CINES
Centre Informatique National de l'Enseignement Supérieur

LA GAZETTE DU CINES

DEUX MISSIONS STRATÉGIQUES : CALCUL INTENSIF ET ARCHIVAGE PÉRENNE

numéro spécial
Février 2013

DOSSIER SPÉCIAL

ARCHIVAGE NUMÉRIQUE PÉRENNE

L'éditorial

Francis DAUMAS

Directeur

De nos jours l'information sous forme numérique est omniprésente, en grande quantité, sous des formats multiples. Elle fait partie du patrimoine de la génération « informatique ». Comme tout patrimoine la question de sa conservation se pose, mais avec une accuité accrue compte-tenu des masses considérables de données concernées qui nécessitent de très gros volumes de stockage et des métadonnées robustes.

Si la sauvegarde apporte une sécurité actuelle des données elle ne constitue en aucune manière une garantie de pérennité dans le temps tant sur la lisibilité du support que sur la capacité d'une restitution compréhensible du contenu. Seule une véritable solution d'archivage peut prétendre répondre à ce défi. En ce domaine les offres clefs en main sont peu nombreuses.

Pour sa part le CINES a développé et mis en œuvre une plateforme (PAC), conforme au modèle OAIS et respectueuse des normes et standards. Elle a reçu l'agrément du SIAF (Service Interministériel des Archives de France) pour les archives intermédiaires et accepte tous types de données (administratives, patrimoniales, scientifiques : issues d'observations ou de calculs) sous tous types de formes (texte, son, image, vidéo, ...).

Dans ce numéro, des partenaires, services versants, témoignent de leur expérience avec le CINES. Le Ministère a désigné le CINES comme site national d'archivage des thèses au format numérique, ainsi le service STAR de référencement des thèses mis en place par l'ABES (Agence Bibliographique de l'Enseignement Supérieur) est interfacé avec PAC, l'outil d'archivage du CINES.

La bibliothèque Cujas, une référence en sciences juridiques, a confié au CINES l'archivage de ses fonds numérisés relatifs à l'histoire du droit. Le CNRS-CCSd (Centre pour la Communication Scientifique Directe) s'adresse au CINES pour la conservation à long terme des dépôts effectués dans l'archive ouverte HAL.

Assurer la pérennité de données dans le temps exige l'adoption d'une démarche qualité rigoureuse afin de prévenir les risques inhérents au défi relevé. Celle-ci est d'autant plus nécessaire que les différentes catégories de données ont leurs spécificités. Quels problèmes soulèvent les données administratives ? les données scientifiques ? comment s'assurer de la validité d'un format ? autant d'interrogations auxquelles il faut apporter des réponses et pour lesquelles une compétence archivistique permet d'éviter les écueils.

Apporter sa contribution à la conservation du patrimoine scientifique public national est l'une des deux missions statutaires du CINES sur laquelle s'appuie l'initiative BSN (Bibliothèque Scientifique Numérique) initiée et pilotée par le Ministère de l'Enseignement Supérieur et de la Recherche et qui est un cadre idéal pour créer une dynamique positive au sein des différents organismes.

Au-delà de nos frontières, l'Europe des grandes masses de données est en marche. Le CINES est partenaire dans les projets européens PRACE, axé sur les très gros moyens de calcul et les données associées, et EUDAT, orienté vers la mise en place d'une infrastructure collaborative pour le partage, la duplication, la diffusion et la conservation des données.

Enfin s'il n'est jamais trop tard pour penser à archiver, il est plus aisé d'intégrer cette préoccupation dès la production des informations. Pour preuve et pour rassurer les contributeurs à cette gazette, le CINES s'engage à l'archiver !

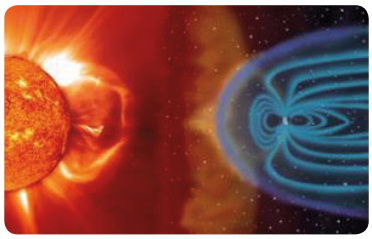
SOMMAIRE

ECLAIRAGE SUR ... 4

Qu'est-ce que le CINES ? 4
État de l'art de l'archivage électronique en France et dans le monde 5
Le service d'archivage électronique à long terme du CINES et la plateforme PAC 10

TEMOIGNAGE DE ... 18

L'agence bibliographique de l'enseignement supérieur (ABES) : l'archivage des thèses dans PAC 18
La bibliothèque Cujas : l'archivage du fonds des ouvrages numérisés en histoire du droit 22
CNRS : archivage des dépôts de l'archive ouverte pluridisciplinaire HAL (Hyper Articles en Ligne) 24

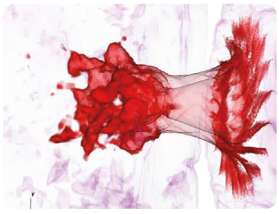


COMPETENCES ... 26

La démarche qualité pour la préservation à long terme des données numériques 26
L'expertise sur les formats de fichiers : une spécificité de l'offre d'archivage du CINES 36

FOCUS SUR... 38

L'archivage des données administratives : 5 questions à Charlotte Maday 38
L'archivage des données scientifiques : des besoins spécifiques 41
L'archivage des données scientifiques : ISAAC, une solution du CINES 46



RELATIONS INTERNATIONALES ... 49

Les collaborations du CINES aux niveaux national et européen 49
L'archivage à l'heure de l'Europe 53

FICHE PRATIQUE

Comment archiver au CINES ? 56

GLOSSAIRE

Les notions utiles 58

Qu'est-ce que le CINES ?

Le C.I.N.E.S. (Centre Informatique National de l'Enseignement Supérieur) est un établissement public national autonome, basé à Montpellier (France) et placé sous la tutelle du Ministère de l'Enseignement Supérieur et de la Recherche. Il est composé d'une cinquantaine d'agents.

Outre l'archivage numérique pérenne, le CINES a une autre mission statutaire stratégique : le calcul intensif. Pour remplir cette mission historique, le CINES offre aux enseignants-chercheurs des moyens de calcul exceptionnels, parmi les plus puissants au monde, pour relever de grands défis scientifiques. Ces équipements de pointe permettent de réaliser des simulations extrêmes de phénomènes difficilement observables ou reproductibles. Le supercalculateur parallèle Jade (SGI) possède 23040 cœurs pour une puissance de calcul de 267 Teraflop/s (247 mille milliards d'opérations flottantes par seconde). Tous les domaines scientifiques sont clients des moyens de calcul du Centre. Les principaux consommateurs sont la mécanique des fluides, la physique des plasmas, l'astrophysique et la géophysique, ainsi que la biologie et les systèmes moléculaires organisés.

Une équipe d'ingénieurs hautement qualifiés travaille avec les chercheurs des universités et des organismes de recherche publics et assure une aide personnalisée et des formations aux techniques de calcul intensif afin de faciliter et d'optimiser l'utilisation de ces ressources exceptionnelles.

Le CINES participe à plusieurs projets européens comme PRACE ou HPC-Europa qui mettent en place une Europe du calcul. Le centre propose également un accueil scientifique.

État de l'art de l'archivage électronique en France et dans le monde

Les termes relatifs à la conservation des données numériques s'enrichissent quotidiennement et témoignent du besoin grandissant des établissements de trouver des solutions afin de gérer ce patrimoine informationnel fragile. La problématique est double : la production documentaire informatique ne cesse d'augmenter et prend diverses formes (bureautique, email, blogs, bases de données, etc.) alors que l'information produite est très volatile et peut se détériorer rapidement si certains moyens ne sont pas mis en œuvre. De plus, la gestion de ces documents requiert un ensemble de compétences jusqu'alors assez éloignées : celles de l'archiviste, expert en gestion de l'information et du cycle de vie du document ; celles de l'informaticien, spécialiste des risques liés à l'environnement numérique ; ou encore celles du juriste, garant du respect de la règlementation en vigueur.

Du fait de ces spécificités, un véritable marché de l'archivage électronique est en train de voir le jour, comparable à celui de la GED[1] il y a une dizaine d'années. Mais il recouvre des besoins différents et plusieurs notions coexistent au risque d'être souvent confondues : archivage probant, système d'archivage électronique, coffre-fort électronique, archivage patrimonial, sauvegarde sécurisée, etc. L'ambition de cet article n'est pas de dresser une liste exhaustive des offres mais plutôt de mettre en lumière quelques solutions intéressantes pouvant servir de point de départ et de comparaison à toute réflexion sur l'archivage électronique.

Archivage patrimonial vs Archivage à valeur probante vs Sauvegarde sécurisée

Tout fichier informatique qui doit être préservé à moyen ou long terme est confronté inéluctablement à quatre menaces principales : l'obsolescence du matériel, la disparition du logiciel de lecture, l'obsolescence du format du fichier et la perte de la signification du contenu.

Préserver l'information numérique consiste à mettre en œuvre les moyens nécessaires pour lutter contre ces risques, c'est-à-dire : conserver le document (garantir son intégrité et son authenticité), le rendre accessible et en préserver l'intelligibilité. La complexité de la tâche dépend en grande partie de la durée de conservation du document.

Dans un premier temps, il convient de faire la distinction entre l'archivage à valeur patrimoniale et l'archivage à valeur probante. L'archivage patrimonial concerne principalement les établissements publics ayant un patrimoine à archiver, tels que les bibliothèques, services d'archives, musées ou universités. Qu'il s'agisse de documents prestigieux numérisés ou d'actes administratifs règlementaires nativement numériques, ces documents sont généralement à conserver sur le long terme, voire définitivement. Lorsque l'enjeu principal est de garantir l'intégrité d'un document et de prouver son authenticité à court ou moyen terme, on parle plutôt d'archivage à valeur probante. Cela concerne surtout les documents engageants d'un établissement tels que les factures, les contrats ou le courrier.

La sauvegarde sécurisée, quant à elle, ne fournit qu'une copie de sécurité d'un ensemble d'informations numériques. Le but n'est pas de documenter les informations à préserver mais d'avoir un site 'miroir' permettant seulement de se prémunir des pertes liées à des incidents au sein de l'infrastructure (pannes du matériel, vols, dégâts sur le bâtiment, etc.).

C'est ce que propose notamment le service de stockage en ligne Amazon S3. Une sauvegarde ne peut donc concerner que des documents à conserver sur du court terme.

(1) Gestion Électronique de Documents

Le point de vue de… Frédérique FLEISCH

Frédérique FLEISCH

Association des Archivistes Français (AAF)
Présidente de la Commission Archivage Électronique

« Lorsque l'on aborde la question de l'archivage électronique avec notre direction et notre service informatique le retour est plutôt favorable mais nous sommes régulièrement confrontés à la méconnaissance de ce qu'implique un véritable archivage électronique. Beaucoup utilisent le terme d'archivage pour désigner en fait une sauvegarde sécurisée ; or les moyens à mettre en œuvre ne sont pas les mêmes dans les deux cas.

Depuis 2004, c'est dans une approche résolument pragmatique que la Commission Archivage Électronique (CAE), organe de l'Association des Archivistes Français (AAF), réunit des experts issus des secteurs publics et privés pour sensibiliser à cette problématique. Elle effectue notamment un lobbying auprès des différents acteurs du monde de l'archivage électronique tout en identifiant un langage commun entre DSI, prestataires, éditeurs de logiciel et archivistes.

Des rencontres interactives avec des éditeurs de solutions d'archivage électronique ont été organisées. La CAE constate qu'il n'existe que quelques solutions sur le marché français, proposant des niveaux de service différents allant de la seule garantie d'intégrité et d'authenticité (coffre-fort électronique) à un engagement sur la lisibilité et l'intelligibilité des données archivées sur le long terme.

La CAE compte maintenir ses échanges et son partenariat avec les différents acteurs du monde de l'archivage électronique afin de leur faire remonter les demandes des archivistes. »

L'archivage électronique dans le secteur public

Au niveau national, la sphère publique est la première à avoir mis en œuvre des solutions afin de gérer son patrimoine informationnel numérique, notamment les données scientifiques.

Depuis 1998, le CNES[2] archive des images de satellites et de grands observatoires, des données numériques issues de logiciels spécifiques ainsi que les documents produits dans le cadre de projets d'études spatiales (missions) initiés par le Centre de Données de la Physique des Plasmas.

Le Système d'Information, de Préservation et d'Accès aux Données (SIPAD) qu'il utilise gère les fonctions d'acquisition, de gestion des données descriptives et d'accès aux données. Le stockage des données est pris en charge par le module STAF (Service de Transfert et d'Archivage de Fichiers). Ce système d'archivage électronique a fait l'objet d'une migration vers une nouvelle plateforme en 2007 (SIPAD-NG). La spécificité des archives scientifiques numériques, outre leur haut niveau de technicité, réside dans les larges volumes à conserver.

(2) Centre National d'Études Spatiales

Début 2010, le système d'archivage de l'IGN[3] archivait plus de 110 To de données cartographiques, photographiques ou d'extractions d'applications. Du fait des volumes importants comme au CNES, le stockage s'effectue sur bandes magnétiques, moins coûteux à grande échelle que le stockage sur disques.

> *« Par ailleurs, la complexité de ces archives oblige, encore plus que pour n'importe quelle autre catégorie d'archives, à les prendre en compte le plus en amont possible, dès la production de la donnée. »*

Ainsi, à l'IGN, l'étape d'archivage intervient entre la production et la diffusion de l'information. Au CINES, un nouveau service ISAAC (Information Scientifique Archivée Au CINES) sera proposé prochainement à la communauté des chercheurs afin d'archiver les résultats de calculs scientifiques ou d'observations, notamment issus des supercalculateurs du CINES. Le système pourra archiver de larges volumes de données sur du moyen terme, tout en offrant des services compatibles avec un archivage définitif (effectué dans PAC – la Plateforme d'Archivage du CINES, après sélection). En plaçant l'archivage très tôt dans le cycle de production de la donnée, ISAAC permet de collecter et de documenter l'information scientifique directement auprès des producteurs.

Si la technicité des archives scientifiques fait que les pratiques d'archivage électronique existent depuis un certain temps déjà dans ce milieu, elles ne sont pas pour autant les seules concernées. De par leur statut juridique, toutes les archives publiques[4] 'papier' ou numériques, doivent faire l'objet d'un traitement spécifique qu'elles soient conservées ou détruites. La gestion de ces archives administratives est assurée par les services publics d'archives territoriaux ou ministériels dont certains ont créé leur propre système d'archivage électronique (SAE), à l'image de la plateforme M@rine du Conseil général de l'Aube, du projet ARCADE du Ministère de l'Économie de l'Industrie et de l'Emploi, ou encore ATLAS de la Direction Générale des Finances Publiques. Les documents à archiver sont en grande partie issus de la bureautique et conservés au format PDF/A suivant les recommandations du Service Interministériel des Archives de France (SIAF) et du Référentiel Général d'Interopérabilité (RGI). Ils sont accompagnés de métadonnées généralement conformes au Standard d'Échange de Données pour l'Archivage[5]. Cependant, toutes les administrations n'ont pas les moyens nécessaires (financiers, humains et techniques) pour mettre en place et gérer leur propre SAE. Alors que certaines se tournent vers les offres de tiers archivage (cf. infra), l'ADULLACT[6], en partenariat avec le SIAF, a développé un logiciel d'archivage libre, As@lae, à destination des collectivités locales. Mis en production en 2010, ce logiciel qui respecte les principales normes du métier rencontre un vif succès auprès des collectivités : Agence landaise pour l'informatique, Archives départementales de la Nièvre et de la Vienne, etc.

Avec le développement de la numérisation, les institutions qui possèdent un patrimoine documentaire important sont également concernées par l'archivage électronique si elles veulent préserver leurs ouvrages numérisés. À titre d'exemple, la Bibliothèque nationale de France utilise SPAR (Système de Préservation et d'Archivage Réparti) pour l'archivage de ses collections numérisées (accessibles dans Gallica), des publications électroniques et du web (obligation de dépôt légal). Cela représente des volumes de données considérables (plus de 500 To début 2013), stockées sur bandes magnétiques. Dans une volonté de mutualisation des expertises et des coûts, il est prévu que la BnF ouvre son système à d'autres partenaires et institutions offrant ainsi un service de «tiers-archiveur» du patrimoine numérique.

(3) Institut Géographique National
(4) « Les archives publiques sont [...] Les documents qui procèdent de l'activité, dans le cadre de leur mission de service public, de l'Etat, des collectivités territoriales, des établissements publics et des autres personnes morales de droit public ou des personnes de droit privé chargées d'une telle mission [...]. »
Code du Patrimoine, art. L211-4
(5) Cf. encadré page 40
(6) Association des Développeurs et des Utilisateurs de Logiciels Libres pour l'Administration et les Collectivités Territoriales

Le CINES se positionne à la croisée de ces différents types d'archivage : à la fois scientifique, patrimonial et administratif. La plateforme PAC propose depuis 2006 des services de tiers-archivage pérenne à destination de l'ensemble de la communauté enseignement supérieur et recherche[7]. Le CINES bénéficie entre autres d'une réelle expertise sur les formats de fichiers et la certification de services d'archivage électronique, qui fait de lui un des acteurs majeurs de la préservation numérique à l'heure actuelle.

Le développement du marché privé

Pour répondre à la demande des administrations et des entreprises, le secteur privé a également développé une offre. Des fonctionnalités de coffre-fort électronique au SAE complet, les sociétés leaders sur le marché du tiers archivage 'papier' proposent désormais une solution d'archivage électronique, majoritairement à valeur probante. Ce type d'archivage qui s'est généralisé avec la dématérialisation des procédures administratives, dans le public comme dans le privé, utilise notamment les procédés de signature électronique afin de garantir l'intégrité et l'authenticité des données. La norme de référence lorsque l'on parle d'archivage à valeur probante, la NF Z 42-013[8], détaille toutes les exigences requises pour assurer la conservation et l'intégrité de documents électroniques.

Certaines offres sont proposées sous la forme de modules complémentaires intégrant un coffre-fort électronique tel que celui de CDC Arkhinéo ou de Cecurity.com auquel on peut ajouter des outils de signature électronique (CertSign de CertEurope) ou de parapheur électronique (i-Parapheur de l'Adullact). D'autres sociétés fournissent des solutions d'archivage électronique plus complètes comme Mnesys de Naoned Systèmes, Spark Archives de Klee Group ou EverSuite Compliance d'Ever Team. Ces SAE proposent des fonctionnalités supplémentaires telles que la conversion automatique des fichiers en PDF/A et pour certains intègrent des briques logicielles externes (le cœur de métier de Locarchives repose sur le coffre-fort développé par Cecurity.com).

Toutefois, la variété de l'offre du marché privé ne doit pas faire perdre de vue qu'elle ne concerne en fait qu'une partie de la production documentaire : les documents issus d'applications bureautiques ; pour lesquels il est parfois proposé une conversion automatique en PDF mais sans contrôler le format du fichier au moment de l'archivage. Les métadonnées à joindre au fichier pour le documenter sont quant à elles totalement libres.

Ainsi, si on compare ces offres avec les services fournis par les SAE mis en place dans les institutions patrimoniales (CNES, BnF, etc.), elles semblent plus adaptées pour un archivage de documents texte ou image sur du moyen terme.

Le CINES par exemple, en plus d'un logiciel d'archivage amélioré par des développements complémentaires, a mis en place une équipe dédiée qui dispose d'une forte expertise sur les problématiques métier informatiques et archivistiques afin d'accompagner les services versants et d'assurer une gestion et un suivi de qualité.

Les systèmes d'archivage électronique à l'étranger

Sur le plan international, les réflexions sur la préservation de l'information numérique sont nombreuses.

(7) Cf. article « Le service d'archivage électronique à long terme du CINES et la plateforme PAC », page 10

(8) AFNOR NF Z42-013 – Archivage électronique : Spécifications relatives à la conception et à l'exploitation de systèmes informatiques en vue d'assurer la conservation et l'intégrité des documents stockés dans ces systèmes (Mars 2009)

« Qu'il s'agisse de systèmes d'archivage électronique open source ou de solutions mises en place dans les institutions patrimoniales d'envergure nationale, certains pays font figure de pionniers comme les États-Unis ou les Pays-Bas. »

FedoraCommons™

Le projet Fedora Commons entrepris dès 1998 par l'Université de Cornell aux États-Unis a permis d'implémenter un logiciel open source pour conserver, gérer et accéder à tout type de contenu numérique, grâce à une architecture flexible et extensible. De nombreuses institutions (bibliothèques nationales, universitaires, services d'archives, groupes de recherche, musées ou encore entreprises) ont choisi Fedora comme socle pour construire leur SAE. D'autres projets du même genre ont ensuite vu le jour au sein d'universités anglo-saxonnes comme E-Prints à Southampton ou DSpace à l'Institut de Technologie du Massachusetts. Toutes ces plateformes de dépôt, d'archivage et de diffusion de publications électroniques sont libres et paramétrables. Elles permettent la constitution d'une base de données des métadonnées associées et la gestion-distribution à long terme des documents.

eprints repository software

DSPACE

archivematica.

Le logiciel Archivematica, quant-à-lui, développé par Artefactual Systems, s'adresse plus à des archivistes et bibliothécaires ayant des moyens limités, en offrant une solution performante, gratuite et libre qui fonctionne comme un ensemble de micro-services en lien les uns avec les autres.

Régulièrement, ces acteurs de l'archivage électronique se retrouvent lors d'évènements internationaux tels que l'IPRES (International Conference on Preservation of Digital Objects), le PASIG (Preservation and Archiving Special Interest Group) ou l'ICA (International Council on Archives) pour faire connaître leurs activités et échanger sur les bonnes pratiques. Parmi les principaux interlocuteurs on peut citer la bibliothèque royale des Pays-Bas et son e-Depot qui archive les publications électroniques, le web et les bruts de scan, les deux plateformes américaines Portico et Hathi Trust chargées de collecter la littérature scientifique numérique, la Safety Deposit Box de Tessella utilisée notamment par les Archives nationales de Grande-Bretagne, le système portugais RODA ou encore la bibliothèque nationale de Nouvelle-Zélande qui a basé sa plateforme sur la solution Rosetta de la société Ex Libris pour y archiver ses collections patrimoniales numériques et numérisées.

PORTICO

ExLibris Rosetta

DEPOT KB

HATHI TRUST

Tessella Technology & Consulting

Ces dernières années l'offre du marché, aussi bien français qu'international, s'est donc largement développée afin de s'adapter aux besoins de toutes les administrations quels que soient leur production documentaire et leurs moyens. Nous sommes tous concernés, parfois sans le savoir, car nous produisons tous les jours des documents électroniques. L'enjeu actuel consiste désormais à sensibiliser le plus grand nombre de personnes à cette problématique, améliorer les services proposés et inculquer les bonnes pratiques. Cela peut se faire notamment au travers du groupes nationaux de travail tels que le groupe PIN (Pérennisation de l'Information Numérique – Association ARISTOTE) ou de l'AAF (Association des Archivistes Français).

Lorène BECHARD

Archiviste
Équipe PAC
CINES

Le service d'archivage électronique à long terme du CINES et la plateforme PAC

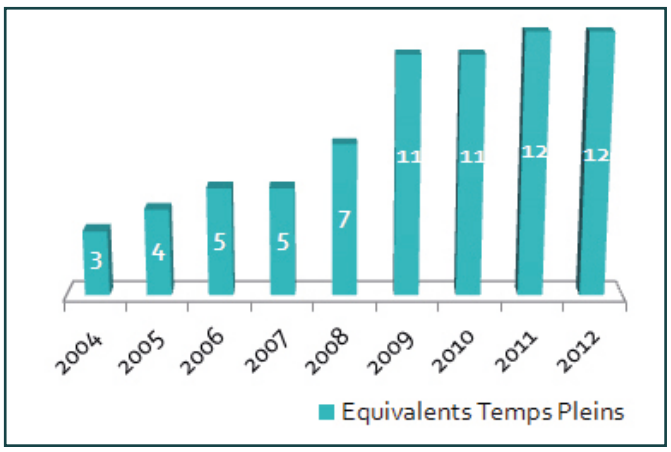
Au cours des années 2000, les missions historiques du CINES ont évolué pour s'étendre au domaine de l'archivage numérique pérenne. Le Ministère de l'Enseignement Supérieur et de la Recherche a confié au CINES la mise en place d'un service d'archivage à long terme des données sous forme numérique destiné à la préservation du patrimoine scientifique.

