualité du papier ne permettait pas d'envisager plusieurs manipulations – la BnF a envisagé alors de passer d'une numérisation de diffusion, dont le but est principalement d'obtenir une numérisation facilement diffusable, sans contrainte de conservation forte à une numérisation de conservation dont le but est de garder le document dans un même format le plus longtemps possible, sans altération. Ainsi, le format JPEG – sa compression destructive étant jugée trop risquée – a été abandonné au profit du format TIFF non compressé multipliant par 4 le volume des fichiers créés.

Dans le même temps, les moyens de production de supports argentiques, notamment pour la couleur, ont rapidement disparu. La conservation des documents « couleur » est progressivement passée au numérique. Cette bascule numérique finira par toucher l'ensemble de la filière microforme qui est devenue pratiquement inexistante.

À ces évolutions se sont ajoutées de nouvelles missions comme la collecte du web Français. Les premières collectes du web apparaissent dès 2002 avec le projet de collecte des sites des élections présidentielles : quelques téraoctets seront enregistrés. Dès lors ces collections ne cesseront d'augmenter pour atteindre plus de 150 To en 2008. De même depuis 2005, la BnF expérimente le dépôt légal électronique en substitution du dépôt papier.

Tous ces changements ont eu des conséquences importantes notamment sur les capacités de stockage, obligeant le département informatique à régulièrement les augmenter. Petit à petit, le parc devint de plus en plus hétérogène et difficile à maintenir. En 2004, la situation étant devenue de plus en plus difficile à gérer, le département des systèmes d'information a décidé de lancer un marché d'acquisition d'une infrastructure pour les fonds numériques qui soit évolutive en matière de capacité de stockage (« scalable ») et réponde aux critères de pérennisation en matière d'ouverture, de distributivité et d'indépendance technologique. Ce qui sera appelé SPAR-infrastructure était né.

Apparut alors à la direction de la BnF le besoin d'aller plus loin dans cette démarche et de lancer le projet d'un véritable système de préservation : SPAR (Préserver le patrimoine numérique de la BnF, archiver l'ensemble de ses données et répartir l'accès à celles-ci).

*Grands principes retenus*

Le numérique signifie pour les détenteurs de fonds et les gestionnaires de collections –

généralement des bibliothécaires – une perte de contact avec les matériels à conserver. Il était absolument nécessaire d'établir un climat de confiance et de responsabilité. C'est pourquoi les groupes de travail mis en place dans le cadre du projet, ont convenu de la mise en place :

• des contrats de service qui définissent l'engagement de l'Archive vis à-vis de ses utilisateurs mais également des utilisateurs vis-à-vis de l'Archive ; ainsi, changer le nombre de copies d'une typologie de document ou augmenter la rapidité de diffusion ont un impact sur les contrats de service qu'il convient de budgétiser et donc de faire arbitrer par la direction ; l'idée étant de ne prendre aucun engagement qui ne pourrait être tenu.

• de « l'auditabilité » du système vis-à-vis des normes, principalement de la norme OAIS et des engagements (contrats de service) ; il s'agit de démontrer que le système et les procédures mises en place peuvent être vérifiés régulièrement par des personnes ou des organismes externes aux personnes en charge de l'Archive ; cette transparence est un gage pour les utilisateurs comme pour les responsables de l'Archive de s'assurer du respect des engagements.

Les groupes de travail ont défini qui sont les Producteurs indépendamment du schéma organisationnel actuel de la BnF.

Ils ont retenu huit types de Producteurs qui se distinguent par leurs spécificités techniques et juridiques. Par exemple, les conséquences sur les contrats de service ne sont pas les mêmes si dans le cadre du dépôt légal, la BnF négocie de gré à gré le dépôt électronique ou si, comme pour le dépôt légal du web, elle collecte les sites Internet qui ont un intérêt patrimonial.