Cette problématique étant au cœur des préoccupations actuelles, plusieurs décisions ont depuis renforcé cette mission :

- 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 ; qui désigne le CINES comme centre de conservation des thèses électroniques françaises.

- Convention du 2 mai 2007 (faisant suite à celle du 15 octobre 2003) relative à la mise en ligne et l'archivage pérenne des données numérisées dans le cadre du programme Persée.

- Lettre de cadrage du 12 février 2008 recentrant les activités du CINES autour de deux missions stratégiques : le calcul intensif et l'archivage pérenne.



ÉVOLUTION DU NOMBRE DE COLLABORATEURS

PAC EN QUELQUES CHIFFRES…

- 10 services versants
- Plus de 8 000 000 de fichiers archivés (300 000 documents)
- Chaque document est conservé en 4 exemplaires
- Près de 25 To utiles d'archives (réel : ~ 100 To)
- 12 personnes dédiées
- Volumétrie : +500 % sur les années 2011 et 2012 (accroissement exponentiel)

Une équipe de 12 personnes aux profils divers et complémentaires travaillent dans ce service : un chef de projet, une archiviste, des experts en formats de fichiers, des développeurs informatiques, des administrateurs des systèmes et infrastructures, des techniciens hautement qualifiés (pour une surveillance 24h/24 et 7j/7), etc. Cette solution est proposée à l'ensemble de la communauté de l'enseignement supérieur et de la recherche dès lors que le corpus de documents à archiver est reconnu d'intérêt national[9].

Après trois années de conception et de développement, une première version du système PAC - Plateforme d'Archivage au CINES (PAC v1.0) - a été mise en service au printemps 2007 avec comme axe initial l'archivage des thèses électroniques. Toutefois, l'infrastructure matérielle ne permettant pas de gérer le volume de données prévu pour de nouveaux projets d'archives, un appel d'offres a été lancé pour l'acquisition d'une plateforme capable de gérer de larges volumes (plus de 40 Téraoctets utiles).

(9) Cf. fiche pratique « Comment archiver au CINES ? », page 56

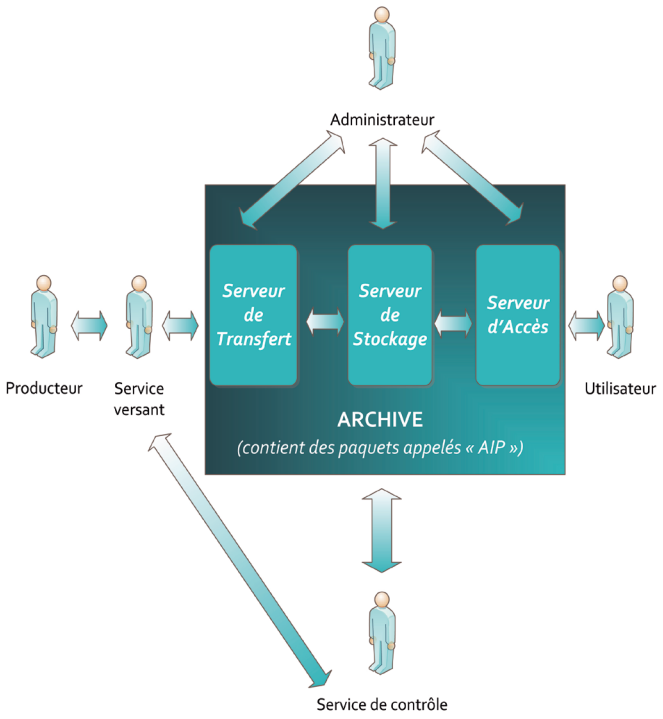
C'est donc une deuxième version du système PAC (PAC v2.0) qui est exploitée depuis le printemps 2008. Elle privilégie toujours la même approche généraliste, évitant ainsi le traitement des projets d'archives au cas par cas, et permet de mutualiser la plateforme pour tous les projets d'archives. Elle respecte les normes et pratiques du domaine comme l'OAIS[10], le standard d'échange de données pour l'archivage (SEDA[11]) ou le Dublin Core.

Une infrastructure matérielle et logicielle générique

Le système PAC a été conçu comme un ensemble de trois serveurs logiques, s'inspirant du modèle proposé par la norme ISO 14 721 (OAIS – Open Archival Information System) :

- un serveur de versement auquel le producteur des documents pourra transmettre ses archives ;
- un serveur de stockage où sont conservés les documents ;
- un serveur d'accès au travers duquel les utilisateurs autorisés (producteurs par exemple) peuvent consulter les archives et en obtenir une copie.

Le serveur de transfert réalise des opérations ayant attrait au cœur de métier de l'archivage électronique. En effet, il fournit notamment une assurance qualité sur les archives (contrôle de la conformité des métadonnées, contrôle des formats de fichiers, récupération automatique des schémas et DTD qui définissent les métadonnées métier, etc.). C'est la raison pour laquelle le développement de cette brique logicielle de PAC (appelée « module d'entrée » ou « module de versement ») est effectuée par les équipes du CINES.



PLATEFORME D'ARCHIVAGE PAC : POINT DE VUE LOGIQUE

« Ce serveur logique (développé en Java et utilisant une base de données MySQL) est couplé d'une part au progiciel Arcsys (de la société Infotel) et d'autre part au module ArCheck. »

Arcsys gère le stockage des fichiers et fournit une interface d'accès basique aux données. Le module ArCheck (développé en interne) effectue pour sa part des opérations métier complémentaires en vérifiant la qualité de toutes les archives au moins une fois par an.

D'un point de vue matériel, la plateforme actuelle utilise des serveurs d'application SUN X4150, des baies de stockage SUN Storagetek ST6140-4G (en double exemplaire sur les supports disques) et des bandothèques IBM 3584 (un exemplaire des archives dans chaque robothèque).

(10) Cf. encadré « Le modèle OAIS », page 14-15
(11) Cf. encadré «le SEDA», page 40

Les différents échanges dans PAC

Les communications entre les trois serveurs ou avec les services versants et les utilisateurs se font par le biais d'échanges.

Transfert

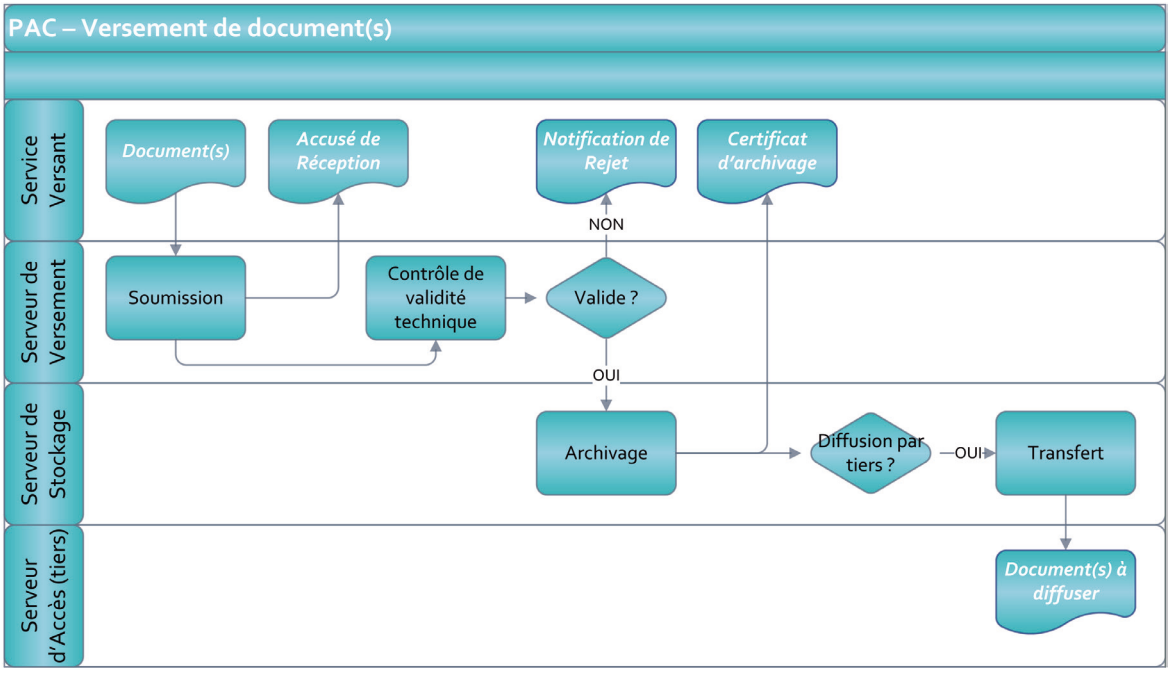
Le service versant transmet par réseau (via SFTP) ou sur support amovible, les informations à verser au service PAC du CINES.

Le transfert est repéré par l'application PAC qui procède alors à une série de contrôles de validité technique des données transférées : conformité des formats, structure des versements, etc. Si le contrôle est positif, le versement se voit enrichi de métadonnées propres à l'archivage telles que l'identifiant unique et pérenne ou la date d'archivage. Le sip.xml devient alors aip.xml car le paquet à envoyer au serveur de stockage est un AIP (Paquet d'Informations Archivé). Lorsque le paquet est archivé, le service PAC renvoie au service versant un certificat d'archivage comportant entre autres l'identifiant unique et pérenne de l'archive ainsi que ses métadonnées.

Mon archive/
- sip.xml (fichier de métadonnées décrivant l'objet à archiver)
- DÉPOT (qui contient l'objet à archiver et ses éventuelles métadonnées métier)
  - Fichier 1 à archiver (exemple : le mémoire de thèse)
  - Fichier 2 à archiver (exemple : une annexe de la thèse)
  - DESC
    - Fichier de métadonnées métier complémentaires

STRUCTURE D'UN DÉPÔT

L'identifiant de l'archive est un identifiant international unique qui permet au service versant ou à des utilisateurs de PAC de retrouver l'archive. Dans le cas où la qualité du versement n'aurait pas été jugée satisfaisante, une notification d'erreur est envoyée au service versant afin qu'il modifie les éléments non conformes et reverse le paquet d'archives.



LES ÉTAPES DU VERSEMENT D'UNE ARCHIVE DANS PAC

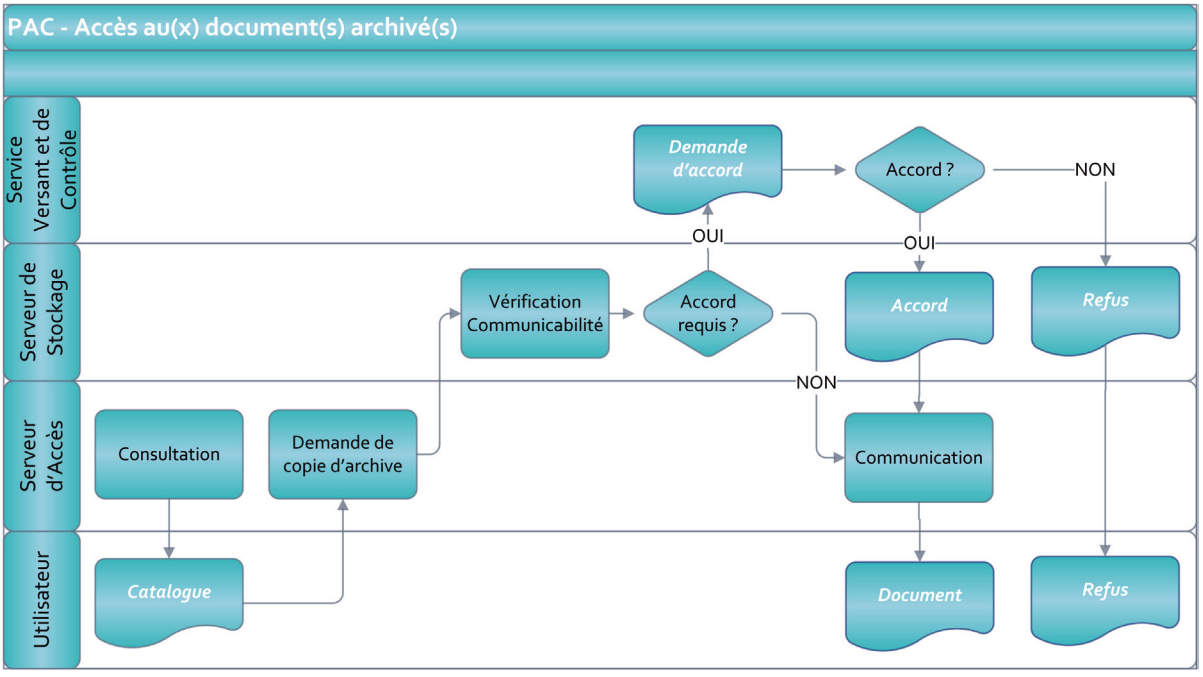
Stockage

Lorsqu'il arrive sur le serveur de stockage, tout document à archiver est copié simultanément sur deux disques distincts puis sur deux bandes magnétiques. Chacune de ces copies est conservée dans une salle distincte et une autre copie disque sur site distant (CC-IN2P3 à Lyon - en cours). Une réplication croisée de données avec la Bibliothèque nationale de France (BnF) est également en cours. Elle permettra au CINES d'avoir un site distant de plus de 800 km et de réduire d'autant plus le risque de pertes de données dans le cas de catastrophes majeures (attentats, inondations, incendies...). Pour faciliter les recherches, les métadonnées associées au document sont indexées. Il n'y a pas de suppression des documents archivés dans la plateforme. Les mises à jour sont possibles et se superposent aux documents déjà archivés. Le CINES réalise une veille technologique sur les évolutions des technologies, des environnements de stockage, et des formats. Durant la vie du document dans le système PAC, des opérations d'exploitation courante peuvent être menées (contrôle des supports, migrations de support à support, restaurations, contrôle des métadonnées). Le service versant n'en sera pas informé. Cependant, lorsque les modifications concernent le format des fichiers ou une mise à jour des métadonnées, le service versant en est averti.

Communication / Accès aux données

La communication se fait via le serveur d'accès en conformité avec le modèle OAIS et le standard d'échanges de données pour l'archivage (SEDA).

On s'assure en premier lieu que l'utilisateur a bien les droits d'accès à la plateforme avant de lui communiquer le catalogue des archives auxquelles il a accès. Il fait alors une demande d'archives qui, après vérification de la communicabilité (i.e. restrictions d'accès sur le document en application de la loi sur les archives)[12], lui est accordée ou non. Si l'archive n'est pas communicable, l'utilisateur a la possibilité de demander, via le CINES, une dérogation au service versant/producteur. Si la communication est autorisée, le serveur d'accès demande une copie de l'archive au serveur de stockage, la prépare et la met à disposition de l'utilisateur dans un répertoire dédié.



LES ÉTAPES DE LA COMMUNICATION D'UNE ARCHIVE À UN UTILISATEUR DE PAC

(12) Code du Patrimoine, art. L213-1 à L213-8

Les stratégies de pérennisation

Pour pouvoir garantir un archivage électronique pérenne, il faut mettre en place un certain nombre de mesures et planifier son action. Cela prend la forme d'un plan de préservation. Il définit un ensemble d'actions de conservation à mettre en place. Il doit prendre en compte la politique de conservation du service d'archives, la réglementation, les contraintes techniques et organisationnelles et les besoins des utilisateurs. Il évalue les différentes stratégies possibles pour en choisir une et expliquer les raisons de ce choix. Au CINES, la stratégie retenue est centrée sur deux mesures principales :

- L'utilisation de métadonnées et la conservation des informations de représentation, pour garantir la compréhension des documents à travers le temps.

- La réalisation de migrations physiques (i.e. de support à support) et logiques (i.e. conversion de formats de fichiers) lorsque les activités de veille détectent des signes d'obsolescence.

Pour tenir compte de l'hétérogénéité des projets d'archives, tout en garantissant la compréhension des documents archivés à travers le temps, le CINES a fait le choix d'allier plusieurs jeux de métadonnées de plusieurs types. Au niveau le plus haut, ces métadonnées permettent de décrire le contexte du projet d'archives : le(s) fonds à archiver, le(s) producteur(s), le(s) service(s) versant(s) ainsi que les choix retenus pour l'archivage. Elles sont regroupées dans un document appelé PPDI (Project Preservation Description Information). Les métadonnées du sip.xml et de l'aip.xml décrivent quant-à-elles le document archivé, au moyen de métadonnées descriptives (normalisées selon le Dublin Core) et de métadonnées de gestion (propres à PAC). Ce jeu de métadonnées étant volontairement généraliste, les services versants sont par ailleurs encouragés à joindre un fichier de métadonnées métier à chaque document archivé. Les métadonnées métier décrivent les documents selon leurs spécificités, généralement en suivant le standard utilisé par la communauté, comme OLAC pour les ressources linguistiques. Au CINES, la notion de métadonnées est interprétée dans son sens le plus large, à savoir : toute donnée permettant d'expliquer et de comprendre une autre donnée. Ainsi, PAC préserve également de manière pérenne ce que l'OAIS appelle des informations de représentation. Il s'agit notamment des spécifications des formats de fichiers archivés au CINES, des schémas XSD et DTD des fichiers également archivés, ainsi que de toute la documentation inhérente au service (spécifications de la plateforme, normes utilisées, processus métier, etc.).

LE MODÈLE OAIS

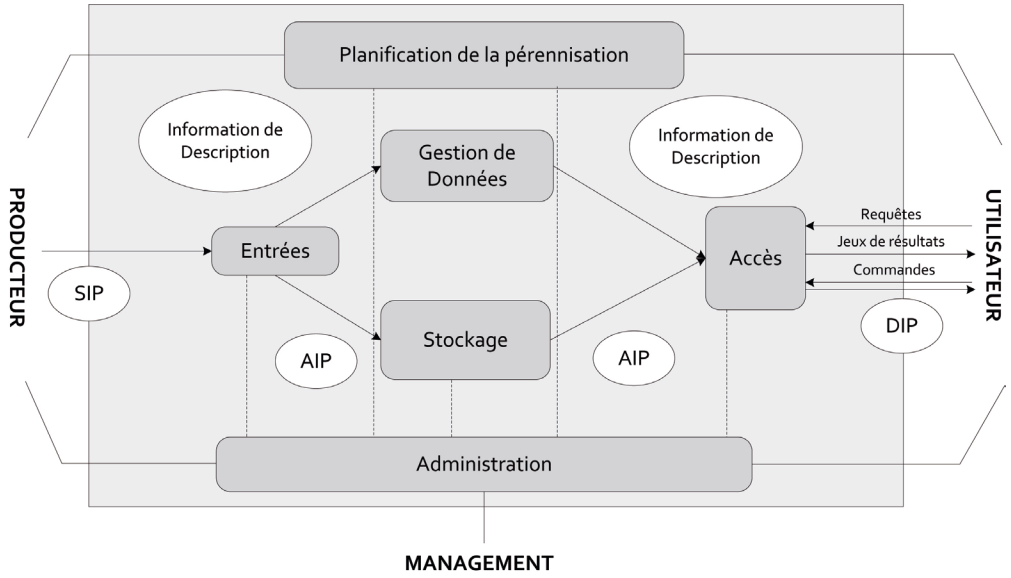
Le modèle de référence OAIS - Reference Model for an Open Archival Information System - est le résultat d'un travail du CCSDS (organisme international de normalisation des agences spatiales) - réalisé à la demande de l'ISO.

Ce travail, auquel ont été associés des représentants de bibliothèques et d'archives institutionnelles, a abouti en 2002 à la rédaction d'un document qui spécifie de manière très générale l'architecture logique et les fonctionnalités d'un système d'archivage. C'est aujourd'hui une norme internationale ISO (ISO 14721).

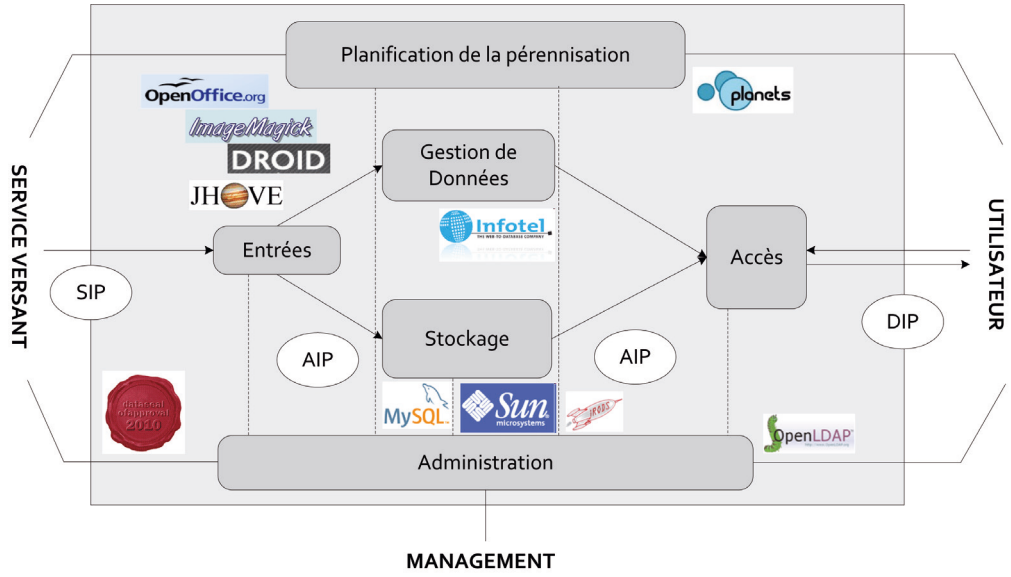
L'OAIS est un modèle abstrait. Il définit une terminologie et des concepts communs à l'ensemble des acteurs de la préservation, qu'ils soient informaticiens, archivistes, bibliothécaires ou manageur. Il identifie les acteurs, décrit les fonctions et les flux d'information et propose un modèle d'information particulièrement adapté à la problématique de l'archivage numérique, même s'il ne préjuge pas de la nature des objets à archiver. L'OAIS n'est pas un recueil de spécifications techniques destinées à être directement implémentées. C'est un guide qui cerne la problématique dans son ensemble et qui contraint à se poser toutes les bonnes questions.

En tant que modèle conceptuel de référence, l'OAIS s'est aujourd'hui largement imposé, au niveau international, auprès de tous les acteurs institutionnels de l'archivage numérique.

L'autre action du CINES en matière de pérennisation porte sur la problématique de l'obsolescence des supports de stockage et des formats de fichiers. Dans le premier cas, le CINES effectue une surveillance permanente de l'évolution et du vieillissement des médias de stockage afin de procéder à des migrations physiques le moment venu. Dans le cas de l'obsolescence des formats de fichiers, outre la sélection de quelques formats pour l'archivage dans PAC et le contrôle des formats des fichiers lors de leur versement, le CINES a mis en place une procédure de migration des fichiers qui se déclenche lorsque le format est obsolète. Cette procédure est composée de plusieurs étapes, de l'identification du format cible à la conversion effective des fichiers et à leur archivage dans PAC. Elle se déroule en étroite collaboration avec le service versant et les utilisateurs afin d'identifier notamment les caractéristiques essentielles à préserver dans le document lors de la conversion.



Le schéma ci-dessous présente les outils mis en œuvre dans PAC pour répondre aux exigences du modèle OAIS.



Les utilisateurs du service d'archivage PAC

Suite à l'arrivée des premiers projets en 2007 et 2008, les procédures d'archivage électronique ont pu être validées. 2010 et 2011 ont été des années charnières pour le service PAC. En effet, de nombreux projets ont débuté en 2010 et sont passés en phase d'archivage en 2011. Aussi, en deux ans, le nombre d'établissements archivant au CINES (« services versants ») a plus que quadruplé.

2007

LES UTILISATEURS HISTORIQUES

Les premiers documents archivés dans la plateforme furent les thèses électroniques délivrées en France et collectées par l'ABES – Agence Bibliographique de l'Enseignement Supérieur – auprès des bibliothèques universitaires par l'intermédiaire de l'outil STAR. Depuis 2007, près de 15 000 thèses ont été archivées dans la plateforme.

2008

Les revues en sciences humaines et sociales numérisées dans le cadre du programme PERSEE ont constitué le deuxième projet d'archivage du CINES. Les premières revues ont été archivées dès 2008. La volumétrie de ce projet est relativement considérable puisqu'elle dépasse les 16 To au début de l'année 2013.