Les groupes de travail ont défini les modèles de données et ont retenu :

• le modèle METS pour le format d'empaquetage,

• le modèle PREMIS pour les métadonnées de préservation,

• et le format ODRL pour formaliser les licences d'utilisation des documents :

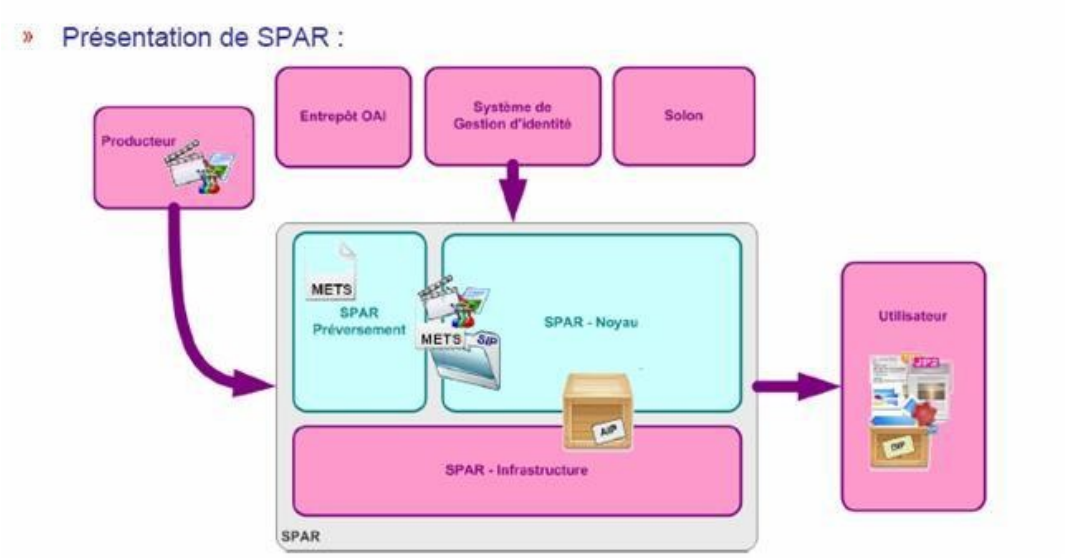
en effet, c'est une particularité de SPAR qui n'est pas couverte de manière précise par le modèle OAIS (cette partie est développée dans la nouvelle version du modèle OAIS en cours d'écriture) mais nécessaire pour gérer les droits de communication des documents ; SPAR n'est pas directement le système de diffusion aux utilisateurs finaux;

les documents sont diffusés aux utilisateurs par des systèmes dédiés : Gallica, la banque d'images, les postes de consultations internes, etc. ; autant SPAR peut s'assurer que le demandeur – généralement un système – est bien habilité à faire cette demande et que le type de document et le format correspondant sont accessibles pour ce demandeur et donc refuser de diffuser un document qui ne respecterait ces règles, autant certaines contraintes de diffusion lui échappent : par exemple, si un document ne doit pas pouvoir être imprimé, SPAR informe le demandeur via la licence ODRL que le document demandé n'est pas imprimable, charge au système de diffusion de mettre en œuvre les moyens qui permettent d'appliquer cette contrainte ; dans les faits, les responsables de l'Archive n'habilitent pas de système pour ce type de document sans s'assurer qu'il est capable de respecter la licence ; la licence est un moyen de contractualiser le transfert de responsabilité.

Les groupes de travail ont établi le nombre de versions qui seraient conservées d'un document. Ainsi, au maximum trois versions seront conservées. La version initiale qui ne sera plus jamais touchée. La version en cours et la version immédiatement précédente. Cette précaution permettra à la BnF de revenir à la version précédente en cas d'erreur constatée après une transformation. Seuls les documents qui ont des contrats de service particuliers – comme c'est le cas pour les archives administratives qui ont des délais de rétention définis – peuvent être détruits.

Le groupe en charge de la gestion des risques a établi une première liste des risques. Chaque risque a été pondéré. Le groupe a tenté de définir une maîtrise, mais constatant que cette tâche de définition était immense, il a également établi des procédures de veille et de résolution de risque. Elles permettront de faire évoluer le référentiel de risques.