2009

L'OUVERTURE DE LA PLATEFORME À UNE COMMUNAUTÉ PLUS LARGE

Après 6 mois de mise en œuvre, les premiers documents de l'archive ouverte HAL (Hyper Article en Ligne) gérée par le CNRS-CCSD (Centre pour la Communication Scientifique Directe) ont été archivés dans PAC le 29 mars 2010. Il s'agit d'articles scientifiques pluridisciplinaires de niveau recherche ainsi que de thèses françaises ou étrangères disponibles en accès libre sur le portail HAL.

Le projet pilote de l'archivage des données orales des CRDO (Centre de ressources pour la description de l'oral)[13] de Paris et d'Aix est entré en production début mars 2010 après deux ans de phases de tests nécessaires pour implémenter notamment la prise en charge de formats audio et vidéos dans PAC.

Après validation du processus d'archivage et de diffusion grâce à ce pilote, le CINES et le TGE Adonis travaillent ensemble à l'industrialisation et à un passage à l'échelle.

La Bibliothèque Universitaire Pierre et Marie Curie est entrée en production le 1er février 2011 pour archiver durablement ses ouvrages numérisés sur les neurosciences, la biologie, la géologie et la physique-chimie.

La bibliothèque inter universitaire de droit Cujas en tant que pôle de référence, couvre l'ensemble des sciences juridiques. Elle archive au CINES depuis fin février 2011 les ouvrages fondamentaux en droit qu'elle numérise ainsi que certains documents pédagogiques.

2010

2011

La Bibliothèque Inter Universitaire Santé (BIUSanté) possède un fonds d'ouvrages spécialisés sur l'histoire de la médecine et de l'art dentaire. Elle a entrepris depuis quelques années la numérisation de ses collections : bibliothèque numérique Medic@, banque d'images et revues spécialisées numérisées. L'ensemble de ces documents sont archivés au CINES depuis juillet 2011 et représentent déjà près de 3 To de données.

La Bibliothèque interuniversitaire Sainte -Geneviève (BSG) poursuit la numérisation de ses collections de documents patrimoniaux. Elle les archive dans PAC depuis août 2011. À terme, cela représentera près de 4 To de données.

Depuis le 15 mars 2011, l'École française d'Extrême-Orient archive dans PAC un important fonds numérisé de photographies, plaques de verre et plaques stéréoscopiques, etc. prises depuis le 19ème siècle dans le cadre de missions patrimoniales en Extrême-Orient. Environ 17 000 images sur près de 85 000 ont déjà été archivées.

Marion MASSOL

Chef du projet PAC
CINES

Lorène BECHARD

Archiviste
équipe PAC
CINES

(13) Depuis fin 2011, le CRDO-Aix est devenu le Speech Language and Data Repository (SLDR). Le CRDO Paris est devenu COCOON (COllection de COrpus Oraux Numériques)

L'Agence Bibliographique de l'Enseignement Supérieur (ABES) : l'archivage des thèses dans PAC

FICHE DESCRIPTIVE

Établissement : Agence Bibliographique de l'Enseignement Supérieur

L'ABES a été créée en 1994 pour mettre en œuvre le Sudoc, catalogue collectif des bibliothèques de l'enseignement supérieur. Ouvert en 2001, le Sudoc est une réussite éclatante : avec 1 419 bibliothèques déployées auxquelles s'ajoutent les 1793 bibliothèques publiques ou privées du réseau Sudoc-PS, spécialisé dans le signalement des publications en série, plus de 10 millions de notices bibliographiques, 32 millions de documents localisés et 24 millions d'interrogations publiques en 2010, il occupe une place éminente dans le dispositif documentaire français de l'enseignement supérieur et de la recherche.

À partir de 2002, l'ABES a diversifié son action dans plusieurs domaines de l'Information Scientifique et Technique (normalisation documentaire, conversion rétrospective des catalogues, thèses, réseaux et projets coopératifs). Suivant les lignes structurantes de la bibliothèque scientifique numérique[14] (BSN), une très grande infrastructure, elle se coordonne avec les autres opérateurs.

- **Site Internet : http://www.abes.fr/**
- **Passage en production :** octobre 2006
- **Durée de l'avant-projet :** 2 ans
- **Volumétrie :** près de 15 000 thèses archivées 230 Go (au 31/12/2012)

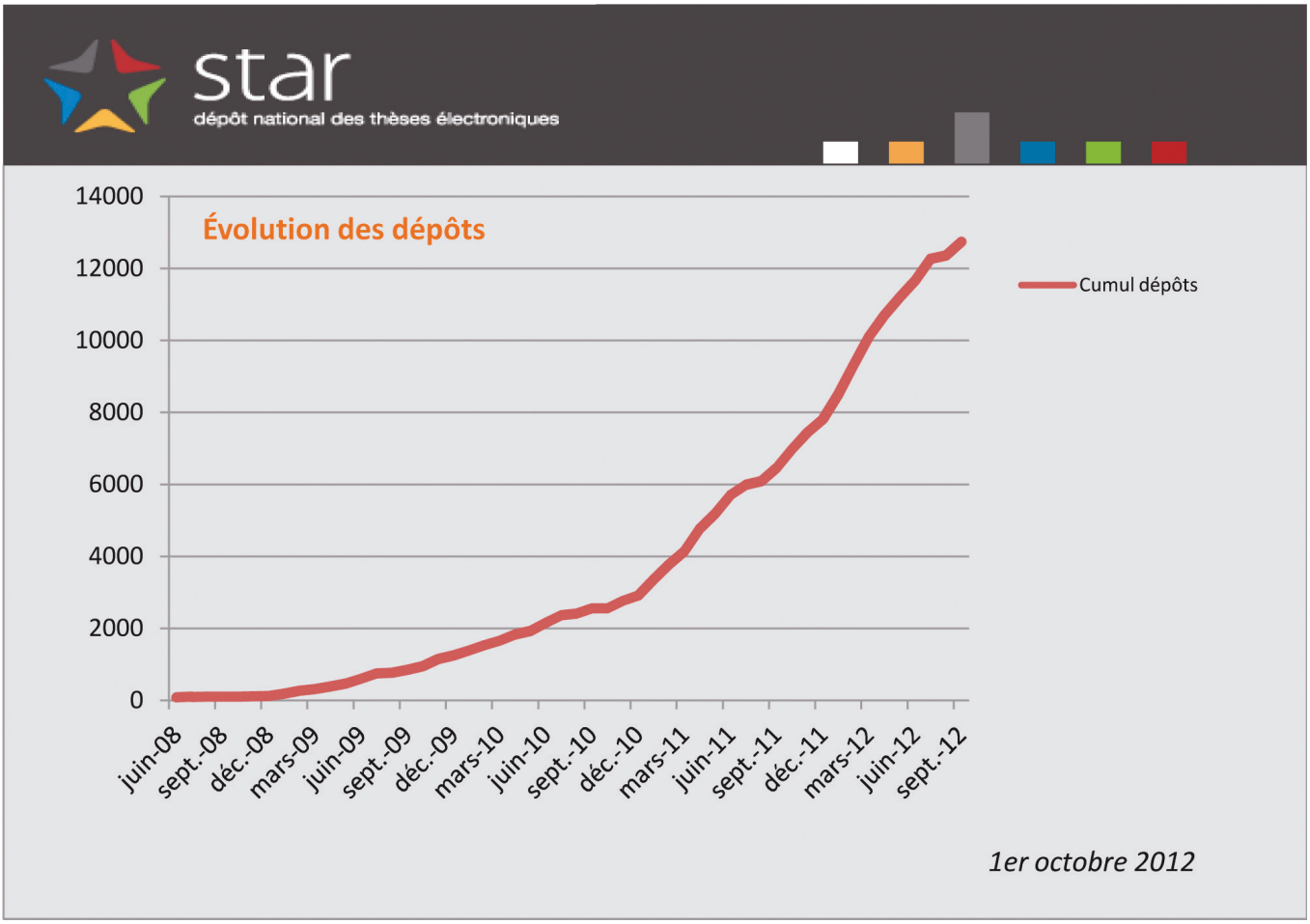
Spécificités

- Archivage patrimonial (thèses électroniques françaises)
- Premier service versant du CINES - projet pilote
- Seul établissement à avoir utilisé PAC-V1 et PAC-V2 (migration de plateforme)

« Depuis la publication de l'arrêté du 7 août 2006[15] relatif aux modalités de dépôt, de signalement, de reproduction, de diffusion et de conservation des thèses ou des travaux présentés en soutenance en vue du doctorat, chaque établissement habilité à délivrer des diplômes de doctorat doit faire le choix entre opter pour le dépôt des thèses sous leur forme papier ou sous leur forme électronique. Les établissements ayant fait le choix du dépôt électronique se voient offrir le service STAR[16] (application nationale pour le Signalement des Thèses, Archivage et Recherche développée par l'ABES) leur permettant de procéder au

(14) Cf. article « Les collaborations du CINES aux niveaux national et européen », page 49
(15) http://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000000635069&dateTexte=
(16) http://www.abes.fr/Theses/Les-applications/STAR

signalement et à l'archivage pérenne de ces documents. L'application a été ouverte en 2006. Depuis, 108 établissements ont choisi d'être déployés dans le réseau STAR, 78 sont effectivement en production et ont choisi officiellement le mode de dépôt électronique pour leurs thèses. La réflexion autour de ce nouveau support pour le signalement, la diffusion et l'archivage des thèses avait commencé avant la publication de cet arrêté. Dans le cadre du dépôt sous forme papier, chaque établissement avait la charge de l'archivage pérenne de ses thèses. Avec l'arrivée de l'électronique, il paraissait absurde d'imaginer que chaque établissement mette en place une plateforme d'archivage. C'est pourquoi, le CINES d'une part pour sa mission d'archivage et l'ABES d'autre part pour sa mission de signalement ont commencé dès 2004 à réfléchir à une organisation cohérente de ce nouveau circuit.



THÈSES DÉPOSÉES PAR STAR DANS PAC DEPUIS 2008

Un gros travail sur les métadonnées

Un groupe d'experts rassemblant des documentalistes, des bibliothécaires et des informaticiens issus de divers établissements dépendant du Ministère de l'Enseignement Supérieur et de la Recherche (dont l'ABES et le CINES) s'est penché sur ce qui fait la spécificité des thèses. Au carrefour de plusieurs mondes, la thèse est à la fois un objet documentaire, un document sanctionnant des études et une œuvre intellectuelle protégée par le droit d'auteur. Sous forme électronique, elle obéit à certaines caractéristiques techniques.

La recommandation TEF[17] (thèses électroniques françaises), jeu de métadonnées

XML, vise à prendre en compte ces différentes facettes :

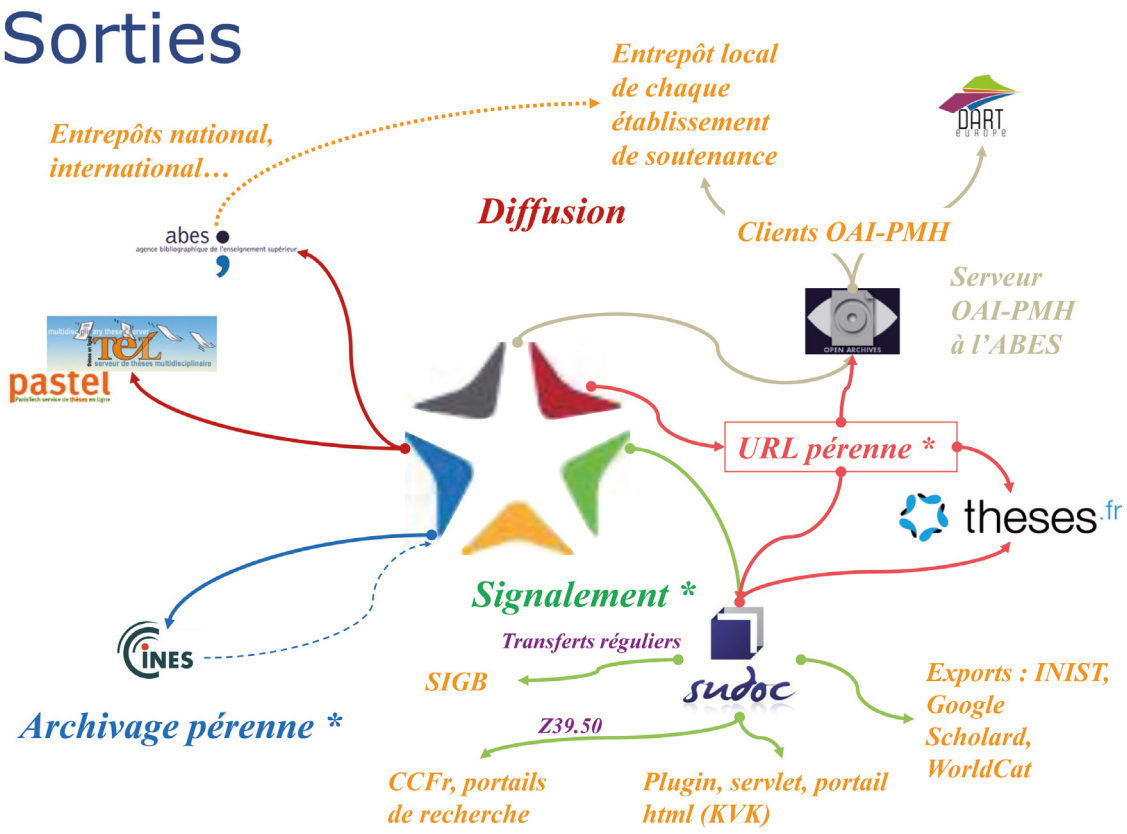
- *les métadonnées descriptives (résumé de la thèse, mots clés…) sont basées sur le Dublin Core ;*
- *les autres métadonnées, qualifiées de métadonnées de gestion, servent à décrire : le contexte administratif de la thèse (noms des membres du jury de soutenance, de l'école doctorale…), les conditions juridiques associées à la thèse (droits de propriété intellectuelle), et les impératifs de conservation (type d'encodage des caractères, format de fichiers…).*

Le tout est « encapsulé » dans un vocabulaire METS qui présente l'avantage de proposer une structure modulaire tout en gérant le(s) fichier(s) électronique(s) de la thèse. Ces métadonnées, certes spécifiques mais obéissant à des standards informatiques répandus sur internet, seront utiles à la diffusion de la thèse qu'il s'agisse d'une diffusion dans les établissements eux-mêmes ou dans des entrepôts (institutionnels ou non).

Format pivot, TEF permet d'assurer les fonctionnalités obligatoires de STAR décrites dans l'arrêté du 7 août 2006 :

- *Le signalement automatique des thèses électroniques dans le Sudoc, Système universitaire de documentation[18] (conversion TEF ➜ Unimarc)[19] ;*
- *L'archivage pérenne au CINES (conversion TEF ➜ PAC, plateforme d'archivage du CINES).*

La version 2 de la recommandation est parue en mars 2006. Tout était donc prêt pour la mise en œuvre des applications STAR et PAC dès la parution de l'arrêté.



EXPORTS GÉNÉRÉS PAR STAR

(17) http://www.abes.fr/abes/documents/tef
(18) Catalogue Sudoc : http://www.sudoc.ABES.fr
(19) http://fr.wikipedia.org/wiki/Unimarc

Les deux systèmes qui communiquent

L'application STAR (dans sa 1ère version) a été ouverte aux communautés en octobre 2006. De ce fait, dès l'ouverture de PAC-V1, au printemps 2007, STAR était en mesure de faire ses premiers versements automatiques. Au moment du changement de plateforme (PAC-V1 vers PAC-V2), le 24 juin 2008, nous avions déjà 85 thèses à migrer.

Un travail en partenariat

Nous avons, tout au long de ces phases de mise en œuvre, travaillé conjointement avec les équipes du CINES, avec beaucoup d'échanges, de mails, de réunions pour finaliser l'ensemble de la chaîne :

- *enregistrement par les établissements des thèses dans STAR ;*
- *envoi par STAR des métadonnées techniques au CINES (SIP)[20] avec le(s) fichier(s) thèse ;*
- *envoi par le CINES d'un accusé de réception ;*
- *après vérification de la validité du dépôt, envoi par le CINES du certificat d'archivage contenant l'identifiant unique et pérenne de la thèse archivée (AIP)[21].*

Au fur et à mesure des tests entre STAR, PAC et les établissements, il nous est apparu rapidement nécessaire, voir indispensable de donner à la communauté un outil permettant de vérifier, avant envoi, la validité des fichiers thèses que ce soit pour le format PDF ou pour le format XML (les 2 formats recommandés pour le corps de la thèse).

Dans STAR, nous avions introduit un contrôle avec l'outil Jhove[22] qui permettait un premier niveau de contrôle, mais qui s'est révélé largement insuffisant. Il est apparu plus cohérent que ce soit le CINES qui prenne en charge ce type de contrôles. Le CINES a développé un système plus complet (FACILE)[23] qui est librement accessible sur le web. Il permet aujourd'hui de vérifier la validité des fichiers à archiver.

En complément, nous avons également intégré dans STAR les contrôles FACILE afin de refuser les fichiers invalides dès le dépôt dans STAR. L'assistance de second niveau mis en place par le CINES ultérieurement, fournit aux établissements du réseau STAR une assistance précieuse en les aidant à remettre aux normes des fichiers invalides.

Aujourd'hui, nos deux établissements continuent ce travail de collaboration. Depuis les premières thèses archivées qui représentaient les premiers documents archivés par le CINES, la plateforme PAC reçoit des volumes de données bien plus importants que les thèses de STAR. Mais nous restons, historiquement, le premier partenaire du CINES pour l'archivage.

L'ABES et le CINES ont en commun cette mission nationale, il nous tient à cœur de la mener à bien pour l'ensemble de la communauté de l'enseignement supérieur. »

Marianne GILOUX

Ingénieur de recherche

Responsable fonctionnel de l'application STAR

Agence Bibliographique de l'Enseignement Supérieur

(20) Submission Information Package
(21) Archival Information Package
(22) http://hul.harvard.edu/jhove/
(23) Cf. article « L'expertise sur les formats de fichiers », page 36

La bibliothèque Cujas : l'archivage du fonds des ouvrages numérisés en histoire du droit

FICHE DESCRIPTIVE

Établissement : Bibliothèque Cujas

La bibliothèque Cujas possède l'un des fonds les plus importants de France en sciences juridiques et économiques. Elle est également dépositaire des publications de l'ONU, centre de documentation européenne et centre d'acquisition et de diffusion de l'information scientifique et technique (CADIST) en sciences juridiques.

- **Site Internet :** http://biu-cujas.univ-paris1.fr
- **Passage en production :** février 2011
- **Durée de l'avant-projet :** 8 mois
- **Volumétrie :** 150 documents archivés
  1 To (au 31/12/2012)

Spécificités :

- Bibliothèque numérisée par différents prestataires
- Projet d'archivage = assurance de la qualité globale des fonds (formats des fichiers, documentation du fonds…)

« La bibliothèque interuniversitaire Cujas (Paris 1 – Paris 2), bibliothèque de référence en sciences juridiques et CADIST en droit, a entrepris dès 2006 la numérisation d'une partie de son fonds patrimonial à partir d'une liste établie par deux enseignants en droit, MM Pfister et Roumy et au travers de projets tels que l'exposition virtuelle Jean Carbonnier. Actuellement, environ 150 documents numérisés sont en ligne sur notre bibliothèque numérique[24]. Bien que conscient de l'importance de l'archivage pérenne des données numérisées dès le début du projet, la collaboration avec le CINES n'a démarré concrètement qu'en octobre 2009. Elle a fait l'objet d'une convention entre les deux institutions en mai 2010. Fin 2011, tous les documents numérisés et en ligne sur la bibliothèque numérique utilisant XTF étaient archivés au CINES. Ce projet a été l'occasion de corriger les erreurs accumulées au cours de la numérisation et d'identifier les procédures à respecter impérativement lors des différents traitements des documents. En effet, l'automatisation des étapes de l'archivage pérenne ne tolère aucune erreur tant au niveau du nommage des répertoires et des fichiers que des formats des fichiers. Nous ne mentionnerons dans cet article que les phases essentielles du projet d'archivage.

Définition de l'objet numérique

Notre première question a été : « qu'allons-nous archiver ? ». L'unité intellectuelle que nous avons retenue était le volume physique traité (i.e. 1 volume physique = 1 paquet archivé). A celui-ci, correspond un numéro de code-barres qui nous sert d'identifiant tout au long du processus de numérisation. Cela signifie autant de paquets d'archives que de volumes pour une œuvre en plusieurs volumes par exemple. Suite aux recommandations faites par l'équipe du CINES, nous avons restreint l'archivage aux fichiers

(24) http://cujasweb.univ-paris1.fr/

suivants : les fichiers master (fichiers TIFF couleurs détourés avec une compression LZW), le fichier table des matières au format CSV (indiquant la concordance entre les pages et les images) et les fichiers XML de métadonnées (fichier Dublin Core, fichier METS).

La constitution du PPDI (Project Preservation Description Information)

La conception de ce fichier contenant les informations de pérennisation de niveau projet a permis à l'équipe « numérisation » de la bibliothèque de synthétiser et de pérenniser toutes les informations relatives au projet de numérisation. Cette étape descriptive très fine est l'occasion de retracer l'historique des choix techniques et intellectuels. Elle permet également, pour certains points, de formaliser ce qui a pu se construire de façon informelle au fil des mois.

La mise en œuvre de l'automatisation de la procédure

Une autre étape importante du projet d'archivage électronique était l'automatisation des procédures d'archivage afin de réduire à la fois la charge de travail de l'équipe informatique et les sources d'erreurs. Pour cela, les informations nécessaires à la génération du SIP (submission information package) ainsi que les fichiers à archiver sont récupérés automatiquement dans les répertoires sources de notre bibliothèque. Une fois les fichiers sip.xml et METS générés et après un dernier contrôle qualité, le SIP est envoyé automatiquement au CINES. Le certificat d'archivage (AIP - archival information package) est récupéré automatiquement dans une base de données qui permet de gérer les différentes versions des archives, si nécessaire.

Les enseignements du projet d'archivage

Nous avons pris le parti de concevoir, dès le départ du processus d'archivage, un contrôle qualité strict des différents fichiers du paquet d'archives afin de minimiser les rejets ou les versions successives.

De cet objectif général, a découlé une démarche qualité « rétrospective » sur l'ensemble de la chaîne de numérisation. Ce projet a permis à l'équipe d'identifier les lacunes ou erreurs qui s'étaient accumulées au fil du temps (lors de la numérisation et du traitement documentaire du premier lot archivé). Ceci a été l'occasion de nombreuses corrections mais surtout de la rédaction de procédures qualité aux principales étapes du projet de numérisation : contrôle des tables des matières, contrôle des notices Dublin Core, contrôle du nommage des fichiers... Un an après la mise en production de l'archivage, les procédures qualité ont été réunies dans un référentiel de numérisation qui permet de contrôler à chaque étape la qualité des données et des fichiers et qui constitue la mémoire de ce projet. De même, l'étape finale de l'archivage pérenne fait également l'objet d'une procédure et d'une « checklist » de contrôles à effectuer qui permettent de générer des paquets d'archives de plus en plus fiables. Cette dernière s'incrémente en fonction des dysfonctionnements rencontrés au cours de l'archivage.

Conclusion

Pour l'équipe, le projet d'archivage a été l'occasion de revenir sur le projet de numérisation afin de revoir et d'améliorer les différentes étapes du processus. »

Catherine RENARD

Professeur certifié affecté dans l'enseignement supérieur

Service informatique,
Bibliothèque
interuniversitaire Cujas

Brouk LEULMI

Ingénieur de recherche

Service informatique
Bibliothèque
interuniversitaire Cujas

CNRS : archivage des dépôts de l'archive ouverte pluridisciplinaire HAL (Hyper Articles en Ligne)

FICHE DESCRIPTIVE

Établissement :
Centre National de la Recherche Scientifique

Le CNRS, principal organisme de recherche à caractère pluridisciplinaire en France, a confié au CCSd - Centre pour la Communication Scientifique directe (une unité propre du CNRS créée fin 2000) la mission de faciliter la diffusion de la production scientifique et les échanges entre chercheurs, notamment par le biais de HAL.