**Présentation**




*Présentation globale de SPAR : constitué par le noyau logiciel de SPAR (gestion des données descriptives), l'espace dans lequel les archives sont transférées pour contrôle (SPAR Préversement), l'infrastructure de stockage proprement dite, la gestion des droits d'accès des utilisateurs d'une part et des archives d'autre part, la constitution d'un entrepôt OAI permettant à SPAR d'être moissonné par des ressources externes.*

*Etude de faisabilité*

Il a fallu s'assurer que la mise en œuvre d'un système OAIS était possible pour la BnF. La BnF n'envisageait pas de construire un système de toutes pièces. L'idée était clairement de s'appuyer sur une solution existante qui pourrait être partagée avec d'autres. C'est pourquoi en parallèle des autres instructions et sans présager des options retenues, une étude de faisabilité partant du modèle OAIS a été menée. Cet exercice était délicat car l'offre n'était pas mûre, pas plus qu'aujourd'hui d'ailleurs. C'est pourquoi, il a fallu dans un premier temps explorer les produits existants qui pourraient correspondre au besoin.

Vingt produits ont été évalués selon trois axes : qualité fonctionnelle, qualité technique et pérennité affichée. Ces produits étaient de trois types différents :

• les logiciels de gestion de contenu (content manager) qui pourraient être adaptés aux besoins, aucun n'a été retenu dans cette catégorie,

• les logiciels dits de « institutional repository » tels que DSPACE ou FEDORA qui sont généralement des logiciels dédiés à l'accès mais qui comprennent des mécanismes de versement et de gestion de données intéressants,

• et finalement, les logiciels clairement affichés comme logiciel d'archivage.

De cette première exploration, six logiciels ont été retenus. Ils ont fait l'objet d'une évaluation approfondie basée sur un questionnaire de 132 critères répartis en 9 catégories couvrant les fonctionnalités définies par la norme OAIS mais également des éléments techniques et organisationnels. Cette étude a permis de montrer que des solutions pouvaient répondre aux besoins de la BnF.

L'ensemble de ces travaux a permis la rédaction d'un marché qui fut lancé en 2007. Le système devrait être livré fin 2009.

*Préservation des informations numériques à la BnF*

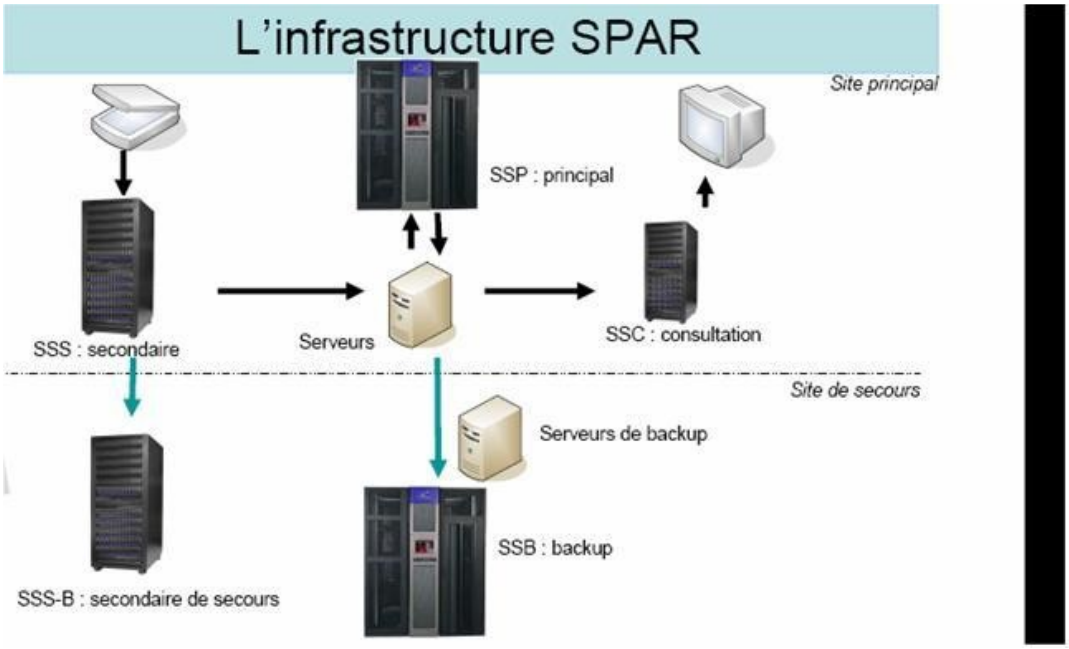
La BnF a la responsabilité de la conservation des savoirs patrimoniaux qui lui sont confiés. Ainsi, elle se doit également de tout mettre en œuvre pour conserver ses collections numériques ; elle doit déployer autant d'énergie à les conserver qu'à conserver ses collections traditionnelles (livres, estampes, globes, monnaies, photographie, etc.). Pour la BnF, cette conservation se place dans une dimension pluri séculaire.