L'archive ouverte pluridisciplinaire HAL est destinée au dépôt et à la diffusion d'articles scientifiques de niveau recherche, publiés ou non, et de thèses émanant des établissements d'enseignement et de recherche français ou étrangers et des laboratoires publics ou privés.

- **Sites Internet :** http://www.ccsd.cnrs.fr et http://hal.archives-ouvertes.fr
- **Passage en production :** mars 2010
- **Durée de l'avant-projet :** 6 mois
- **Volumétrie :** 215 000 articles archivés 475 Go (au 31/12/2012)

Spécificités :

- Nombreux documents de faible volumétrie
- Important arriéré avec des erreurs sur les fichiers PDF

« Le système HAL/Archives ouvertes développé au CCSd, accueille aujourd'hui des documents de tous les champs du savoir, expression des recherches menées dans les laboratoires des établissements publics de recherche (CEA, CEMAGREF, CIRAD, CNRS, INRA, INRIA, INSERM, IRD, PASTEUR,...), des universités ou des grandes écoles. Dans le but d'offrir un service plus complet et d'assurer la pérennité des documents scientifiques déposés dans l'archive HAL, le CCSd a passé un partenariat avec le CINES pour en assurer un archivage à long terme.

Quelles données sont archivées ?

Pour les documents texte, nous archivons le ou les fichiers au format PDF ainsi que les métadonnées le qualifiant dans un format XML. Celles-ci font référence à un schéma de métadonnées qui est vérifié lors de l'archivage de chaque document. Pour les images, nous archivons le fichier au format JPEG et au format TIFF s'il existe. Les documents archivés peuvent faire l'objet de nouvelles versions ou de mises à jour des métadonnées. Les métadonnées relatives aux documents déposés dans HAL sont généralement modifiées dans la période qui suit le dépôt. Pour éviter d'avoir un nombre trop important de mises à jour des métadonnées, nous avons décidé d'archiver les dépôts après un délai de 3 mois.

Mise en œuvre

Nous avons développé une application PHP interfacée avec notre application d'archives ouvertes HAL. Nous travaillons sur la même base

de données. Avant d'envoyer les documents pour archivage, nous effectuons une analyse d'éligibilité de nos fichiers à un archivage sur la plateforme PAC du CINES grâce à l'outil FACILE développé par le CINES (facile.cines.fr). En cas d'échec nous n'envoyons pas les documents à l'archivage et nous stockons l'erreur produite décrivant la raison du rejet.

Grâce à ce test préalable, nous avons pu constater un certain nombre d'erreurs sur les fichiers PDF existants. Or, nous n'avons aucun moyen d'influencer le processus de création des PDF que nous recevons. Nous avons ainsi remarqué une erreur récurrente au niveau des rejets qui concernait le format de date de création du fichier PDF présent dans les métadonnées du fichier : 'Improperly formed date' qui rendait de nombreux fichiers non archivables. Le CINES a mis au point des procédures de correction et nous avons lancé une campagne de corrections de PDF spécifique à ce type d'erreurs. Le résultat en a été très bénéfique car nous avons corrigé 92% des fichiers concernés, ce qui correspond environ à 7000 dépôts.

Nous utilisons le protocole SFTP pour les transferts. Nous analysons les mails renvoyés par l'application du CINES. Nous conservons dans une base de données un historique de tous les échanges. La procédure de versement est totalement automatisée. L'exploitation consiste actuellement en la surveillance et la mise à jour de la machine, en une analyse des logs, un suivi des volumes de rejets et en l'envoi ponctuel de documents.

Calendrier

La mise en place de ce projet aura duré 6 mois répartis en :

- *une phase préparatoire de conversion en version 1.2 des PDF existants dans des versions non acceptées par la plateforme d'archivage ;*
- *une phase de développement de l'application de sélection des dépôts, de transfert et d'historisation des échanges effectuée par un ingénieur à 80% ;*
- *une phase d'archivage massif de l'existant sur environ 3 mois soit 110 000 paquets d'archives versés ;*
- *et enfin une phase de correction des PDF en « Improperly formed date » sur 2 mois effectuée par un ingénieur à 50%.*

Nous sommes actuellement en phase d'exploitation avec des campagnes de mises à jour des métadonnées qui durent environ une semaine.

Volume

En décembre 2012, les lots archivés représentent un volume de 475 Go répartis en 215 000 lots. Nous avons un taux de rejet global d'environ 10%. Nous aimerions être capables à terme d'archiver tous les PDF déposés dans HAL.

Le retour de nos utilisateurs

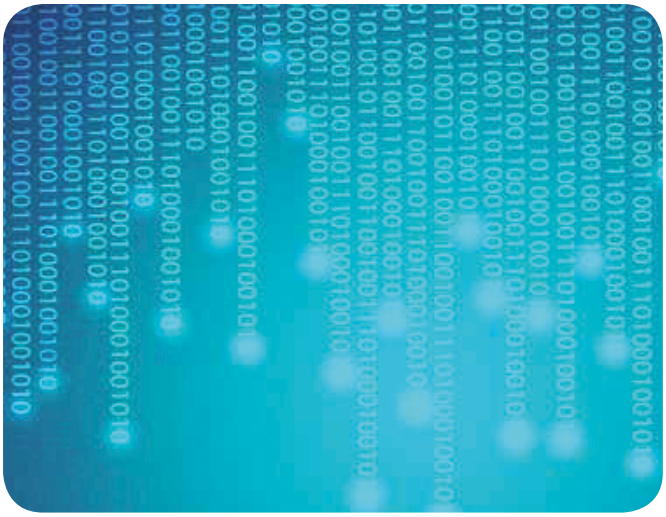
L'intérêt de l'archivage à long terme des données numériques n'est pas évident chez nos utilisateurs a priori. Lors de rencontres d'information sur HAL, nous expliquons pourquoi il nous semble important d'offrir ce service. Les retours sont alors extrêmement favorables et suscitent une motivation supplémentaire chez nos utilisateurs pour déposer leur production scientifique sur notre archive ouverte.

À NOTER : Lors du dépôt d'un document dans HAL, nous signalons à nos utilisateurs que leur fichier n'est pas éligible à l'archivage si tel est le cas. Nous les informons de la possibilité de vérifier leur fichier sur l'interface du CINES (FACILE). Cette mesure est uniquement informative car nous ne voulons pas mettre un frein au dépôt. »

Hélène GABORIAUD

Ingénieur
Chargée du projet d'archivage
CNRS - Centre pour la Communication Scientifique directe

La démarche qualité pour la préservation à long terme des données numériques



Introduction

L'information créée, manipulée, traitée, stockée, échangée, diffusée dans notre société est une information majoritairement sous forme numérique. Cette forme de représentation de l'information est incroyablement puissante et le coût de son stockage est de plus en plus faible. Il est à présent possible de la préserver sans la moindre altération et des outils existent pour créer des documents complexes et pour y rechercher les informations utiles. Et pourtant, en arrière plan de ces avantages immenses, se cache un risque majeur, celui d'une grande vulnérabilité au temps, explicable et avérée.

Cette vulnérabilité tient à la réalisation inéluctable[25], si aucune action préventive n'est prise, d'un ou de plusieurs risques liés à la nature même de l'information, à savoir :

- La connaissance perdue du contenu des objets numériques : pour pouvoir gérer et retrouver une information, il est nécessaire de disposer des détails permettant de la signaler et de la localiser. Sans cette description, la possibilité de retrouver l'information repose sur la mémoire de celui qui l'a créée.

- L'impossibilité de lire le format des fichiers contenant l'information : une information sous forme numérique n'est rien d'autre qu'une suite de 0 et de 1, et il faut se référer à un format standard du marché pour pouvoir en restituer le contenu informationnel. Or, ces standards existent en très grand nombre et sont aussi en constante évolution ; certains formats largement utilisés il y a quelques années ont aujourd'hui disparu.

- La dégradation et le vieillissement des supports de stockage : un dispositif de lecture complexe impliquant du matériel et du logiciel est indispensable pour l'accès à une donnée numérique. Ce dispositif peut s'altérer dans le temps, à cause d'effets physiques ou chimiques, tels que les rayonnements électromagnétiques, la chaleur ou encore la poussière, rendant impossible la lecture correcte de l'information.

- La disparition des matériels ou logiciels de lecture : la pérennité d'un tel dispositif de lecture dépend de celle de son fabricant ou de son bon vouloir ; le bon vouloir de l'éditeur de logiciel pouvant être de vous garder sous sa dépendance. L'indisponibilité d'un seul élément au sein du dispositif de lecture peut compromettre définitivement un patrimoine de données. Chacun sait à quelle vitesse évoluent les technologies numériques et à quel rythme les nouvelles éliminent les anciennes. Nous sommes confrontés ici à la faible pérennité des supports de données, à la disparition pure et simple des technologies de stockage les unes après les autres. Une telle disparition signifie en pratique que les matériels utilisés par cette technologie ne sont plus fabriqués et disparaissent du marché avec pour conséquence un accroissement exponentiel des coûts de maintenance des équipements existants.

(25) C. Huc, F. Banat-Berger, L.Duplouy, « L'archivage numérique à long terme, les débuts de la maturité ? », La Documentation Française, 2009

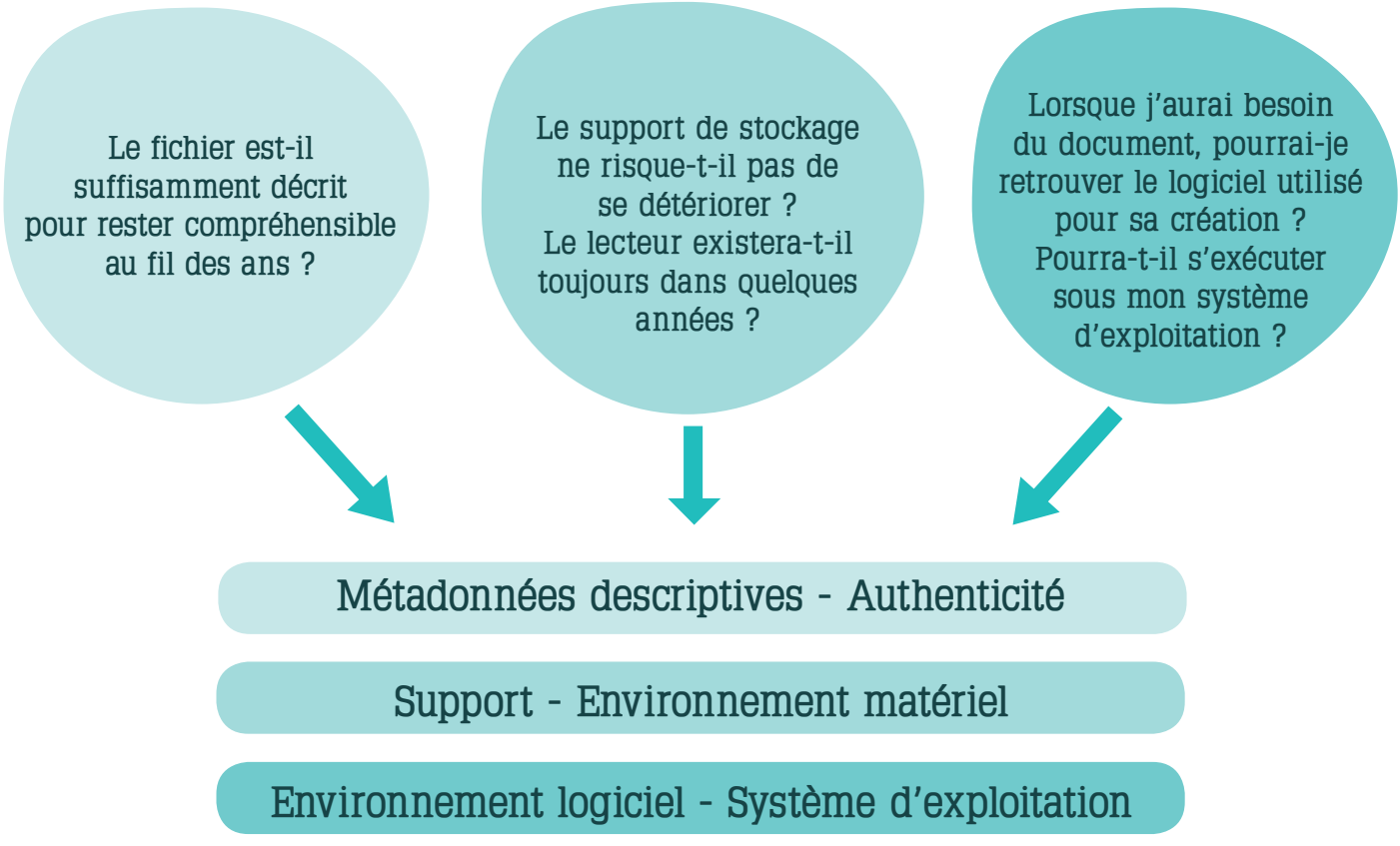
L'échelle de temps est donc ici un paramètre majeur eu égard au problème posé. Si on se situe à un horizon de l'ordre de dix ans, le problème est - relativement - simple à traiter. En effet, un stockage informatique de bonne qualité et sécurisé garantit contre la perte accidentelle du document. L'évolution des technologies n'aura sans doute pas conduit à un bouleversement tel que le document sera devenu irrémédiablement illisible. Et enfin, la communauté des utilisateurs potentiels du document sera vraisemblablement assez proche, scientifiquement et culturellement, de celle qui aura créé le document dix ans plus tôt.

Si on se place à un horizon de l'ordre de trente ans ou plus, rien de tout cela n'est plus assuré si personne ne s'est préoccupé d'accompagner le document dans le temps. C'est bien cet horizon du très long terme qui constitue le cœur du défi de l'archivage numérique pérenne, à savoir atténuer l'impact des quatre risques évoqués précédemment le jour où ils se réalisent, et ce sont bien les effets d'une démarche d'assurance qualité qui permettront de garantir l'intelligibilité et l'accessibilité des documents numériques.

Les méthodes essentielles à mettre en œuvre sont maintenant éprouvées :

- L'utilisation de métadonnées et d'identifiants persistants pour la capacité à garder la connaissance des contenus.
- Le choix de formats de fichier durables pour en garder la maîtrise et la capacité à les migrer vers de nouveaux formats (lorsque la conversion est la stratégie de préservation).
- La gestion proactive du vieillissement des médias de stockage pour la capacité à conserver correctement le train de bits constituant les fichiers et les migrer vers de nouveaux supports.
- La veille technologique permanente et l'anticipation des changements technologiques.

L'approche qualité au travers de bonnes pratiques dans ce domaine peut être abordée sous deux angles : technique et organisationnel.



L'ESSENTIEL

Bien qu'elle offre toujours plus de fonctionnalités pour un coût toujours moins important, l'information numérique est inéluctablement très vulnérable dans le temps. En effet, sans action préventive spécifique, elle s'expose à quatre risques principaux :

- la perte de la connaissance du contenu des objets numériques ;
- l'impossibilité de lire le format des fichiers ;
- la dégradation et le vieillissement des supports de stockage ;
- et la disparition des matériels ou logiciels de lecture.

L'atténuation de ces risques est le cœur de métier de l'archivage électronique. Pour cela, les méthodes éprouvées à mettre en œuvre sont :

- l'utilisation de métadonnées et d'identifiants persistants ;
- le choix de formats de fichiers durables ;
- la gestion proactive du vieillissement des médias de stockage ;
- la veille technologique permanente et l'anticipation des changements technologiques ;
- et la mise en place d'une stratégie autour de la qualité (formalisation des processus).

La qualité technique

L'approche technique de la qualité pour la préservation de documents numériques recouvre l'ensemble des procédures visant à garantir un haut niveau de qualité de l'objet numérique lui-même. Elle peut se décliner en trois niveaux.

La qualité des métadonnées

Les métadonnées sont les données qui permettent de préserver les informations décrivant les objets numériques ; on parle alors d'informations de pérennisation (description du contenu de l'information, de sa provenance ou sa source, et son historique) et d'informations de représentation (aspect technique de l'information, de sa structure ou sa forme, et de ses droits d'accès). Plusieurs types de contrôles peuvent être effectués, lesquels pourront être facilités par l'adoption de standards de métadonnées : en effet, les métadonnées normalisées sont décrites dans un référentiel qu'il est facile d'utiliser comme base de vérification pour s'assurer du niveau de qualité. L'éventail de tels jeux de métadonnées est très large, du plus générique pour la description de ressources numériques (ex. Dublin Core - ISO 15836 : norme utilisée pour la description des objets dans les paquets versés au CINES) aux plus spécifiques à un domaine précis comme le commerce électronique (ex. ebMXL, ISO 15000) ou les données géographiques (ISO 19115), en passant par les métadonnées techniques de préservation (PREMIS, METS) ou encore les métadonnées administratives pour la propriété intellectuelle et les droits d'auteur (MPEG-21).

La qualité des formats de fichiers

Pour pouvoir être lisibles et convertibles dans le temps, les fichiers doivent respecter scrupuleusement les spécifications de leur format ; des outils libres (Jhove, DROID), ainsi que des bases de données spécifiques aux formats (PRONOM) faisant autorité dans le domaine de l'archivage électronique permettent l'identification, la validation et la caractérisation des fichiers et assurent leur éligibilité à une conservation à long terme.

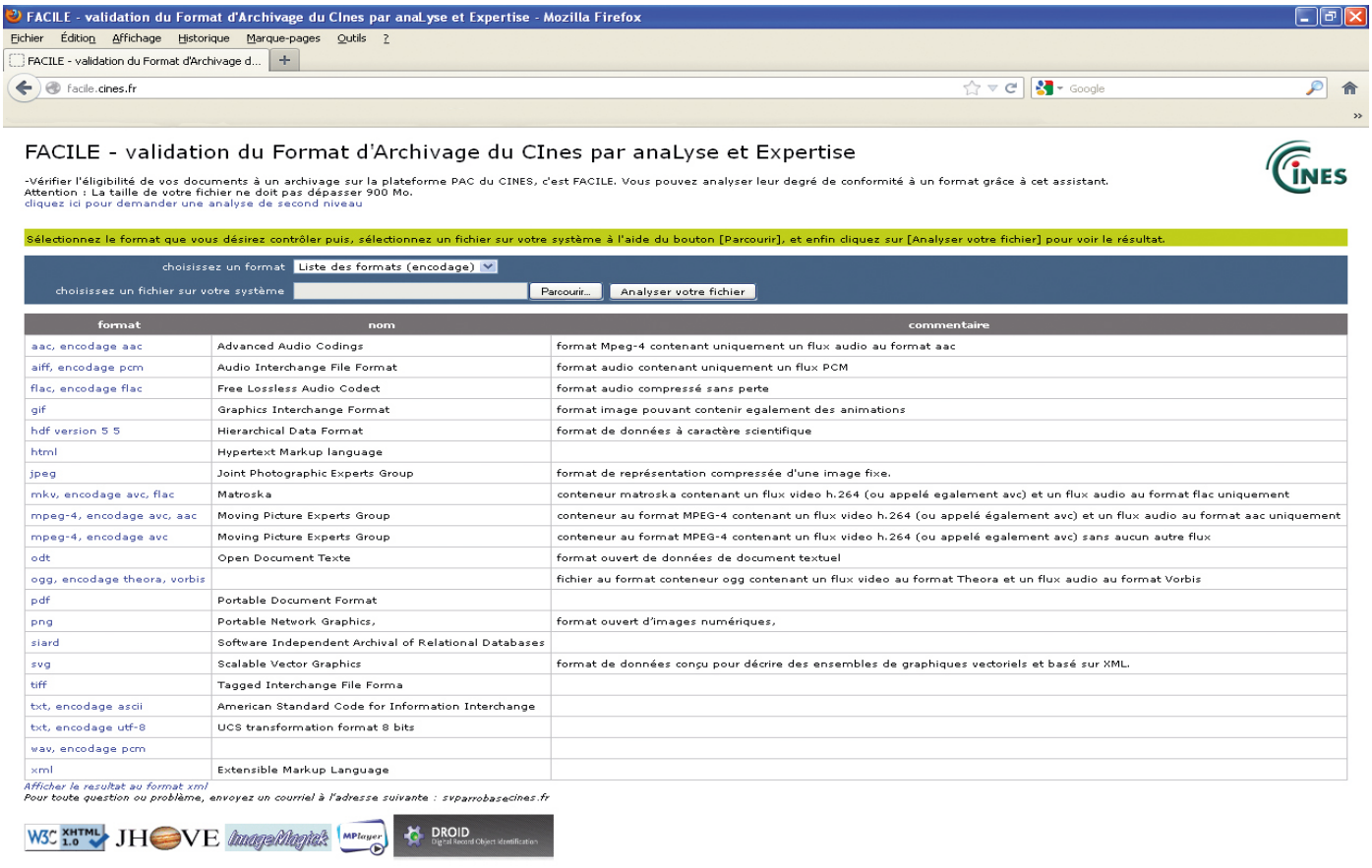
Afin de permettre le contrôle de la qualité d'un fichier, celui-ci doit donc être dans un format identifié et vérifiable. Il s'agit par conséquent de privilégier des formats de fichiers maîtrisés, c'est-à-dire dont les spécifications sont publiées (ex. PNG - ISO 15948) et normalisées si possible (ex. PDF - ISO 32000-1), et surtout largement utilisés par la communauté des producteurs d'information.

Dans cette optique, la plateforme de préservation numérique du CINES prend en charge une liste restreinte de formats de fichiers (une quinzaine au total à l'heure actuelle), et intègre ces logiciels libres pour effectuer toute une série de contrôles qualité sur les fichiers déposés avant leur archivage à long terme. Pour sensibiliser les utilisateurs potentiels des services d'archivage pérenne, des outils en ligne de contrôle qualité des formats de fichiers existent. Parmi eux, celui du CINES est destiné plus particulièrement à la communauté Enseignement Supérieur et Recherche.

L'outil FACILE

FACILE (validation du Format d'Archivage du CINES par anaLyse et Expertise, ci-dessous) est un projet initié dans le cadre de l'archivage des thèses électroniques. Développé par l'équipe du Département Archivage et Diffusion, il permet de valider la qualité du format des fichiers à archiver. Il intègre les mêmes outils et la même liste de formats que ceux qui sont supportés par la plateforme d'archivage du CINES.

En conséquence, beaucoup de producteurs d'archives utilisent FACILE pour détecter en amont de l'archivage la non-conformité de leurs fichiers. Le code de l'application est libre et ouvert, pour permettre une intégration dans les applications métiers d'autres établissements ou organismes intéressés. Par ailleurs, une assistance de deuxième niveau par l'équipe du CINES est possible pour résoudre les problèmes de non-conformité plus complexes.



L'OUTIL FACILE : HTTP://FACILE.CINES.FR/

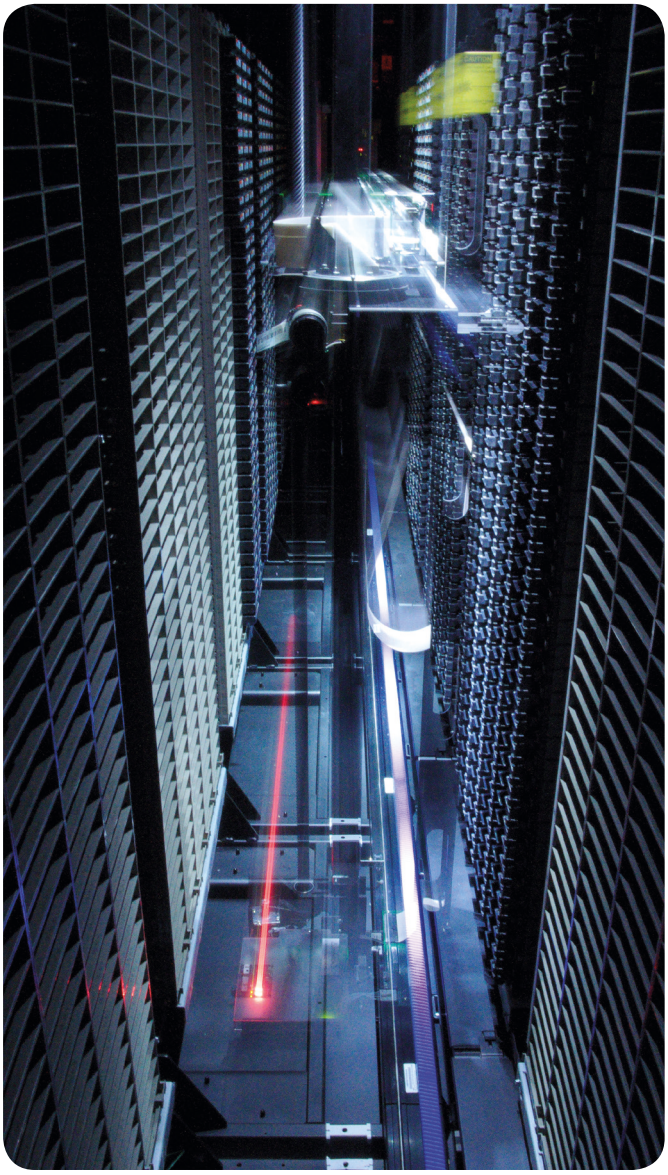
La qualité du stockage

La bonne conservation du train de bits qui compose les fichiers contenant les données à préserver est cruciale ; elle constitue le niveau le plus bas, le plus proche du média physique utilisé pour entreposer l'information. La première stratégie consiste à effectuer des copies multiples d'un même document, si possible dans un nombre supérieur à deux. La détermination du nombre de copies peut être problématique, notamment sur un plan financier, car lorsque la volumétrie des documents à préserver augmente, les coûts de stockage augmentent en proportion, réduisant potentiellement le nombre de copies[26] qui peut être réalisé.

Il convient également de d'utiliser plusieurs technologies de support de stockage, idéalement différentes, pour éviter d'hypothétiques erreurs de conception ou structurelles qui pourraient apparaître au fil du temps. Ainsi, le service PAC a une stratégie de stockage mêlant copies disques et copies bandes, avec un maximum de quatre copies pour un document archivé, dont une externalisée hors du CINES. Enfin, un audit régulier de toutes les répliques permet de détecter les dommages éventuels, et d'effectuer les actions correctives si besoin est.

Le contrôle de l'intégrité des fichiers permet donc d'anticiper la corruption de l'information. Il peut se faire au niveau matériel, par la vérification CRC par les contrôleurs de disques ou les contrôleurs réseau, ou au niveau logiciel par la vérification des sommes de contrôle (en anglais *checksum*) au travers du calcul des empreintes numériques par échantillonnage et comparaison avec l'empreinte initiale avec l'utilisation d'algorithmes de hachage (MD5, SHA-256, etc.).

Plusieurs études ont été menées sur le thème de la fiabilité des supports de stockage, leur résistance au vieillissement ou encore la corruption silencieuse des données, parmi lesquelles le rapport de l'Académie des Sciences sur la longévité de l'information numérique[27], les études des Archives de France sur les CD-R / DVD-R / Blu Ray et graveurs du marché (2008)[28] ou encore les études de l'Université Carnegie Mellon[29], Google[30] et



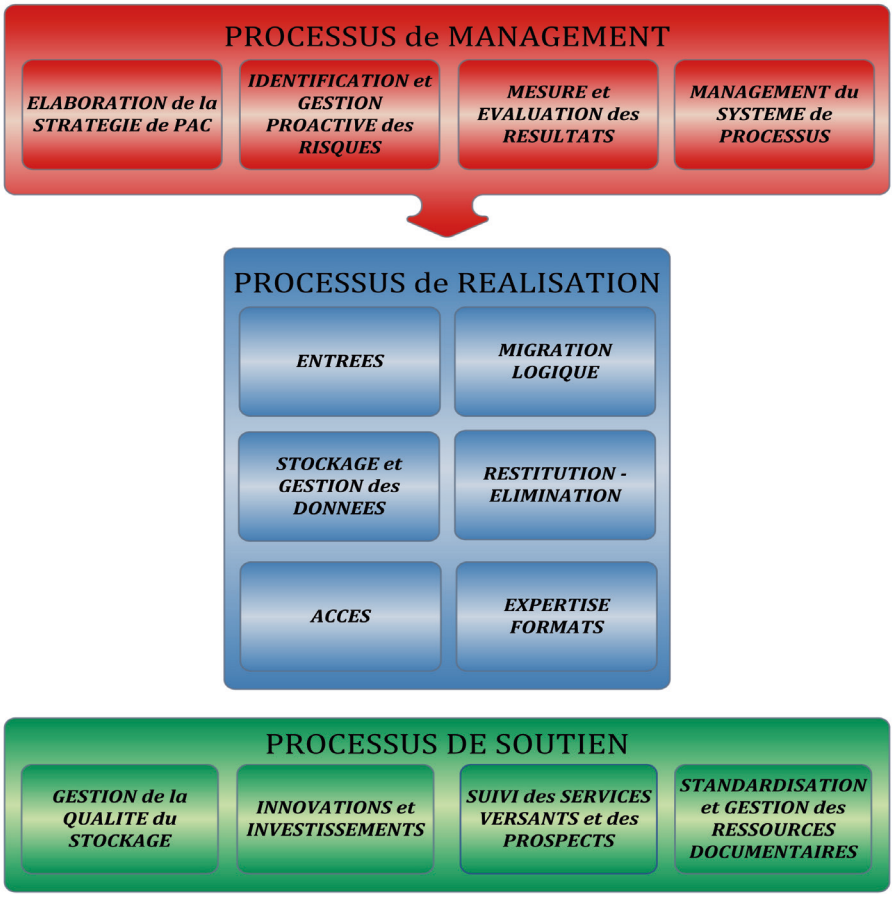
LE STOCKAGE SUR BANDES

(26) D. S. H. Rosenthal, « Bit Preservation : A Solved Problem? », Stanford University, 2008
(27) J.C Hourcade, F. Laloë, E.Spitz, « Longévité de l'information numérique, les données que nous voulons garder vont-elles s'effacer », EDP Sciences, 2010
(28) J.M. Lambert, J. Perdereau, « Qualité des CD-R disponibles sur le marché pour l'archivage des données numériques », « Qualité des DVD+R et des DVD-R disponibles sur le marché pour l'archivage des données numériques », « Qualité des disques Blu-Ray enregistrables pour l'archivage des données numériques », Ministère de la Culture et de la Communication – Direction des Archives de France, 2008
(29) B. Schroeder, G. A. Gibson, « Disk failures in the real world: What does an MTTF of 1,000,000 hours mean to you? », Carnegie Mellon University – Computer Science Department, 2007
(30) E. Pinheiro, W. D. Weber, L. A. Barroso, « Failure Trends in a Large Disk Drive Population », Google Inc, 2007
(31) L. N. Bairavasundaram, G. R. Goodson, B. Schroeder, A. C. Arpaci-Dusseau, R. H. Arpaci-Dusseau, « An Analysis of Data Corruption in the Storage Stack », University of Wisconsin-Madison, 2008

l'Université Wisconsin-Madison[31] sur les disques durs. Les résultats sont édifiants et incitent à la plus grande vigilance dans ce domaine. La qualité technique est donc incontournable, mais n'a de sens que si l'organisation du processus de préservation dans son ensemble suit les mêmes exigences d'excellence.

La qualité organisationnelle

La mise en place d'une démarche qualité au niveau de l'organisation même d'un service dédié à la préservation numérique est un projet global mobilisant l'ensemble des acteurs qui y sont impliqués. Trois thématiques applicables peuvent y être identifiées.



LA CARTOGRAPHIE MACROSCOPIQUE DE PROCESSUS DU SERVICE PAC

Les processus métier

Le premier axe d'une démarche qualité consiste à inventorier et documenter les différents processus métier[32] mis en œuvre pour la préservation à long terme de documents numériques. L'intérêt d'une telle initiative est multiple : la formalisation des processus métier constitue la colonne vertébrale de la documentation du service.

Dans la perspective d'une conservation de documents à très long terme, les informations relatives à toutes les procédures qui leur sont appliquées ont un sens évident, et constituent la base de la mémoire du service. Cette notion de transmission de la connaissance métier est particulièrement importante lorsque la rotation ou la mobilité des acteurs est importante (mutations, départs à la retraite, contrats de travail à durée déterminée, etc.), et sa bonne gestion est primordiale.

Ensuite, cette formalisation permet une auto-évaluation permanente de l'organisation, par la revue et la critique régulière, et donc une amélioration continue des performances globales du système et du service mis en œuvre.

Dans une même perspective d'évaluation et d'amélioration, elle permet une comparaison plus efficace du mode de fonctionnement de deux structures qui auraient une démarche similaire de documentation de leurs processus métier.

Enfin, la formalisation des processus métier est un excellent support pour l'amélioration de la communication interne ou

(32) J. Segot, « Outils de management : management des processus », Fascicule de documentation FD X50-176, Afnor, 2009

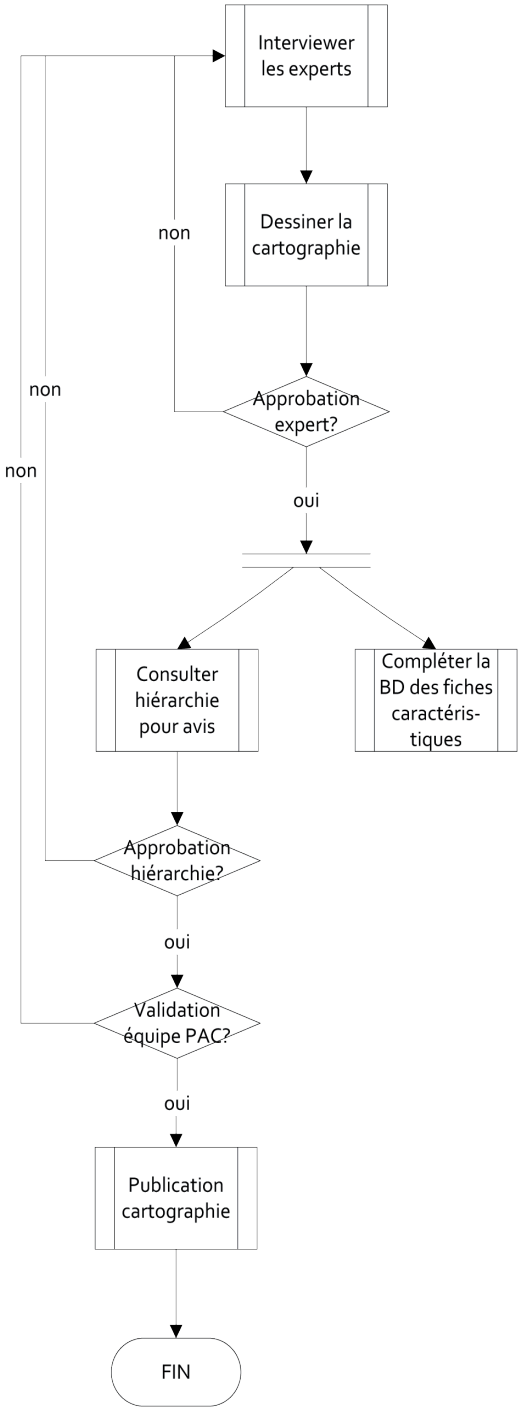
externe : elle constitue un moyen de lutte contre le cloisonnement interne interservices, ou permet une plus grande transparence dans les méthodes de travail, ce qui est un point essentiel pour établir et maintenir des relations de confiance avec les partenaires, les tutelles ou les utilisateurs.

Une telle démarche débute avec le choix d'un ensemble de règles de formalisation et de relations entre les objets à décrire : on parle de métamodèle. Plusieurs formalismes existent, des plus génériques (ex. UML, ISO 9001 ou ISO 9004, etc.) aux plus complexes (ex. BPMN), si bien qu'il peut s'avérer difficile d'en choisir un qui convienne au domaine spécifique de la préservation numérique, même chez les éditeurs de solutions logicielles de BPA. Elle se poursuit avec la décomposition des fonctions métier de l'archivage en processus, sous-processus, activités, et l'établissement d'une cartographie générale des processus au sein de laquelle les processus de réalisation, de soutien ou de pilotage sont différenciés.

Chaque processus identifié est à rapprocher des groupes fonctionnels OAIS (ISO 14721) et de ses trente-trois groupes fonctionnels théoriques. Dans le cas de l'implémentation faite au CINES, quatorze processus identifiés ont été jugés applicables (cf. schéma de la page 31). Un pilote est affecté à chaque processus à documenter, dont il a si possible la responsabilité dans son exécution quotidienne, et qui est chargé de le formaliser et le mettre à jour au fur et à mesure de l'avancée du projet.

Les différentes étapes de la description d'un processus sont résumées ci-contre :

- La formalisation du processus par l'expert et/ou le pilote au travers de réunions, d'entretiens, et à partir des cartographies existantes. Dans son implémentation, le CINES a choisi un métamodèle basé sur des représentations graphiques (figure ci-contre) pour plus de lisibilité, et des descriptions complémentaires (ou fiches descriptives) pour une plus grande conformité avec les normes ISO 9001 et 9004.
- La validation par l'ensemble de l'équipe impliquée après corrections éventuelles, pour une plus grande responsabilisation et un renforcement de l'expertise de chacun, grâce à une incitation à la réflexion sur la méthodologie quotidiennement adoptée.



LA MÉTHODOLOGIE DE DESCRIPTION D'UN PROCESSUS

(33) M. Massol, O. Rouchon, « Quality insurance through business process management in a French Archive», iPRES 2010, http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/massol-6.pdf

La validation de la cartographie et des caractéristiques associées par la hiérarchie, et la publication éventuelle du système de processus ainsi constitué.

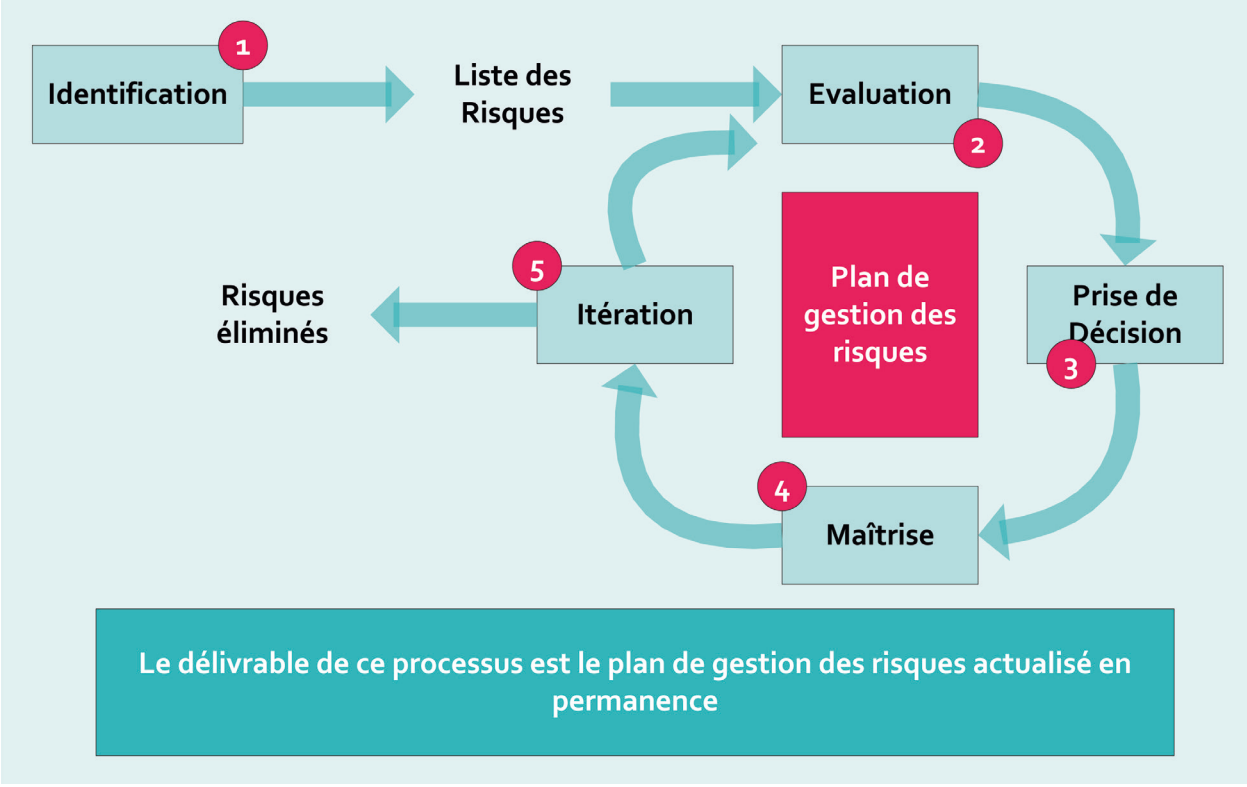
Bien entendu, une veille régulière sur cette cartographie est nécessaire. Les processus ne sont jamais tout à fait figés et évoluent dans le temps ; le référentiel doit être mis à jour en fonction des changements apportés.

Il va de soi que cette initiative requiert l'adhésion et l'implication de tous les agents impliqués, faute de quoi les chances d'aboutir à un résultat concret sont minces. L'implémentation au CINES, démarrée en 2009, a duré près de dix-huit mois. Aujourd'hui, les cartographies comme les fiches descriptives sont disponibles en ligne sur le site institutionnel du CINES et ont fait l'objet de présentation lors de conférences internationales sur la préservation numérique[33].

La gestion des risques

La méthodologie de pilotage de projet par la gestion des risques est issue du monde industriel, et est aujourd'hui très répandue dans les secteurs du management et des services. La préservation des documents numériques est une action principalement préventive, d'où la nécessité d'intervenir avant que les dégradations se produisent. C'est un ensemble de compromis à la fois entre les impératifs immédiats des producteurs d'information et les besoins à long terme des utilisateurs, et entre les besoins et les moyens mis en place.

La préservation des documents numériques est un projet comme un autre, et comme toute autre activité, elle génère des risques. L'objectif n'est pas ici de les supprimer, mais bien de déterminer le niveau de risque acceptable. Une telle méthode est bien définie et a déjà fait ses preuves.



LE PROCESSUS DE LA GESTION DES RISQUES

En préambule, il convient de définir le contexte d'application et de fixer les objectifs de la gestion des risques ; puis les différentes phases sont les suivantes :

- L'identification et la catégorisation des risques. Un risque est un évènement qui, lorsqu'il se réalise, peut empêcher la bonne réalisation de la mission d'un service ; il est défini par la combinaison d'une vulnérabilité et d'une menace.

- L'évaluation des risques, par l'analyse de la probabilité et de l'impact de chaque risque dans le temps, et la combinaison de ces deux facteurs. Il existe un troisième critère d'évaluation pouvant être pris en compte : le degré d'apparition dans le temps, évalué selon le degré d'imminence du risque dans l'évolution temporelle. L'indice de priorité des risques (IPR) peut donc être calculé soit sur deux dimensions (impact x probabilité) comme cela est fait au CINES, soit sur trois dimensions (impact x probabilité x degré d'apparition dans le temps) comme souvent évoqué dans les présentations de la BnF sur la gestion des risques appliquée à la conservation numérique, ou celles de la démarche AMDEC, plus destinée au monde industriel.

- La prise de décision : il s'agit d'identifier les risques prioritaires, en fonction du niveau à partir duquel ils deviennent acceptables, les moyens de leur traitement et le plan d'action ; pour cela, une mise en perspective avec les objectifs et les activités de l'établissement, les coûts et bénéfices des opérations de maîtrise ou encore les obligations légales peut être pertinente.

- La maîtrise des risques, par l'assignation d'une personne physique comme son propriétaire, et la mise en place des actions nécessaires pour diminuer le niveau de risques, soit en diminuant leur probabilité, soit en diminuant leur impact.

Un tel processus est itératif, et n'est efficace que s'il est réévalué régulièrement, car certains risques résiduels peuvent être éliminés alors que d'autres apparaissent au cours du fonctionnement du système. Dans le cas du service PAC du CINES, le plan de gestion des risques, mis en place en 2009, est révisé tous les six mois par l'ensemble de l'équipe.

La certification

La certification est l'aboutissement de la consolidation d'un organisme et/ou d'un service. Elle matérialise la reconnaissance de la qualité et du professionnalisme, donc un moyen d'instaurer des relations de confiance avec les communautés d'utilisateurs, et peut constituer, en outre, un levier pour obtenir des budgets auprès des organismes de tutelle. Plusieurs types de certification sont envisageables :

- Les certifications généralistes, telles que ISO 27000 (sécurité informatique), ISO 9000 (qualité), CMMI(34) (ingénierie), ITIL(35) (services), etc.

- Les certifications spécifiques à l'archivage pérenne, parmi lesquelles l'ISO 16363 (European Audit Framework) avec le CCSDS(36)et la marque NF461 d'AFNOR Certification (basée sur la NF Z42-013) ; et des accréditations : Data Seal of Approval (bonnes pratiques), DRAMBORA (37) (gestion des risques), TRAC(38) (liste de critères).

La démarche de certification est un projet conséquent, tant sur le plan de l'investissement humain qu'elle requiert pour la conduite du projet ou le pilotage des changements requis, que sur le plan de l'investissement financier puisque des audits externes réguliers sont à prévoir.

« Il convient par conséquent de bien identifier le type de certification qui aura le plus d'impact à la fois sur la communauté d'utilisateurs et sur les organismes de tutelle. »

(34) Capability Maturity Model Integration
(35) Bonnes pratiques du management du système d'information (Information Technology Infrastructure Library)
(36) Consultative Committee for Space Data Systems
(37) Digital Repository Audit Method Based on Risk Assessment
(38) Trustworthy Repositories Audit & Certification

Le service d'archivage du CINES a procédé à quatre audits internes ou externes en autant d'années : le premier pour identifier les forces et faiblesses du département dans une perspective de certification, les suivants pour obtenir l'agrément SIAF pour la conservation intermédiaire d'archives publiques ou encore pour tester la future norme ISO 16363 dans le cadre d'un projet – APARSEN – financé par la Commission Européenne.

Conclusion

De la qualité des objets numériques dépend grandement la facilité de leur préservation dans le temps ; il ne faut toutefois pas réduire cette qualité au seul aspect technique des objets numériques, tant la qualité des processus de conservation qui leur sont appliqués est toute aussi cruciale.

Une démarche qualité pour la conservation à long terme représente un investissement conséquent et immédiat dont les effets ne seront perceptibles qu'à long terme ; de plus, une telle initiative, et son aboutissement qu'est la certification, requiert l'adhésion de toutes les ressources humaines impliquées dans le processus de conservation.

Pourtant, les indicateurs et les métriques sur la qualité ne sont pas encore clairement définis.

Alors, rendez-vous dans trente ans ?

Olivier ROUCHON

Responsable du Département Archivage et Diffusion
CINES

(39) Consultative Committee for Space Data Systems
(40) http://www.alliancepermanentaccess.org/index.php/current-projects/aparsen/

LE CINES TESTE L'ISO 16363 SUR LA CERTIFICATION DES SERVICES D'ARCHIVAGE ÉLECTRONIQUE

Publiée en février 2012, l'ISO 16363 « Audit and Certification of Trustworthy Digital Repositories » est une initiative du CCSDS[39] pour identifier des institutions d'archivage pérenne de confiance en définissant et recommandant des bonnes pratiques sur lesquelles baser un processus d'audit et de certification, et en évaluant le niveau de confiance d'un système d'archivage électronique.

Dans le cadre de l'initiative européenne APARSEN[40], le CINES a été retenu parmi six autres centres d'archivage dans le monde pour tester le guide d'audit avant sa publication et évaluer la pertinence et l'applicabilité des critères listés dans le référentiel. L'expérience s'est matérialisée par la réalisation de deux audits en 2011 :

- un interne, sur la base du référentiel proposé par la norme : 91 critères d'évaluation, répartis en trois sections : la structure organisationnelle de l'institution (24 critères), la gestion opérationnelle des objets numériques (43 critères) et la prise en compte des risques liés à l'archivage (24 critères). Pour chaque critère, la norme précise ce qui est attendu et donne également des exemples des preuves qui doivent être fournies afin d'évaluer le degré de complétude.

- et un externe effectué par douze experts internationaux indépendants, sur la base du rapport final d'audit interne.

Outre le test de l'ISO 16363, le projet APARSEN vise la mise en place au niveau européen d'un cadre pour l'audit et la certification de services d'archivage électronique, qui fédèrerait des accréditations et certifications avec des degrés de difficultés différents, tels que le Data Seal of Approval et l'ISO 16363.