C'est pourquoi SPAR sera bien plus qu'un simple entrepôt de données sécurisé. Le système assurera ce niveau minimum de garantie d'intégrité des données. Il effectuera de multiples

copies – de deux à trois selon la nécessité – et assurera une surveillance continue de l'état des supports d'enregistrement permettant d'anticiper les recopies avant la perte définitive. Mais il permettra également, grâce à une reconnaissance précise et complète des formats de données versées, de garantir la continuité d'accès en procédant aux transformations nécessaires en cas d'obsolescence technologique des outils informatiques de restitution. Ainsi, par exemple, lorsque le format d'image JPEG deviendra obsolète, SPAR sera en mesure de transformer les images concernées dans le format image JPEG de demain. Apporter cette garantie implique un travail permanent de veille technologique sur les formats, de prototypage et de tests des outils, de mise en œuvre et de suivi des transformations.

De plus, SPAR permettra à tout moment de revenir en arrière ou de retrouver les objets initialement versés. Ainsi, c'est potentiellement jusqu'à neuf copies d'un même objet numérique qui sont conservées.

*Infrastructure* (cf image)




*L'infrastructure de stockage de SPAR*

## 3.2.6. L'expérience d'un grand organisme de l'enseignement supérieur : PAC, la plateforme du CINES

Le monde de l'enseignement supérieur et de la recherche à son tour se préoccupe depuis plusieurs années de diffusion, publication et accès à l'information numérique. La préoccupation relative à la conservation du numérique est plus récente comme en témoigne l'un des axes d'actions du très grand équipement (TGE ADONIS) pour les sciences humaines, qui est axé sur la conservation pérenne.

### 3.2.6.1. Centre informatique national de l'enseignement supérieur (CINES)

Le centre informatique de l'enseignement supérieur (CINES) se trouve dans le sud de la France, à Montpellier (Hérault). Il a été créé en 1999, succédant au CNUSC (Centre National Universitaire Sud de Calcul), créé en 1980. L'établissement est placé sous la tutelle de la DGRI (Direction Générale de la Recherche et de l'Innovation) et de la DGES (Direction Générale de l'Enseignement Supérieur) du Ministère de l'Enseignement Supérieur et de la Recherche. Ses deux missions principales sont d'une part, le calcul numérique intensif et d'autre part l'archivage pérenne de documents électroniques. Depuis 2004, le CINES travaille sur la mise en place d'un service pour l'archivage pérenne du patrimoine scientifique.

L'information scientifique et technique (IST) désigne l'ensemble des informations produites ou reçues par les secteurs de la recherche et de l'enseignement. Elle est définie par des contenus constitutifs de connaissances, des supports documentaires ainsi que des canaux de communication spécifiques et elle s'inscrit dans des formes documentaires diverses : revues scientifiques, thèses, rapports, actes, ouvrages spécialisés, manuels, bibliographies, résumés, prépublication, brevets, cartes, banques d'images et de vidéos, données statistiques…

Elle est actuellement en phase de mutation, avec d'une part l'évolution des supports documentaires et canaux de communication et d'autre part, le passage massif au numérique (documents nativement au format électronique et opérations de numérisation de masse). Cependant ses limites demeurent imprécises en raison des liens étroits existants avec d'autres types d'informations produites hors du champ de la recherche, ce phénomène étant accentué par les profondes transformations des canaux de communication.