L'expertise sur les formats de fichiers : une spécificité de l'offre d'archivage du CINES

Qu'est-ce qu'un format numérique ?

Un format informatique est une convention pour représenter une donnée sous forme numérique. Il peut être spécifié, ouvert, normalisé, standardisé, propriétaire. On parle de format spécifié lorsqu'il est suffisamment décrit pour en développer une implémentation complète. Un format est ouvert s'il est légalement exempté de droits d'utilisation et que sa description est publique. Ceci le rend alors compréhensible et interopérable.

Il n'est cependant pas simple de qualifier précisément un format car il peut être structuré à l'aide de plusieurs autres formats ; les formats de plus haut niveau étant structurés à partir de ceux de plus bas niveaux. Le XML utilise le langage SGML qui utilise les formats UTF8 ou ASCII. Un format peut également contenir d'autres formats, par exemple le "Matroska" peut contenir, entre autres, un flux vidéo (H.264) et un flux audio (AAC) ainsi que des sous-titres, etc.

L'ESSENTIEL

Un format de fichiers est une convention pour la représentation d'une donnée sur un support numérique. Toutefois, en pratique, un grand nombre de fichiers ne sont pas totalement conformes aux spécifications dont ils se prévalent. En conséquence, afin de faciliter les futures lectures et conversions, les services d'archives tels que PAC :

- sélectionnent des formats dits « archivables » selon certains critères (format spécifié, ouvert, normalisé, standardisé, propriétaire…)
- et contrôlent en conséquence la qualité du (des) fichier(s) candidat(s).

Pour finir, il est important de décorréler la notion de format de celle de fichier. En effet, la description d'un format peut porter sur un ou plusieurs fichiers voire même sur un flux numérique. La présence de plus en plus importante de plateformes nomades en interaction avec des bases de données et des streaming nous invite à penser que la notion de fichier a une place de moins en moins prépondérante.

Les critères de sélection d'un format à l'archivage

En tant qu'acteur reconnu du domaine de la préservation à long terme des documents numériques, le CINES s'est doté d'une cellule d'expertise dédiée aux formats numériques.

> *"Pour qualifier un format de pérenne, le CINES impose qu'il soit ouvert, largement utilisé et contrôlable."*

De plus, la politique de sélection des formats éligibles à l'archivage tend à minimiser le nombre de formats différents, pour en faciliter la gestion en termes de migration et de veille technologique. Aussi, tout document soumis à archivage passera par une phase d'assurance qualité dont l'objectif est de voir si le document est bien en accord avec sa spécification.

Il s'agit de s'assurer que le fichier est valide et bien-formé par rapport à la spécification de son format :

- Un fichier est considéré comme "bien formé" s'il répond aux exigences purement syntaxiques.
- Un fichier est considéré comme "valide" s'il répond aux exigences sémantiques du format.

Sous ces conditions, il sera alors toujours possible à l'aide de la spécification de comprendre l'organisation des données dans le format et ainsi de l'exploiter. De plus, le fait de s'assurer qu'un fichier est en accord avec ses spécifications nous donne les meilleures chances de migrer sans erreur ce format vers un autre lorsque ce sera nécessaire.

Cependant, un grand nombre de fichiers générés par diverses applications ne sont pas produits de manière strictement conforme à leur spécification et même si ces fichiers restent exploitables, ils présentent un risque quand à leur lisibilité ou leur migration future.

Aussi la cellule d'expertise en formats de fichiers s'est dotée de moyens humains, techniques et organisationnels dont le but est de permettre non seulement de qualifier la conformité d'un document par rapport à sa spécification mais aussi de fournir une aide aux utilisateurs en apportant des corrections aux fichiers non conformes.

FACILE

La Validation du Format d'Archivage du CINES par anaLyse et Expertise propose un moyen très simple, via une simple interface web de vérifier l'éligibilité d'un document à un archivage sur la plateforme PAC. Cette application repose sur le validateur de formats utilisé dans PAC afin de rendre un résultat strictement conforme à une analyse dans le cadre d'un archivage.

Cet outil propose également la liste des formats archivables et les outils utilisés pour les valider. Il indique si un fichier est conforme au format que vous sélectionnez et vous précise l'erreur identifiée.

En cas de problème, l'utilisateur peut faire une demande d'assistance au travers d'une interface dédiée : c'est « l'expertise FACILE de second niveau ». La cellule d'expertise formats du CINES accompagnera alors l'utilisateur dans la résolution de son problème.

Pour plus de renseignement : http://facile.cines.fr

LA CONVERSION DE FORMATS

La sélection des formats de fichiers avant l'archivage n'occulte pas la problématique de l'obsolescence inéluctable de tout format, elle en retarde seulement la date. Aussi, le CINES a également mis en place une stratégie de pérennisation qui repose sur la migration logique ou conversion de formats. Ce processus organise et planifie l'ensemble des étapes à suivre pour la conversion d'un format de fichiers jugé non pérenne en un autre format jugé pérenne, de manière à assurer la lisibilité des informations contenues dans le fichier. Les services versants sont associés à cette démarche afin d'identifier les caractéristiques essentielles à préserver lors de la conversion de formats et de valider les choix retenus.

Une nouvelle norme ISO sur le changement de supports et conversion de formats de fichiers…

Utilisée par le CINES, la norme ISO 13008, parue en juin dernier aborde la question des migrations de supports et de formats sous un angle à la fois généraliste et pratique. Elle présente tout d'abord les aspects managériaux de la procédure en termes de gestion des risques, de planification, de responsabilité et de documentation. Elle décrit ensuite les étapes qui composent toute démarche de migration.

Philippe PRAT

Expert formats de fichiers
équipe PAC
CINES

L'archivage des données administratives : 5 questions à Charlotte Maday

Charlotte MADAY

Présidente de la section AURORE (Archivistes des Universités, des Rectorats, des Organismes de Recherche et autres Établissements d'enseignement supérieur) de l'Association des Archivistes Français (AAF)

QU'EST-CE QU'UNE ARCHIVE PUBLIQUE ?

Le Code du Patrimoine (art L211-1 et suivants) définit les archives publiques comme étant « *l'ensemble des documents, quels que soient leur date, leur lieu de conservation, leur forme et leur support, produits ou reçus par toute personne physique ou morale et par tout service ou organisme public ou privé dans l'exercice de leur activité [...], documents qui procèdent de l'activité, dans le cadre de leur mission de service public, de l'Etat, des collectivités territoriales, des établissements publics et des autres personnes morales de droit public ou des personnes de droit privé chargées d'une telle mission* ».

1 CINES : Quelles archives peut-on rencontrer en établissements d'enseignement supérieur et recherche ? Sous quelle forme ?

Charlotte Maday : Les activités d'un tel établissement sont d'abord axées sur le domaine pédagogique et le domaine de la recherche scientifique ; ces deux activités étant accompagnées et soutenues par des infrastructures administratives (services centraux, services communs etc...). Les archives sont donc logiquement classées en fonction de ces 3 grandes catégories :

- *les archives administratives procèdent de l'activité des services de scolarité, de ressources humaines, de finances, etc.*
- *les archives pédagogiques sont souvent identifiées comme les documents supports de cours, et plus généralement les données et documents issues de l'activité d'enseignement ;*
- *les archives de la recherche sont assimilables à tous les documents produits par les chercheurs dans le cadre de leur activité de recherche : cahiers de labos, résultats d'expériences, journaux de fouilles, calculs, etc.*

On peut trouver ces documents, ces informations, sous toutes les formes possibles : papier, électronique, vidéo, son, matériel de laboratoire, etc.

2 CINES : Comment s'organise la collecte de ces archives ? Par quels moyens ? Qui s'occupe des archives ?

Charlotte Maday : Les structures d'archives en établissement d'enseignement supérieur et de recherche ont pour mission la collecte, le classement, la conservation et la communication des archives de l'établissement, conformément à la législation et la réglementation en vigueur.

Lorsqu'un archiviste arrive en poste, il procède d'abord à un état des lieux puis recense tous les documents stockés (ce qu'on appelle l'arriéré d'archives) en examinant leur contenu. Lorsque tous les documents sont identifiés, ils sont triés selon leur valeur administrative ou historique : les documents qui n'ont plus aucune valeur directe pour l'établissement ou pour son histoire sont éliminés selon une règlementation très précise et

des procédures strictes. Après la reprise de l'arriéré, la collecte s'organise directement dans les services, chaque année, selon des calendriers établis en collaboration avec l'ensemble des acteurs de la chaîne d'archivage.

L'archiviste bénéficie ainsi d'un encadrement apporté par les normes, les lois et les règlements, pour élaborer des procédures qui seront par la suite appliquées à l'ensemble de l'établissement.

3 CINES : Quel est l'intérêt d'une telle collecte et conservation ?

***Charlotte Maday** : Pour les archives de la recherche, il s'agit d'abord de protéger le patrimoine de la recherche et les chercheurs : cette démarche d'archivage des documents scientifiques est d'ailleurs complémentaire aux actions des structures de direction de la recherche. Les services d'archives offrent ainsi des infrastructures sécurisées (ou en voie de l'être) pour conserver des documents et surtout, en garantir le contrôle de l'accès (qui, comment, dans quelles conditions). Par la suite, ces documents de la recherche serviront de preuve ou de témoin pour reconstituer l'histoire scientifique d'une discipline, d'un chercheur, d'une institution, d'un pays.*

Pour les archives administratives, l'intérêt de l'archivage est immédiatement compréhensible : c'est celui de démontrer que les établissements satisfont aux exigences légales et réglementaires et à l'obligation de service public, dans le cadre d'une bonne gouvernance.

Pour les archives pédagogiques, la collecte et la conservation de ces documents permet entre autres d'enrichir l'histoire de l'enseignement et de donner des éléments sur l'évolution des contenus et du savoir.

Une fois collectées et classées, les archives historiques de ces établissements sont transmises aux Archives Départementales ou Archives Nationales, pour une conservation pérenne.

L'AGRÉMENT SIAF POUR LA CONSERVATION D'ARCHIVES PUBLIQUES COURANTES ET INTERMÉDIAIRES

En septembre 2009, un décret d'application (2009-1124) est paru concernant la compétence des services d'archives publics et la coopération entre les administrations pour la collecte, la conservation et la communication des archives publiques. Il permet à des institutions telles que le CINES de conserver sur le territoire national les archives qui lui sont confiées, dans des locaux conformes aux prescriptions du Service Interministériel des Archives de France (SIAF) à condition d'obtenir un agrément.

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE

Ministère Culture Communication

Ainsi, en 2010, le CINES a constitué un dossier de demande d'agrément qui a fait l'objet d'une étude de plusieurs mois par le SIAF. Il a ensuite été audité par une commission composée d'une dizaine d'experts nationaux et a obtenu son agrément le 14 décembre 2010 pour une durée de trois ans.

Cet agrément a constitué une étape préalable dans le processus de mise en place de conventions d'archivage comme celle entre le CINES et le CNRS par l'intermédiaire du TGE (Très Grand Equipement) Adonis.

4 CINES : Quels problèmes rencontre-t-on dans la collecte et la conservation ?

Charlotte Maday : Il faut tout d'abord rappeler que la présence d'archivistes en universités est relativement récente et donc que tout le travail de collecte et de conservation reste à faire dans la majorité des universités qui n'ont pas encore de service d'archives.

En ce qui concerne la collecte, on se rend compte que le principal problème rencontré par les archivistes a tendance à s'estomper dès lors que leurs missions sont explicitées aux producteurs de documents. Une fois sensibilisés, ces derniers sont souvent favorables à ce que les documents soient collectés et conservés dans les meilleures conditions.

En ce qui concerne la conservation, c'est généralement l'absence de structure d'accueil et de stockage ainsi que la précarité des moyens accordés aux structures archives qui posent problème. Ainsi, notamment pour l'archivage électronique, le besoin se heurte à l'importance des moyens financiers et humains à engager : aussi, la solution de mutualisation sur ce domaine apparaît la plus favorable à tout point de vue.

5 CINES : Qu'est-ce qu'on peut souhaiter aux archivistes universitaires pour 2013 ?

Charlotte Maday : Le réseau d'archivistes dans les établissements d'enseignement supérieur et recherche, appelé AURORE, est désormais bien structuré et s'est équipé d'outils fonctionnels élaborés à partir des expériences de collecte sur le terrain. Ces outils permettent aux professionnels arrivant en poste de débuter rapidement leur activité sans avoir à tout créer. Comptant actuellement plus de 110 membres, on ne peut qu'espérer son développement de manière à améliorer l'efficacité de l'action de ces archivistes.

LE STANDARD D'ECHANGE DE DONNÉES POUR L'ARCHIVAGE (SEDA)

Le standard d'échange de données pour l'archivage modélise les différentes transactions (transfert, demande de transfert, modification, élimination, communication, restitution) qui peuvent avoir lieu entre des acteurs (producteur, service versant, service d'archives, service de contrôle, demandeur d'archives) dans le cadre de l'archivage de données. Ce standard définit de manière formelle les scénarii dans lesquels les acteurs s'échangent des messages (diagrammes de séquences en UML) et la forme des messages échangés au cours de ces transactions (schémas XML). Les scénarii mis en œuvre dans PAC sont entièrement conformes au SEDA.

Initialement développé dans le cadre de l'action 103 du programme ADELE, ses évolutions sont actuellement réalisées par un comité de pilotage national (avec des représentants du SIAF, du réseau public des archives, des éditeurs, des consultants, des tiers archiveurs, et d'autres institutions telles que l'ADULLACT ou le CINES). Le RGI (Référentiel Général d'Interopérabilité) recommande l'utilisation du SEDA dans le processus d'archivage. Par ailleurs, suite à la publication de la version 1.0 en septembre 2012, il fait actuellement l'objet d'un processus de normalisation auprès de l'AFNOR.

Plus de renseignements sur :
http://www.archivesdefrance.culture.gouv.fr/gerer/archives-electroniques/standard/seda/

L'archivage des données scientifiques : des besoins spécifiques

Il existe un lien fort entre la puissance de calcul d'un établissement et la quantité de données qu'il consomme et produit. En proposant à la communauté scientifique française une puissance de 267 Tflop, le CINES doit assurer la gestion d'une quantité de données scientifiques considérable qui ne cesse de croître.

Conscient de l'importance de cette donnée, le CINES reste attaché à consolider son expertise sur les problématiques inhérentes au cycle de vie de cette donnée afin de pouvoir proposer à la communauté scientifique des outils permettant sa valorisation et sa conservation, au-delà d'un simple stockage.

Cette expertise se fonde sur la gestion des données scientifiques liée à sa mission principale et historique de calcul intensif, mais également à sa mission d'archivage, ou encore à son implication dans des projets européens à la problématique similaire, tels que EUDAT. Le CINES a lancé, en 2011, une enquête(41) auprès de 150 laboratoires français qui utilisent ses calculateurs portant sur la gestion de leurs données.

L'ESSENTIEL

- Le volume d'informations scientifiques ne cesse de croître et en l'absence d'actions spécifiques, une part conséquente risque d'être définitivement perdue.
- Pour la pérenniser, il est nécessaire de décrire la donnée au travers de métadonnées et de formats maitrisés et, si possible, standardisés.
- Dans le monde scientifique, il y a peu de standards de métadonnées et de formats de fichiers qui soient publiés, largement utilisés et standardisés.

Tous ces éléments nous permettent de vous proposer un panorama des données scientifiques en France et en Europe.

Les formats

Que ce soit au niveau de la consommation, de la production ou de l'exploitation, une donnée à caractère scientifique implique l'intervention de bibliothèques informatiques, de logiciels ou d'applications parfois « maison » qui en déterminent souvent le format. Une majorité de projets utilisant le calcul intensif ont des données au format binaire en sortie des logiciels. Les fichiers ASCII et textes sont utilisés dans un tiers des projets souvent en complément des données de calcul. Les données aux formats HDF5 et NetCDF sont aussi très présentes. D'autres formats sont utilisés plus rarement comme le FITS, Grib ou CGNS. Dans ce contexte, un problème fréquemment rencontré est celui de l'uniformisation des données.

Le partage des données au cours des travaux est souvent un besoin exprimé par les communautés. En effet, une majorité des projets a besoin d'échanger leurs données mais en priorité dans un cercle restreint de collaborateurs connus.

La volonté de partage avec l'ensemble de la communauté scientifique est plus rare ou alors, elle pourrait se faire dans un second temps, par exemple après une publication.

La donnée n'étant généralement pas exploitable d'un logiciel à un autre, il est nécessaire de la convertir systématiquement à l'aide de formats pivots (qui soient suffisamment spécifiés et génériques pour être compréhensibles et interopérables au sein d'une même communauté) tout en répondant aux contraintes métier de la communauté.

(41) Cf. encadré pages 44-45

PANORAMA DES FORMATS LES PLUS RÉPANDUS

- HDF5, NetCDF sont des formats ouverts, libres de droits et très généralistes. Ils sont auto-descriptifs dans la mesure où les données et les métadonnées sont contenues à l'intérieur même du fichier. Ils sont conçus pour contenir et manipuler des matrices (comme un maillage par exemple).

- FITS est un format ouvert, libre de droits, conçu pour les images scientifiques. Il permet une description poussée des composants de l'image car il encapsule dans ses en-têtes des métadonnées au format ASCII. Chaque bloc de données peut alors être décrit par un couple attribut/valeur. Un certain nombre d'attributs est réservé au format FITS. Si cela ne lui convient pas, l'utilisateur est libre de définir son propre paramétrage.

Les données scientifiques représentent généralement, avec une certaine complexité, des phénomènes très précis. Toute description a ses limites et si elle n'a pas été mûrement réfléchie, il peut apparaitre un risque de perte de connaissance.

Ainsi, certains centres de recherche en France et dans le monde ont récemment dû faire appel à leurs retraités pour comprendre d'anciennes données. On comprend dès lors qu'il est très important d'en faire une description au moins à deux niveaux pour limiter ce risque :

- 1er niveau : la description syntaxique qui permettra de connaître l'organisation des données dans le fichier (ex : les types primitifs, leur taille, une position dans un tableau, etc.). Ces informations font généralement partie de l'entête des fichiers et sont destinées principalement aux systèmes informatiques chargés de les exploiter.

- 2° niveau : la description sémantique qui va donner des informations sur la correspondance entre la donnée et le sens qu'on lui attribue. Telle valeur correspond à une température, une pression, etc. Ces métadonnées peuvent également être directement contenues dans le fichier ou décrites dans un autre.

Que peut-on faire alors pour garantir au maximum que la donnée sera exploitable par un tiers dans un contexte plus ou moins proche de nos travaux ?

Il ne s'agit pas d'imposer un format standard ou une démarche laborieuse de description des données à un laboratoire qui n'a pas les moyens de le mettre en œuvre mais de sensibiliser et d'amener les producteurs à une bonne gestion des risques et des conséquences liées à la perte d'exploitabilité de leurs données.

Il peut être en effet utile de se poser un certain nombre de questions concernant le cycle de vie de sa donnée :

- *Dans quel but a été produite cette donnée ?* (Partage avec une large communauté ou uniquement pour un travail spécifique du laboratoire ?)

- *Qui sont les destinataires de la donnée et ont-ils une base de connaissance et des outils suffisants pour exploiter cette donnée ?* (ex : les membres du laboratoire ou l'ensemble des scientifiques travaillant sur cette thématique seront-ils capables de comprendre ce fichier binaire ?)

- *Pendant combien de temps cette donnée est-elle pertinente, son niveau de sauvegarde est-il à la mesure de son coût de production ?*

Suivant les réponses, il apparaît utile, nécessaire, voire impératif, d'enclencher par exemple une migration du format maison sans description vers un format standard et si ce n'est pas suffisant, d'y associer un ensemble de métadonnées compréhensibles par les destinataires. Notons ici que notre enquête révèle que 40% des fichiers binaires ne posent pas, a priori, de problèmes de migration vers un format standard de type HDF5 ou NetCDF.

Remarquons cependant qu'un format binaire accompagné d'une bonne description de son contenu, restera facilement exploitable par un tiers et pérenne dans le temps. L'outil BEST (http://logiciels.cnes.fr/BEST/FR/best.htm) est une solution proposée par le CNES pour décrire les fichiers binaires que ce soit au niveau syntaxique avec le langage EAST ou au niveau sémantique avec le standard interne de la NASA (DEDSL) jouissant maintenant d'une réputation internationale.

Les métadonnées

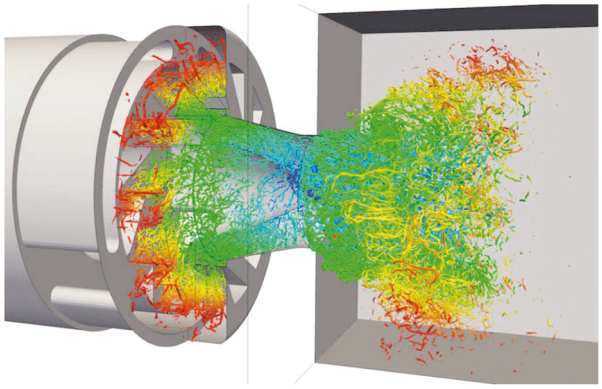
Une donnée est une représentation élémentaire d'une réalité. De fait, elle n'est pas auto descriptive et ne possède pas forcément un sens évident. L'objectif des métadonnées (méta : à côté, au-delà, qui englobe) est d'ajouter un niveau descriptif suffisamment pertinent pour permettre l'exploitation d'un fichier et son partage dans les meilleures conditions. Cependant, les communautés dans lesquelles les chercheurs se regroupent sont très différentes, les laboratoires sont très spécialisés dans leur domaine d'activités et la description des données n'est pas forcement une priorité.

La plupart des laboratoires ne possèdent pas de standard concernant les jeux de métadonnées. Ils décrivent les données principalement à l'aide de références vers des fichiers texte, des notes, des publications, des thèses, des pages web, des codes source, des paramètres de simulation, voire uniquement en utilisant un système de nommage mnémonique des fichiers.

La communication de l'information et le travail collaboratif sont incontournables dans le monde ouvert qui est actuellement le nôtre. Aussi, il est important que les laboratoires et les communautés scientifiques ayant une problématique commune soient fédérés autour de ces enjeux afin de faire émerger des moyens et des référentiels communs capables de garantir une lisibilité et une compréhension suffisante de la donnée qu'ils produisent.

LA DESCRIPTION DES DONNÉES

Voici un exemple de données scientifiques proposé par le CORIA montrant des calculs du brûleur semi-industriel PRECCINSTA conçu par la société Turbomeca (groupe SAFRAN). Ce brûleur a été mesuré expérimentalement au DLR en Allemagne et cette base de données sert aujourd'hui de validation pour les modèles de combustion turbulente. Un maillage de 2,6 milliards de tétraèdres avec une résolution de 100 microns (1 millième de la taille du brûleur) et 16 388 coeurs de la machine IBM Blue Gene/P de l'IDRIS ont été nécessaires pour réaliser ces simulations. Le logiciel YALES2 utilisé pour ces simulations est développé au CORIA depuis 2006 et il a été spécialement conçu pour ces simulations massivement parallèles. De nombreuses informations sont disponibles sur le site :
http://www.coria-cfd.fr



Auriez-vous deviné ce dont il s'agit simplement en regardant l'image ? Comprenez-vous l'ensemble de ces explications ?

Cet exemple montre l'importance d'une bonne expertise en ce qui concerne la qualification des informations de représentation d'une donnée à caractère scientifique.

L'objectif principal reste simple : décrire les données de manière à ce qu'elles soient exploitables par l'ensemble des personnes auxquelles elles sont destinées. Pour juger de l'importance de cela, imaginez les conséquences s'il n'y avait pas de langue de référence à savoir l'anglais pour décrire les articles scientifiques ! Pour y parvenir, des actions à plusieurs niveaux sont à mettre en place :

- Organisationnel : trouver les interlocuteurs, articuler les échanges selon les compétences.

- Informatique : mettre en place des infrastructures logicielles, matérielles et les protocoles permettant le traitement à proprement parler des métadonnées.

- Archivistique : trouver les normes et les formats d'échange pertinents dans le domaine pour permettre à l'information de sortir indemne des outrages du temps.

- Méthodologique : il est important d'identifier les destinataires de la donnée (communauté cible) et de mesurer leur capacité à comprendre cette donnée (base de connaissance). Ensuite il faut mettre en place un ensemble d'informations de représentation qui constituera un lien sémantique entre la donnée et la communauté.