La réussite au niveau national d'une stratégie pour l'archivage pérenne de documents électroniques produits par cette communauté de l'information scientifique et technique passe par la réalisation d'un certain nombre de défis : d'une part l'acquisition d'une nouvelle compétence métier (avec la mise en place d'un service dédié à l'archivage, la définition de nouveaux processus et méthodes à mettre en œuvre et à maîtriser pour piloter un système d'archives), d'autre part la participation à différents groupes de travail ou initiatives sur le thème de la préservation de documents numériques et enfin la sensibilisation de la communauté à la problématique de la préservation à long terme des documents numériques (avec le renforcement des collaborations entre informaticiens, archivistes et bibliothécaires, l'organisation de journées de sensibilisation et d'information à l'archivage pérenne..). Par ailleurs, il était vital de sensibiliser les décideurs sur l'importance de l'enjeu.

Cette mise en place devait en outre s'effectuer en liaison avec les autres services d'archives qui interviennent dans cette chaîne selon l'origine et la nature des documents (Archives départementales, Archives nationales, BnF) dans le respect du contexte législatif.

#### 3.2.6.2. Mise en place de la plateforme PAC (Plateforme d'archivage du CINES)

Pour remplir cette mission d'archivage, le CINES a ainsi mis en place le projet PAC, qui vise à se doter d'un service d'archivage numérique pérenne, avec une équipe composée d'un chef de projet, de quatre ingénieurs et d'un archiviste.

Durant la phase 1, a été développée en interne une première plateforme pour valider les services attendus sur le projet d'archivage des thèses électroniques, avec une capacité de stockage réduite (300 Go). La plateforme était basée sur les standards du domaine et notamment sur le modèle OAIS, le format de métadonnées Dublin Core ainsi qu'une liste des formats de fichier acceptés volontairement limitée : formats publiés, largement utilisés, normalisés si possible (HTML, PDF, TXT, XML ; GIF, JPEG, TIFF, PNG ; WAV). L'architecture était basée sur les logiciels libres (Java, PostgreSQL, Jhove, ImageMagick) avec des premiers tests d'archivage des thèses en mars 2007 et un début de l'exploitation en production fin 2007.

Durant la phase 2, un appel d'offres a été notifié fin 2007 pour l'acquisition d'une plateforme de stockage capable de gérer de larges volumes (20 To extensibles à 40To). Cette plateforme est toujours basée sur les normes et standards du domaine, l'architecture étant basée sur du matériel SUN, le logiciel Arcsys et des logiciels libres comme Java, MySQL, ainsi que Jhove et ImageMagick pour l'identification, la validation et la caractérisation des formats. Les premiers tests de versement et de migration (documents archivés sur PAC v1.0) ont débuté en mars 2008 avec un début de l'exploitation à la fin du deuxième trimestre 2008.

La plateforme comprend les fonctionnalités suivantes : réception et contrôle des SIP (conformité des métadonnées au schéma sip.xsd, correspondance entre la description et les fichiers composant le document transféré, contrôle et validation du format des fichiers, calcul de l'empreinte de chaque fichier), création de l'AIP, stockage (copie multiple de l'AIP sur différents médias et supports, envoi du certificat d'archivage, vérification périodique de l'intégrité des AIP, migration technologique, fourniture d'états et statistiques), accès (contrôle d'authentification des demandeurs, consultation et communication).

#### 3.2.6.3. Archivage des thèses électroniques et des revues en sciences humaines et sociales

Concernant les thèses, le principe a été initié suite à l'arrêté du 7 août 2006 relatif aux modalités de dépôt, de signalement, de reproduction, de diffusion et de conservation des thèses ou des travaux présentés en soutenance en vue d'un doctorat.

Les doctorants déposent leur thèse au format électronique dans la bibliothèque universitaire de leur lieu de soutenance. Les bibliothèques de leur côté versent les thèses électroniques à l'agence bibliographique de l'enseignement supérieur (ABES) via l'outil STAR. Après trois étapes de validation, les thèses éligibles à l'archivage sont transférées sur la plateforme PAC, une copie étant éventuellement disponible en ligne pour la communauté des internautes sur un site de diffusion.