Philippe PRAT

Chef du projet ISAAC

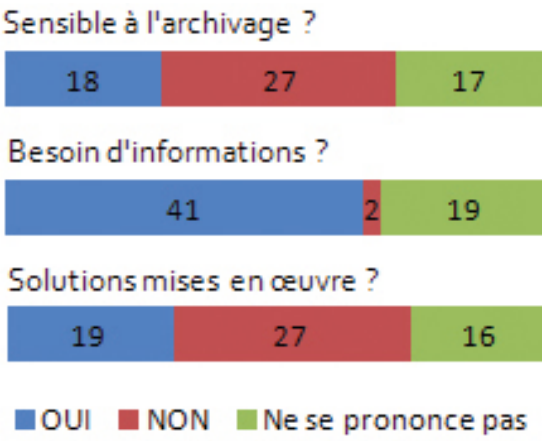
RÉSULTATS D'UNE ENQUÊTE SUR...

Afin de mieux cerner les besoins des utilisateurs dans le domaine de l'archivage des données scientifiques, le CINES a lancé une enquête en 2011 auprès des laboratoires de recherche dont voici les résultats :

- 155 laboratoires sollicités

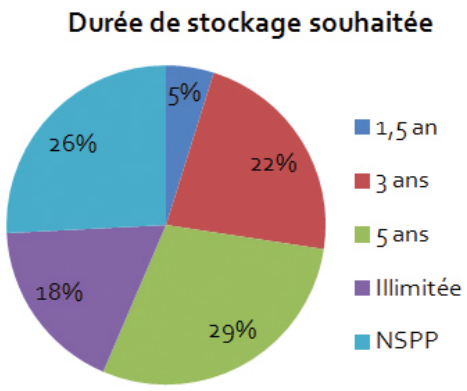
- 40 % de réponses

Sensibilisation à l'archivage et solutions mises en œuvre



Peu de laboratoires effectuent un archivage de leurs données. On note une inquiétude quant à la sécurisation des données archivées face au risque de perte et un besoin fort d'avoir des informations sur l'intérêt de l'archivage et de la diffusion.

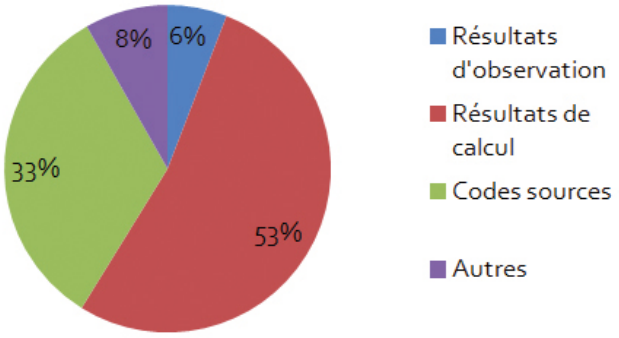
Besoins en termes de durée de stockage



LES BESOINS EN ARCHIVAGE DES DONNÉES SCIENTIFIQUES

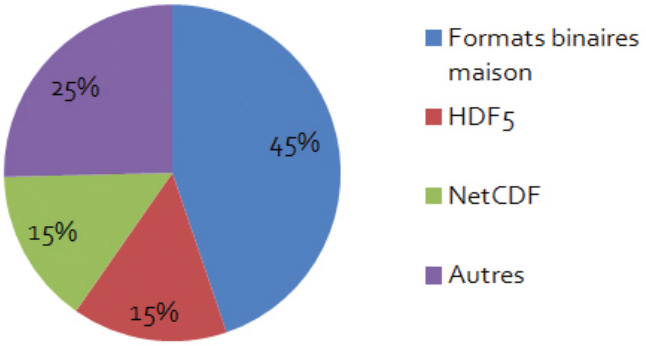
Une forte majorité des laboratoires a besoin d'un accès aux données pendant au moins 3 à 5 ans, par exemple pour avoir le temps d'analyser et de publier leurs résultats. La demande d'un stockage illimité est aussi assez forte. L'intérêt porté à l'archivage tourne autour de la reprise de travaux, la comparaison de résultats, l'amélioration de la sécurisation des données, la possibilité de faire du retraitement des données sur du plus long terme, le fait de ne pas refaire des calculs longs et coûteux et le partage des données entre scientifiques.

Types de données à archiver



La plupart des laboratoires ont besoin d'archiver leurs résultats de calcul, contrairement aux codes sources qui représentent une volumétrie moindre et qui peut donc être gérée directement par le laboratoire. Le faible nombre de demandes pour l'archivage des résultats d'observation est à relativiser avec le nombre de projets réalisant ce type de mesures au CINES. Les autres types de données concernent : les données fournies pour calculer, les résultats d'analyse, les visualisations et les sources des publications (simulations publiées).

Formats utilisés



Pour la majorité des projets, les logiciels ou les codes de calcul génèrent des données au format binaire. Les fichiers ASCII et textes sont utilisés dans un tiers des projets, souvent en complément des données de calcul. HDF5 et NetCDF semblent être plus utilisés que des formats tels que FITS, Grib, CGNS.

Les métadonnées descriptives

Les métadonnées ne sont généralement pas ou très peu utilisées dans les projets. Les données peuvent être accompagnées de fichiers explicatifs de leur contenu. On trouve aussi des métadonnées embarquées dans les entêtes des fichiers de données. L'utilisation de jeux de métadonnées standards est inexistante et cette notion est même inconnue.

La diffusion et le partage des données

Une majorité des projets a besoin d'échanger leurs données, en priorité dans un cercle restreint de collaborateurs connus. La volonté de partager avec l'ensemble de la communauté scientifique est plus rare ou alors elle serait souhaitable dans un second temps, par exemple après une publication. La gestion des droits d'accès est importante, avec en majorité un accès en lecture uniquement ou lecture/écriture selon la personne ou le groupe de personnes.

L'archivage des données scientifiques : ISAAC, une solution du CINES

Un constat

La position du CINES comme l'un des plus grands centres de calcul intensif français et européen lui permet d'être au cœur des préoccupations de la communauté scientifique quant au devenir des données scientifiques. Pour ce type de données, qu'elles soient le résultat de simulations ou le résultat d'observations scientifiques, plusieurs facteurs justifient de s'intéresser de près à leur conservation. Les besoins de réutilisation des résultats (à des fins de comparaison par exemple), les besoins d'échanges de données entre les équipes scientifiques ou encore l'importance des coûts de production font de l'archivage électronique des données scientifiques un enjeu majeur pour les acteurs de la recherche.

L'ESSENTIEL

- Conserver les données scientifiques : un enjeu majeur pour la communauté
- Identifier les besoins des communautés de chercheurs : pour proposer un service utile.
- Fournir un cadre de travail structuré administrativement.
- Une conservation à moyen terme (3 à 5 ans)
- Proposer un outil simple d'utilisation pour avoir l'adhésion des utilisateurs.
- ISAAC : un service du CINES ouvert courant 2013.

Un projet global

À partir de ce constat, et en s'appuyant sur son expertise dans le domaine de l'archivage pérenne (projet PAC), le CINES a initié le projet ISAAC (l'Information Scientifique Archivée Au CINES). Ce projet a pour ambition d'offrir à la communauté de l'enseignement supérieur et de la recherche un service d'archivage électronique spécifique aux données scientifiques. La durée de conservation sera de 3 à 5 ans (conservation à moyen terme). ISAAC est un projet global. Il comprend en effet toutes les étapes de la mise en œuvre d'un projet informatique classique (de la définition des spécifications techniques et fonctionnelles à leur implémentation sur une plateforme de production), mais il a aussi pour but la mise en place d'une organisation administrative structurée faisant intervenir différents acteurs (chercheurs, experts des domaines scientifiques, représentants du CINES).

Une réponse liée aux besoins de la communauté

Cerner les besoins des utilisateurs est essentiel même si cela peut s'avérer complexe. L'étude du projet Européen EUDAT(42) a permis d'identifier un certain nombre de fonctionnalités minimales que devrait apporter un service d'archivage de données scientifiques. De manière complémentaire, le CINES a réalisé en 2011 une enquête auprès de 150 laboratoires de recherche. Ce travail a permis de poser les bases d'une offre de service qui soit à la fois conforme aux directives européennes, en accord avec les techniques d'archivage, mais aussi et surtout qui réponde au mieux aux besoins des utilisateurs.

(42) Cf. fiche de présentation du projet EUDAT, page 55.

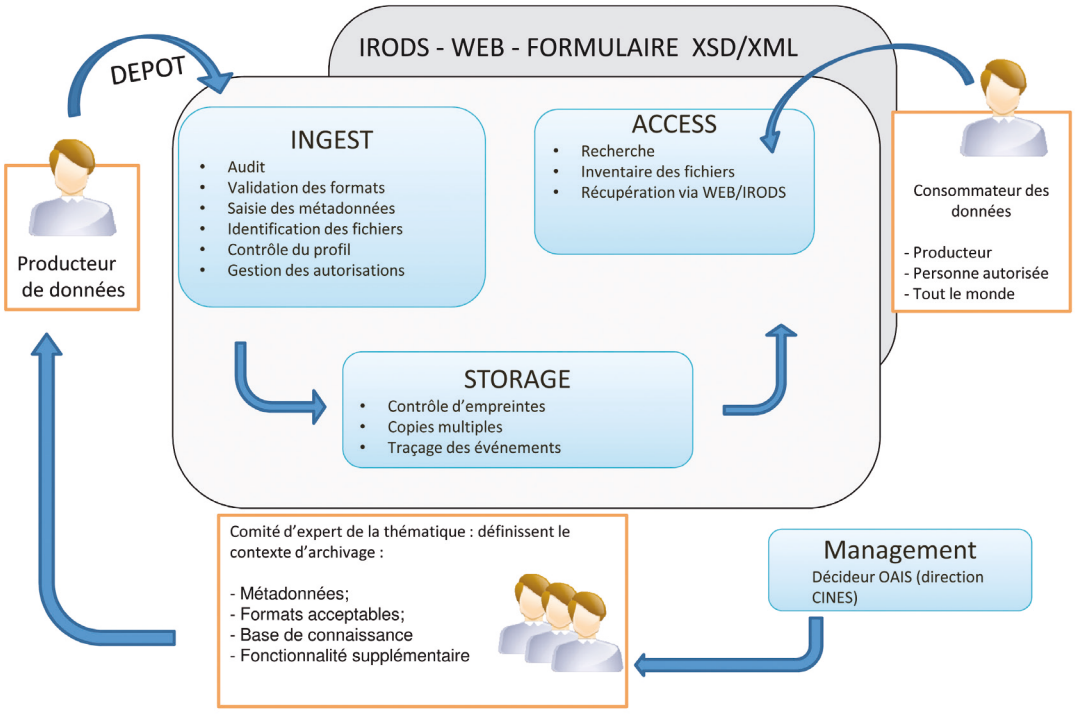
Les principaux besoins identifiés sont :

- La conservation des données à moyen terme (3 à 5 ans)
- La description des données grâce à des métadonnées
- Le partage des données au sein d'un groupe d'utilisateurs
- Le contrôle d'authenticité (login, mot de passe)
- Une gestion des permissions d'accès (groupe, lecture, écriture, etc.)
- La conservation de l'historique des évènements
- La création d'un identifiant unique sur tous les objets numériques
- Une interface Web pour le dépôt, l'accès et le partage des archives
- Une implication forte des utilisateurs notamment pour les demandes d'amélioration

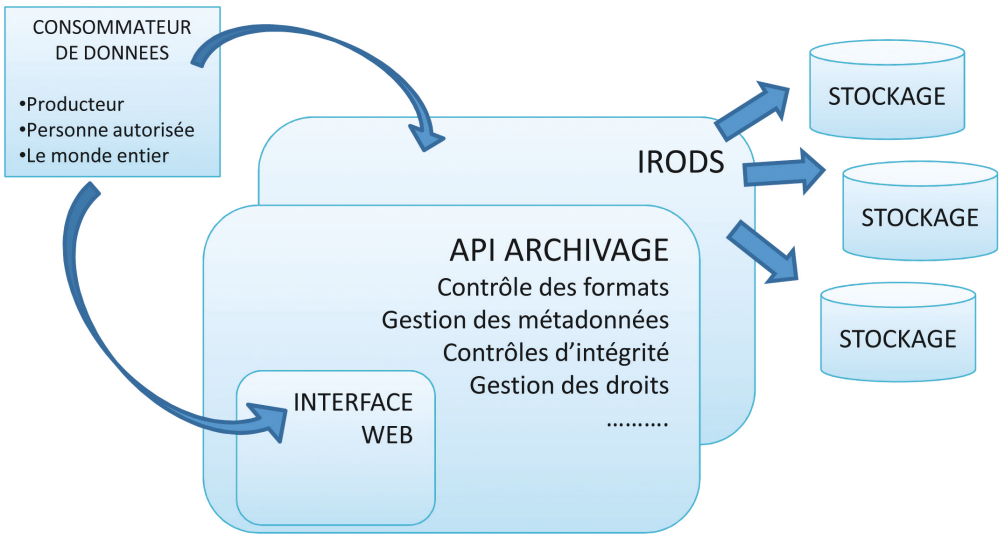
Une organisation structurée d'envergure nationale

L'un des principaux défis du projet ISAAC réside dans son aspect organisationnel et administratif. En effet, toute donnée produite doit être rattachée à un contexte structuré et reconnu, ce qui permet d'en garantir sa pertinence et son intégrité. Ainsi, et dans le même esprit que les comités thématiques scientifiques du CINES du calcul intensif, le projet ISAAC propose la création de « Comités Thématiques d'Archivage ». Chaque CTA est composé d'un président, d'un représentant de la plateforme d'archivage, d'un ou plusieurs experts dans le domaine scientifique concerné. Les rôles essentiels du CTA sont :

- L'étude et le choix des formats de fichiers acceptés dans le système d'archivage
- L'étude et le choix des jeux de métadonnées à utiliser pour la description des données
- L'étude et l'acceptation des projets d'archivage



ISAAC, FONCTIONNALITES



ISAAC, ARCHITECTURE TECHNIQUE

Une architecture technique orientée «données réparties» avec iRods

Ce projet portant sur de grandes quantités de données ayant pour vocation à être partagées sur différents sites avec probablement des environnements hétérogènes, il nous parait opportun de se positionner sur une infrastructure technique offrant la possibilité d'une organisation de type grille de données. Le choix de la technologie « open source » iRods a été fait en ce sens car elle offre des fonctionnalités très performantes en particulier en ce qui concerne la gestion de gros volumes de données réparties. Une interface Web conviviale et simple à utiliser est proposée pour :

- Le dépôt des jeux de données à archiver
- La description des données : compte tenu de la configuration des laboratoires qui n'ont pas toujours les moyens techniques et humains pour générer eux-mêmes leurs fichiers de description XML, ISAAC encapsule un système de génération automatique de formulaires à partir de schémas XML. Ce module permet de faire de manière flexible des formulaires de saisie web pour décrire les données, les projets, les informations spécifiques à une thématique.
- L'analyse et le contrôle des jeux de données selon les critères du CTA
- La recherche, la consultation et l'échange des archives (accès sécurisé)

Un projet qui avance

Actuellement, le projet ISAAC est en phase de développement d'une première plateforme de tests qui permettra de valider tous les choix techniques et organisationnels. Ces tests se feront sur des jeux de données de simulations fournis par le CORIA (COmplexe de Recherche Interprofessionnel en Aérothermochimie).

L'année 2013 verra donc la mise en œuvre d'une solution concrète pour la conservation des données scientifiques au CINES… À suivre !

Jean-Pierre THERON

Ingénieur - Équipe ISAAC

Les collaborations du CINES aux niveaux national et européen

Fort d'une mission nationale confiée par le Ministère de l'Enseignement Supérieur et de la Recherche, le CINES est appelé à jouer un rôle clé dans la réussite d'une stratégie pour l'archivage pérenne des documents électroniques produits par la communauté Université-Recherche. Il reçoit aujourd'hui de nombreuses sollicitations émanant d'organismes publics, d'universités, de bibliothèques ou de laboratoires, pour des services d'aide et de conseil, des retours d'expérience ou encore des collaborations au sein de groupes de travail ou de projets financés par la Commission européenne.

Le projet de Bibliothèque Scientifique Numérique

BSN BIBLIOTHÈQUE SCIENTIFIQUE NUMÉRIQUE DIGITAL SCIENTIFIC LIBRARY

La Bibliothèque Scientifique Numérique – BSN – est un projet initié par le Ministère de l'Enseignement Supérieur et de la Recherche en 2008 et inscrit dans la feuille de route des infrastructures de recherche, qui vise deux cibles principales :

- répondre aux besoins des chercheurs et enseignants-chercheurs en portant l'offre qui leur est fournie en information scientifique et technique – IST – à un niveau d'excellence mondiale ;
- améliorer la visibilité de la recherche française.

L'importance de la BSN, la transversalité des problématiques abordées et l'existence d'un comité de pilotage ont entrainé la création d'un groupe de travail spécifique. Composé des grands décideurs de l'IST et de membres du comité de pilotage, il a pour mission de proposer des pistes d'organisation et de gouvernance, des indicateurs d'activité et des critères d'évaluation de l'activité.

L'ESSENTIEL

La Bibliothèque Scientifique Numérique est un projet porté par le Ministère de l'Enseignement Supérieur et de la Recherche depuis 2008. Elle réunit les principaux acteurs nationaux des neuf segments d'activité qui la composent en vue d'améliorer l'accès à l'information scientifique et technique numérique. Dans ce cadre, le CINES anime le groupe de travail consacré à la préservation numérique à long terme, qui doit constituer un des lieux visibles des coopérations et mutualisations requises par la politique nationale portée au niveau interministériel.

L'association Aristote, francophone, sans but lucratif et indépendante des fournisseurs de solution, a pour objet de regrouper les utilisateurs des nouvelles technologies de l'information et de communication, les organismes d'enseignement supérieur, de recherche et développement, et d'innovation. Parmi les groupes de réflexion qui la composent, le CINES est l'animateur principal de celui pour la Pérennisation de l'Information Numérique, lieu de rencontres entre informaticiens, archivistes et bibliothécaires pour la formation et le partage de retours d'expérience, savoir-faire, succès et échecs dans le domaine de la préservation numérique à long terme.

Le Data Seal of Approval est une initiative menée par quelques institutions européennes pour garantir l'accessibilité et la réutilisabilité des données numériques dans le futur. Elle est basée sur un ensemble de bonnes pratiques sur la qualité apportée à la conservation de l'information, laquelle est évaluée pour l'obtention d'une accréditation. Le CINES y apporte son expertise pour l'évaluation des institutions candidates.

Ses objectifs portent à la fois sur la mise en place d'un pilotage politique partagé par les grands acteurs de l'enseignement supérieur et de la recherche, offrant des services à l'ensemble des communautés (quel que soit leur statut), et sur la création de nouveaux modèles et équilibres économiques entre acteurs publics et privés dans l'édition scientifique.

Ce groupe de travail est décomposé en neuf segments d'activité (acquisitions, diffusion, signalement, archives ouvertes, numérisation, archivage pérenne, formation, etc.) pour lesquels des actions ont été initiées par différents opérateurs nationaux.

Dans ce cadre, le CINES anime le segment consacré à la préservation numérique à long terme, qui doit constituer un des lieux visibles des coopérations et mutualisations requises par la politique nationale portée au niveau interministériel. Ainsi, plusieurs points seront étudiés parmi lesquels :

- la typologie de données à conserver sur le très long terme ;
- les moyens de rationalisation des pratiques des opérateurs existants et ainsi de réduction des coûts ;
- la prise en compte de l'archivage pérenne dans les contrats des établissements et son financement dans la durée ;
- les conditions de l'obtention de l'agrément des Archives de France ;
- ou encore l'organisation et l'assise juridique de l'ensemble.

Ce groupe de travail a repris pour l'essentiel les préconisations faites dans le cadre du Schéma Numérique des Bibliothèques[43] dans lequel le CINES était impliqué (volet conservation numérique). Elles ont servi de base à l'élaboration de la stratégie de la BSN, et des actions à entreprendre en 2013.

Aristote et le groupe de travail « Pérennisation de l'Information Numérique » – PIN

Aristote est une société savante et philotechnique dans le domaine de l'informatique et des réseaux. Elle regroupe des organismes et des entreprises impliqués dans les derniers développements et nouveaux usages des technologies de l'information ; elle organise également des séminaires et des formations.

Les groupes de travail de l'association sont des lieux d'échanges privilégiés autour des nouvelles technologies de l'Information et de la Communication – NTICs.

Ainsi, le groupe PIN est un lieu de rencontres entre informaticiens, archivistes et bibliothécaires pour le partage de retours d'expérience, savoir-faire, succès et échecs dans la pérennisation de l'information numérique. Il essaie de contribuer à la définition et au développement de produits répondant aux besoins. L'animation principale est effectuée par le CINES, avec l'aide d'un comité d'animation qui regroupe des représentants de cinq organismes moteurs (BnF, CEA, CINES, CNES, SIAF).

Le site web de PIN rassemble les compte-rendus et les interventions des réunions plénières organisées par le groupe depuis près de dix ans sur les thèmes et les problématiques de la préservation du numérique. Il est assorti d'une liste de diffusion qui rassemble environ cent cinquante inscrits.

Chaque année, une session de formation d'une semaine est organisée en partenariat avec de grandes institutions telles que l'Institut National du Patrimoine ou encore l'Ecole Nationale des Chartes, dans laquelle des représentants du CINES interviennent régulièrement. Par ailleurs, des réflexions sont en cours pour l'élaboration d'un cursus de formation de longue durée.

Les collaborations dans le domaine du logiciel libre

Fort du succès rencontré avec l'outil FACILE, le CINES a ouvert le code de cette application pour permettre à d'autres établissements ou organismes intéressés de l'intégrer dans leurs applications métier. Dans cette perspective, il collabore avec l'ADULLACT, dont l'objectif est de promouvoir, développer, mutualiser et maintenir un patrimoine commun de logiciels libres utiles aux missions de service public, notamment au travers de l'offre as@lae, son système d'archivage électronique à valeur probante.

De la même manière, des travaux et des études autour des formats de fichiers sont régulièrement menés en partenariat avec les acteurs du domaine, tels que le SIAF ou encore le Très Grand Equipement – TGE – Adonis, et font l'objet de publications sous forme de guides méthodologiques ou de codes à intégrer dans les outils de validation, tels que :

- un guide pour le choix de formats numériques pérennes dans un contexte de données orales et visuelles publié en 2010 (44) ;
- un prototype de validation de formats de fichiers pivots pour l'archivage de bases de données relationnelles réalisé en 2011 (44) ;
- une étude sur le format de fichiers PDF et les outils de conversion ou de contrôle associés, menée en 2012 (44).

Les initiatives au niveau européen

Au niveau européen, le CINES est impliqué depuis 2007 dans les travaux de l'Alliance pour la préservation de l'information scientifique(45).

Cette initiative a pour objectif la coordination de la mise en place d'une infrastructure européenne pour l'archivage pérenne des données scientifiques, et est à l'origine de plusieurs projets financés par la Commission européenne, parmi lesquels APARSEN, évoqué dans l'encadré sur l'ISO 16363, page 35.

À cette même date, le CINES a été référencé comme centre de compétences sur l'archivage numérique par l'initiative Digital Preservation Europe – DPE(46) – mettant ainsi en avant ses activités et son expertise dans le domaine.

Peu après, le CINES a rejoint le comité directeur du Data Seal of Approval – DSA. Il collabore avec d'autres acteurs européens concernés par la durabilité de l'information scientifique à la mise en place d'une accréditation simple des centres de préservation numérique. Ce sceau est attribué aux institutions ayant instauré des procédures d'assurance qualité pour garantir l'accessibilité et la lisibilité de l'information qui leur est confiée. Une telle accréditation s'obtient après une demande en ligne(47), une autoévaluation selon seize critères et une revue finale par un des membres du comité directeur. Cette initiative a été reprise par la Commission européenne dans son projet de mise en place d'un cadre pour l'audit et la certification de systèmes d'archivage électronique, dont elle constitue le premier niveau.

Olivier ROUCHON
Responsable du Département Archivage et Diffusion
CINES

(43) B. Racine, Président de la Bibliothèque nationale de France, rapport du Schéma numérique des bibliothèques élaboré dans le cadre du Conseil du Livre (2009) : http://www.ddm.gouv.fr/IMG/pdf/SNB_Rapport_Racine.pdf
(44) Toutes ces études sont disponibles sur : http://www.archivesdefrance.culture.gouv.fr/gerer/archives-electroniques/formats-et-supports
(45) Site web de l'Alliance for Permanent Access : http://www.alliancepermanentaccess.org/
(45) http://www.digitalpreservationeurope.eu/
(47) Site web du Data Seal of Approval : http://www.datasealofapproval.org/

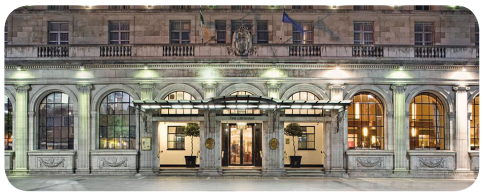
LE CINES DANS LES GRANDES CONFÉRENCES INTERNATIONALES…

CONFÉRENCE ICA 2012
août 2012, Brisbane

Les travaux du CINES sont régulièrement présentés dans les conférences internationales du domaine comme le PASIG (Preservation and Archiving Special Interest Group) ou encore l'iPRES (International Conference on Preservation of Digital Objects). Ainsi, les services et expertises du CINES en matière de préservation numérique ont été présentés en novembre dernier à Vienne (Autriche) lors de la conférence DARIAH, le projet de réplication croisée de données avec la BnF au PASIG 2010 à Madrid (Espagne), le travail de formalisation des processus métiers du service d'archivage lors de l'iPRES 2010 à Vienne (Autriche), et l'initiative de certification de PAC lors de l'iPRES 2011 à Singapour et de l'ICA 2012 à Brisbane (Australie). Tous les actes de ces conférences sont disponibles en ligne.

À cette même date, le CINES a été référencé comme centre de compétences sur l'archivage numérique par l'initiative Digital Preservation Europe – DPE – mettant ainsi en avant ses activités et son expertise dans le domaine. Peu après, le CINES a rejoint le comité directeur du Data Seal of Approval – DSA.

Il collabore avec d'autres acteurs européens concernés par la durabilité de l'information scientifique au travers d'associations ou de groupes de travail tels que l'Alliance for Permanent Access (APA), l'Association des Archivistes Français, le groupe PIN (association Aristote), le DPHEP (groupe international d'études sur « Data Preservation and Long Term Analysis in High Energy Physics »)…



CONFÉRENCE PASIG - octobre 2012, Dublin

FORUM DES UTILISATEURS D'EUDAT
mars 2012, Barcelone

CONFÉRENCE EUDAT - octobre 2012, Barcelone

DPHEP : WORKSHOP
novembre 2012, Marseille

CONFÉRENCE DE IPRES
novembre 2011, Singapour

L'archivage à l'heure de l'Europe

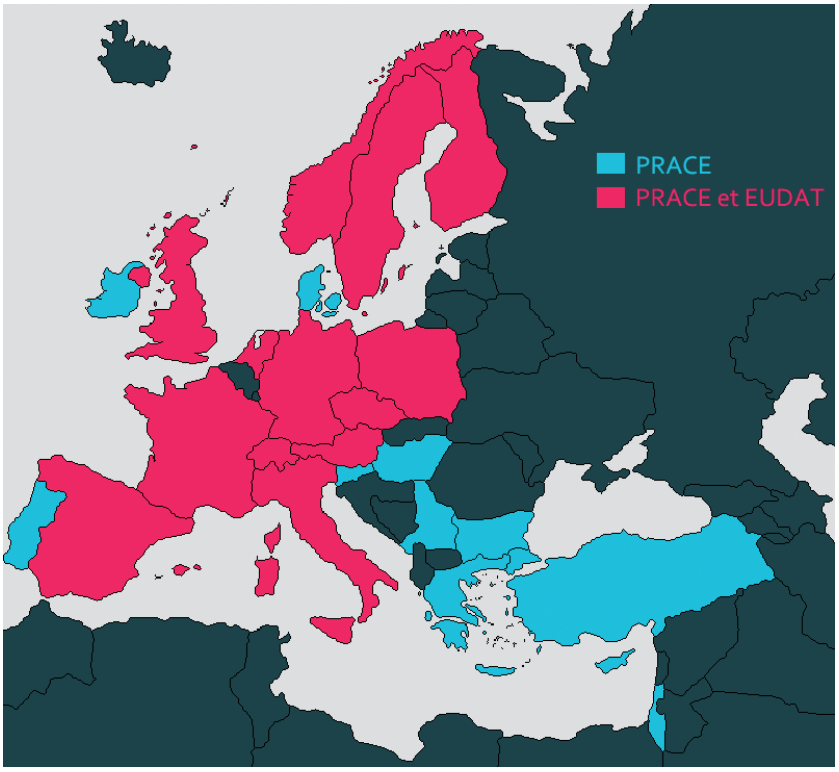
Le CINES est un centre national disposant d'une expertise dans le domaine du calcul intensif (HPC) et de la préservation à long terme de données numériques. Il participe donc à des programmes européens dans ces domaines d'activité.

Les derniers projets auxquels il a contribué sont : PRACE-1IP, PRACE-2IP, PRACE-3IP (projets de 2 ans chacun) et EUDAT (projet de 3 ans).

PRACE (Partnership for Advanced Computing in Europe) est un projet européen ayant pour objectif de faciliter la recherche et le développement en Europe dans tous les domaines scientifiques. Cet objectif est réalisé grâce à la mise en place de grilles de calcul à échelle européenne. Avec 21 pays membres dont les organisations représentatives travaillent ensemble pour former cette infrastructure de calcul intensif à grande échelle (distribuée) permettant aux scientifiques d'utiliser des applications nécessitant d'importantes ressources de calcul. Le projet est enrichi par des connaissances provenant de toute l'Europe dans le but d'exploiter au mieux ces ressources.

Les centres de calcul sont décrits en fonction de leurs utilisateurs. Les centres Tier-2 sont des centres régionaux, les centres Tier-1 sont nationaux et les centres Tier-0 sont européens. Grâce aux machines JADE (Tier-1) et CURIE (Tier-0), le CINES et le CEA sont ainsi des nœuds de cette grille de calcul européenne. Le CINES a fait bénéficier le projet PRACE-1IP de son expertise en matière d'infrastructures de stockage et d'archivage en rédigeant un livre blanc sur les systèmes existants pour préserver les données scientifiques.

La préservation à long terme des données présentée dans PRACE-1IP n'est plus présente dans PRACE-2IP : ces problématiques sont reprises dans le projet EUDAT. En effet, parallèlement à cela, les communautés scientifiques européennes ont exprimé auprès des centres de calcul le besoin de préserver et de partager les données qu'elles produisent chaque année. L'objectif du projet EUDAT est de mettre en place une grille européenne d'archivage de ces données scientifiques. La problématique est donc multiple : il s'agit à la fois de documenter, stocker, partager et répliquer des données de taille importante.



FICHE DESCRIPTIVE

Projet européen PRACE 1IP

- **Objectif :** réalisation d'une grille de calcul + état de l'art sur la conservation de données
- **Montant global du projet :** 400 M€ (investissements via le projet PRACE-RI) + 28 M€ pour PRACE 1IP (développement / fonctionnement)
- **Représentation CINES / France / Total Europe : 25 PM / 163 PM / 2305 PM.** 21 partenaires de 20 pays. Partenaire français : GENCI avec l'appui des 3 centres de calcul nationaux : CCRT/TGCC, IDRIS et CINES.
- **Partie archivage :** 5 rapports remis
- **Date de début / date de fin :** 01/07/2010 – 31/06/2012 (2 ans)

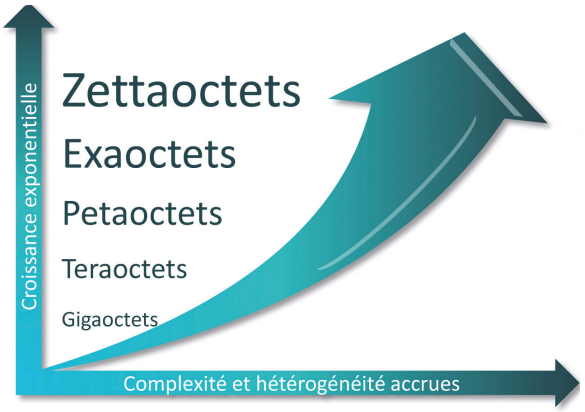
Les centres de calcul européens qui participent aux projets PRACE-1IP et PRACE-2IP ne disposent pas nécessairement de la technologie et de l'organisation appropriées pour préserver à long terme le patrimoine scientifique. Le livre blanc livré par le CINES dans le cadre du projet PRACE-1IP présente les bonnes pratiques, les équipements et les technologies pour archiver à long terme ces informations.

Une étude sur les principaux supports de stockage utilisés par les centres de calcul a été menée. Elle traite des supports disques et des bandes magnétiques. Il apparait que différents éléments environnementaux interviennent dans le processus de dégradation de ces supports. Mais que dans un environnement maitrisé, les bandes constituent un meilleur support de stockage que les disques pour la conservation des données sur une durée de 10 ans. En effet, des études menées par des laboratoires tels que le LNE (Laboratoire National de métrologie et d'Essais) montrent un fort taux de défaillance des disques à partir de 5 ans.

> *« La mise en place d'une grille de calcul européenne requiert une bonne connaissance des supports de stockage de l'information numérique. »*

Compte-tenu de la fiabilité et du coût de ces supports, les supports bandes sont préconisés pour l'archivage d'importantes volumétries de données. Toutefois, les disques, qui offrent d'excellentes performances en termes d'accès (I/O), sont souvent utilisés dans le cadre de la diffusion des données archivées. De plus, comme le mentionne l'article sur la démarche qualité dans un contexte de préservation à long terme du présent numéro de la gazette du CINES, il est préférable d'utiliser des technologies de support de stockage différentes pour éviter d'hypothétiques erreurs de conception ou structurelles qui pourraient apparaître au fil du temps.

Le rapport préconise également l'utilisation de métadonnées, de contrôle de formats de fichiers (afin de faciliter les futures opérations de conversion de formats), de moyens fiables de réplication de données, de systèmes de contrôle périodique de non corruption silencieuse (checksum) et de mise en œuvre de processus de veille technologique.



PRÉSERVER LES BIG DATA : UN NOUVEAU DÉFI

FICHE DESCRIPTIVE

Projet européen EUDAT

- **Objectif :** réalisation d'une grille de partage, diffusion et préservation de données
- **Montant global du projet :** 16.3 M€ (partie du FP7 e-Infrastructure Call 9 (WP11) : INFRA-2011-1.2.2 dont le budget est de 43 M€)
- **Représentation CINES :** 57 MH / 1 436 MH (soit 4% = 1/25°).
  25 partenaires de 13 pays.
  1 seul centre de calcul partenaire du projet en France : le CINES.
- **Date de début / date de fin :**
  01/10/2011 – 30/09/2014 (3 ans)

Ces dernières années, conscients qu'un véritable « déluge de données numériques » va s'abattre sur toutes les communautés de chercheurs, d'importants investissements ont été réalisés par la Commission Européenne et les états membres européens pour créer une infrastructure d'archivage commune.

Une e-infrastructure européenne de préservation de données numériques commence actuellement à prendre forme au travers des réseaux de communication, des systèmes distribués et des installations de calcul intensif (HPC) proposés aux chercheurs européens de toutes les disciplines : c'est le projet EUDAT. En effet, l'objectif de ce dernier est ambitieux puisqu'il s'agit d'une part de définir les bonnes pratiques en matière de conservation à long terme de données scientifiques et, d'autre part, de mettre en place une infrastructure distribuée répondant aux besoins exprimés par les communautés de chercheurs.

Le CSC (Centre informatique Finlandais pour la science) coordonne les actions menées par les 25 partenaires européens du projet. Il s'agit de centres de calculs, de communautés de chercheurs et d'organismes de financements implantés dans 13 pays. On notera que ce projet entre dans le cadre d'un programme européen de financement de l'union européenne : le « framework programme 7 ».

> *« Dans toutes les disciplines de la recherche internationale, on observe depuis plusieurs années une augmentation exponentielle du volume des données numériques produites. En conséquence, il en est de même des archives électroniques intermédiaires et définitives. »*

Le caractère unique du CINES réside dans le fait qu'il est le seul partenaire à disposer d'une « vraie » expérience en matière d'archivage à long terme de données. Mais sa contribution ne se limite pas à un partage de ses savoirs et savoir-faire avec les autres centres de calcul. En effet, le CINES projette d'être un nœud de la grille d'archivage européenne courant 2013.

En conséquence, le projet européen EUDAT mobilise divers types de compétences au sein du CINES. En abordant la question de la préservation à long terme de données sur un système distribué, les ressources humaines mises à disposition doivent maitriser à la fois les problématiques archivistiques (principes généraux de la préservation à long terme, spécificités des données numériques, cadre légal...) et celles concernant les infrastructures et les échanges de données (iRODS, administration de systèmes et réseaux, sécurité, monitoring, assistance aux utilisateurs, systèmes pour le stockage de gros volumes de données, gestion des accès...).

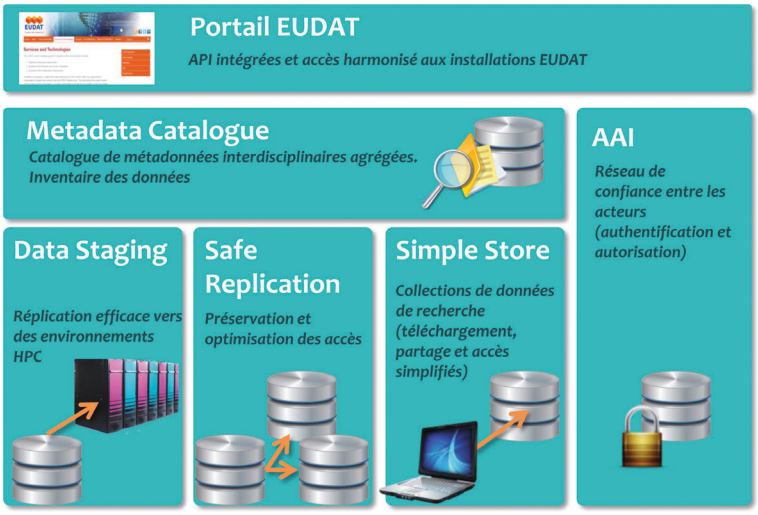
La définition de l'architecture européenne collaborative de données requiert une analyse préalable des besoins et des systèmes existants au sein des communautés de recherche. En 2012 (première année du projet), ce travail a été mené de manière particulièrement intensive avec cinq communautés pilotes : CLARIN (linguistique), ENES (Climatologie), EPOS (sismologie), LifeWatch (Ecologie) et VPH (Physiologie humaine).

Après des études et des tests sur les technologies disponibles sur le marché, les centres de calcul déterminent celles qui répondent aux besoins exprimés par l'ensemble des communautés. Ils identifient également les fonctionnalités nécessitant un développement informatique spécifique. L'implémentation des différents modules requis par la grille est réalisée par les centres de calcul afin de fournir rapidement aux partenaires scientifiques les services de partage, diffusion et préservation de données.

Fin 2012, les premiers services offerts par plusieurs nœuds de la grille s'apprêtaient déjà à entrer en phase de production.

L'architecture collaborative de données est composée de cinq modules principaux accessibles au travers d'un unique portail. Le « data staging » est le module de transfert de données depuis la grille EUDAT vers un système spécifique comme par exemple un environnement de calcul intensif (HPC). Le module « Safe replication » duplique les données sur plusieurs nœuds de la grille EUDAT. Le « Simple store » est un module offrant des fonctionnalités simplifiées pour le téléchargement, le partage, la recherche et l'accès aux données. Un catalogue de métadonnées interdisciplinaires agrégées (« metadata catalogue ») ainsi qu'un système d'authentification et d'autorisation (« AAI ») sont les deux autres modules de l'architecture de la grille EUDAT.

LES ÉCHANGES ENTRE LES NOMBREUX PARTENAIRES



LES PRINCIPAUX MODULES DE L'INFRASTRUCTURE COLLABORATIVE DE DONNÉES D'EUDAT

Marion MASSOL

Chef des projets PAC et EUDAT@CINES
CINES

Comment archiver au CINES ?

Le CINES, grâce à sa plateforme d'archivage pérenne PAC, répond à une demande importante de conservation de documents numériques sur une très longue période, 30 ans et plus. Les documents qu'il conserve sont en lien avec le monde de l'enseignement supérieur et de la recherche.

Pour archiver au CINES, il faut tout d'abord avoir une stratégie pour un projet de préservation à très long terme, à ne pas confondre avec une sauvegarde ou un stockage sécurisé de données.

L'organisme souhaitant verser ses documents effectue une demande par courrier. Cette demande est examinée par la direction du CINES qui étudie la pertinence et la faisabilité du projet. Après acceptation, une convention est signée et l'organisme devient un « service versant » du CINES.

Les données à archiver doivent satisfaire à certaines conditions :

- Les documents ne doivent plus avoir d'utilité courante.

- Il doit y avoir une sélection intelligente des données à archiver : priorité aux données « brutes » (il n'est pas nécessaire d'archiver des documents qui peuvent être reconstitués à partir de données déjà présentes dans le système d'archivage).

- Pour assurer leur pérennité, les données doivent être dans un des formats de données pris en charge par la plateforme d'archivage. Ceci facilitera les futures migrations des documents.

- Les documents doivent être accompagnés de métadonnées afin de pouvoir compléter le bordereau de versement, dont le schéma de référence est disponible à l'adresse suivante : http://www.cines.fr/pac/sip.xsd.

- L'organisme doit connaître et vérifier le contexte légal de production de l'archive : droits de propriété intellectuelle sur les documents à archiver, délais de communicabilité, etc.

DÉROULEMENT D'UN PROJET D'ARCHIVES AU CINES

Mireille Gay

Ingénieur – équipe PAC CINES

Glossaire

AIP
Archival Information Package - Paquet d'informations archivé. Paquet d'informations conservé dans un OAIS (système ouvert d'archivage d'informations) et constitué d'un contenu d'information et de l'information de pérennisation (PDI) associée.

Archivage électronique / numérique
Ensemble des actions visant à identifier, recueillir, classer, conserver, communiquer et restituer des documents électroniques, pour la durée nécessaire à la satisfaction des obligations légales, pour des besoins d'informations ou à des fins patrimoniales.

Archives électroniques
Paquet d'informations sous forme d'enregistrements électroniques reçu, conservé et communiqué par un service d'archives dans une disposition telle qu'elle en permette l'exploitation directe par des outils informatiques. Il est composé de données et de métadonnées :
- un contenu de données, c'est-à-dire le ou les objets numériques qui sont l'objet principal de la pérennisation ;
- des informations de représentation qui traduisent le contenu des données en des concepts plus explicites (par exemple la définition du code ASCII décrit comment une séquence de bits est convertie en caractères).
- des informations de pérennisation, c'est-à-dire des informations nécessaires à la bonne conservation du contenu d'information (provenance, identification, intégrité, contexte...).

Le contenu de données et les informations de représentation correspondantes forment le contenu d'information.

Base de connaissances
Ensemble d'informations préalablement acquises par une personne ou un système lui permettant de comprendre l'information reçue.

Collection
Regroupement volontaire de documents, d'objets, d'informations de provenances diverses, rassemblés en raison de la similitude d'un ou de plusieurs de leurs caractères (contrairement au fonds d'archives qui est constitué organiquement). Exemple : collection d'affiches, de journaux, d'ouvrages.

Communicabilité
Qualité d'un document ou d'un fonds d'archives que son régime juridique rend accessible au public, conformément au Code du Patrimoine.

Cycle de vie
Succession des différentes phases de traitements appliqués à un document, de la création de celui-ci à son sort final (destruction ou conservation).

DIP
Dissemination Information Package - Paquet d'informations diffusé. Paquet d'informations reçu par l'utilisateur en réponse à sa requête au service d'archives. Ce paquet provient d'un ou plusieurs AIP.

Données
Représentation formalisée de l'information, adaptée à la communication, l'interprétation ou le traitement. Exemple : une séquence de bits, un tableau de nombres, les caractères d'une page, un enregistrement de paroles ou un échantillon de roche lunaire.

Durée d'utilité administrative
Durée légale ou pratique pendant laquelle un document est susceptible d'être utilisé par le service producteur ou son successeur, au terme de laquelle est appliquée la décision concernant son traitement final. Le document ne peut être détruit pendant cette période qui constitue sa durée minimale de conservation.

Empreinte numérique
Ensemble de bits caractéristique d'un document numérique. L'empreinte est obtenue par une fonction de hachage. Toute modification du document numérique entraînera une empreinte différente qui révélera la modification par comparaison avec la première empreinte.

Émulation
Imitation du comportement physique d'un matériel par un logiciel pour pouvoir exécuter un programme ou lire un fichier à nouveau.

GLOSSAIRE

Fidélité d'un document
Un document est considéré comme fidèle au document d'origine s'il permet de reconstituer toute l'information nécessaire aux usages auxquels le document d'origine était destiné.

Fonds d'archives
Ensemble des documents de toute nature qu'une personne physique ou morale a produit ou reçu dans l'exercice de ses activités, rassemblés et organisés en conséquence de celles-ci, et conservés en vue d'une utilisation éventuelle.

Format ouvert
Format de fichiers pour lequel il existe une spécification librement accessible et sans restriction d'accès ni de mise en œuvre.

Intégrité
Caractéristique d'une information qui n'a subi aucune destruction, altération ou modification intentionnelle ou accidentelle.

Interopérabilité
Capacité d'un produit ou d'un système à fonctionner avec d'autres produits ou systèmes existants ou futurs et ce sans restriction d'accès ou de mise en œuvre.

Métadonnées
Ensemble structuré d'informations techniques, de gestion et de description attachées à un document servant à décrire les caractéristiques de ce document en vue de faciliter son repérage, sa gestion, son usage ou sa préservation.

Migration logique
Conversion d'un format de fichier jugé non pérenne en un autre format jugé pérenne, de manière à assurer la lisibilité des informations contenues dans le fichier.

OAIS - Open Archival Information System
Système ouvert d'archivage d'information.
Modèle conceptuel destiné à la gestion, à l'archivage et à la préservation à long terme de documents numériques. L'OAIS est officiellement chargé de conserver l'information et de la mettre à disposition d'une communauté d'utilisateurs cible.

Politique d'archivage
Exigences en termes juridiques, fonctionnels, opérationnels, techniques et de sécurité, qu'un système d'archivage électronique interne ou externe doit respecter afin que celui-ci puisse être considéré comme répondant aux besoins exprimés par l'entreprise ou l'organisation qui le met en œuvre.

Producteur
Toute personne physique ou morale, publique ou privée, qui a produit, reçu et conservé des documents/archives dans l'exercice de son activité.

Service versant
Organisme qui transfère des documents d'archives à un service d'archives ; se confond en général avec le service producteur.

SIP
Submission Information Package - Paquet d'informations à verser. Paquet d'informations livré par le producteur à l'OAIS pour l'élaboration d'un ou plusieurs AIP.

Système d'Archivage Electronique (SAE)
Système consistant à recevoir, conserver, communiquer et restituer des archives et qui s'appuie sur une plateforme informatique. Par extension, on parle souvent de Plateforme d'archivage électronique.

Traçabilité
Faculté de pouvoir présenter l'historique des traitements opérés sur un document durant tout le cycle de vie.

Versement
Opération matérielle et intellectuelle par laquelle la responsabilité de la conservation d'archives passe de l'administration à un service de préarchivage ou à un service d'archives, ou bien d'un service de préarchivage à un service d'archives. Ce terme désigne aussi, par extension, les documents ainsi transférés en une seule fois.

Centre Informatique National de l'Enseignement Supérieur
950, rue de Saint-Priest - 34097 Montpellier Cedex 5 - France
Tél. : 04 67 14 14 14 - Fax : 04 67 52 37 63
www.cines.fr - services utilisateurs : svp@cines.fr

Directeur de la publication : Francis Daumas
Co-rédactrices en chef : Marion Massol et Lorène Béchard
Contact - responsable de communication : Claude Kindou