Le projet d'archivage des revues a été initié en 2006 pour répondre à un projet de numérisation massive et de préservation de collections rétrospectives de revues en Sciences Humaines et Sociales par l'équipe Persée (Université Lumière – Lyon 2) (http://www.persee.fr/). La chaîne de numérisation assure une numérisation de masse, une centralisation et une robotisation des traitements ainsi qu'un archivage pérenne des données.

La chaîne de documentation comprend un outil de description des collections, des outils de documentation et de suivi, des étapes de contrôle qualité et de validation et enfin des outils pour la diffusion des données générées.

L'archivage de ces deux sources de données est actuellement en production sur la plateforme PAC.

#### 3.2.6.4. Nouveaux projets en cours

Trois projets sont actuellement en cours de réalisation : d'une part l'archivage de documents sonores issus de la recherche dans le domaine de l'oral : projet pilote du centre de ressources pour la description de l'oral (CRDO) dans le cadre du programme sciences humaines et sociales du Très Grand Equipement (TGE) Adonis du Centre national de la recherche scientifique (CNRS).

D'autre part, l'archivage de cours universitaires de Canal-U (documents vidéos) et enfin un projet d'archivage des documents déposés dans les archives ouvertes (HAL – Hyper Article en Ligne du centre pour la communication scientifique directe (CCSD).

Le projet pilote s'appuie sur trois acteurs (le CRDO, le CINES et le centre de Calcul de l'IN2P3 (CC-IN2P3), Institut national de Physique Nucléaire et de Physique des Particules.

Un partage des responsabilités a été mis en place avec les fonctions de préparation des versements pour le CRDO, les fonctions de prise en charge des versements et de contrôle et validation, stockage et de préservation pour le CINES et les fonctions de diffusion pour le CC-IN2P3.

L'objectif est la mise en place d'une infrastructure mutualisée pour la préservation et la diffusion de données de Sciences Humaines et Sociales, l'expérimentation portant sur des données «orales» qui rassemblent à la fois des enregistrements sonores, du texte (transcriptions, translitérations, annotations) et parfois des enregistrements vidéos.

Ce projet a permis notamment d'enrichir la PAC de nouvelles fonctionnalités : prise en charge de nouveaux formats de fichiers son et vidéo (WAV, AIFF, OGG, MPEG4, Matroska) et d'encodage de ces données (PCM, AAC, Vorbis…), possibilité de lier des enregistrements et notations à un enregistrement d'origine par un lien de parenté, possibilité de mettre à jour certaines informations (enrichissement des métadonnées) sans re-transférer l'objet complet, transmission au CC-IN2P3 d'informations destinées à la diffusion qui n'ont pas vocation à être archivées dans PAC.

Ce projet pilote doit permettre de disposer de premiers éléments sur les coûts de fonctionnement de la solution, et de valider les fonctionnalités d'ensemble de la solution, son caractère générique en dissociant ce qui est totalement applicable aux autres données des sciences humaines et sociales de ce qui est spécifique des corpus oraux, la répartition des tâches et des responsabilités entre les acteurs, l'infrastructure matérielle et logicielle

mise en place, sur le plan des performances et de la fiabilité, les services permettant la recherche, la sélection, la récupération de données, avec la communauté des utilisateurs de corpus oraux.

A l'issue du projet pilote, après évaluation des résultats produits par un comité scientifique du TGE-Adonis, une extension à d'autres données devrait être décidée et l'utilisation de la solution à d'autres Centres de ressources numériques pourra alors être envisagée.

# Bibliographie

***[Premier ouvrage de synthèse sur l'archivage numérique en langue française.]***

• BANAT-BERGER F., HUC C., DUPLOUY L., L'Archivage numérique à long terme, les débuts de la maturité? Paris, La Documentation française, 2009.

***[Norme de référence essentielle pour comprendre le problème posé par l'archivage numérique]***

http://public.ccsds.org/publications/archive/650x0b1(F).pdf