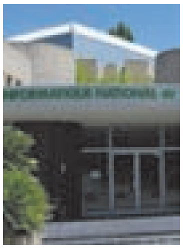
MEDECINE/SCIENCES 2008 ; 24 : 1099-102

La préservation de l'information scientifique et technique

PAC, la plate-forme d'archivage pérenne de documents électroniques du Centre informatique national de l'Enseignement supérieur

Olivier Rouchon

> Pendant près de quinze ans, la question de la conservation à long terme de l'information sous forme numérique n'a été un sujet de préoccupation que pour de rares institutions scientifiques ou patrimoniales qui ont joué un rôle moteur dans la prise de conscience des risques encourus et dans l'émergence de normes de référence dans ce domaine. La progression exponentielle du numérique dans tous les domaines et le caractère impératif de la préservation de l'information ont accéléré cette prise de conscience jusqu'au plus haut niveau de la hiérarchie des administrations et des entreprises. Aussi, ces quatre dernières années ont vu l'émergence de projets, dans les sphères publique ou privée, au niveau national, européen et mondial, visant à développer des infrastructures dédiées à la conservation de l'information électronique. Parmi ces projets, celui du Centre Informatique National de l'Enseignement Supérieur vise à doter la communauté scientifique et technique française d'un véritable service d'archivage à long terme des données sous forme numérique, et est maintenant opérationnel. <



Centre Informatique National de l'Enseignement Supérieur,
950, rue de Saint Priest,
34097 Montpellier Cedex 5, France.
olivier.rouchon@cines.fr

L'archivage pérenne au Centre Informatique National de l'Enseignement Supérieur (CINES)

L'archivage pérenne des documents électroniques consiste à conserver le document et l'information qu'il contient dans son aspect physique comme dans son aspect intellectuel, sur le très long terme (trente ans et au-delà), et de manière à pouvoir le rendre accessible et compréhensible. Or, la plupart des fichiers informatiques de plus de dix ans sont aujourd'hui pratiquement illisibles, conséquence de plusieurs facteurs inéluctables dont la connaissance perdue du contenu des fichiers, l'obsolescence des formats de fichier, la détérioration des supports physiques ou encore la disparition des logiciels ou matériels de lecture [12]. Il y a donc des choix à faire concernant la définition des processus et des pratiques de préservation à mettre en œuvre pour atténuer les effets de ces risques lorsqu'ils se produiront.

Depuis 2004, le CINES *(Figure 1)* travaille à la mise en place d'un service pour l'archivage pérenne du patrimoine scientifique, qui permettrait à tout organisme produisant ou collectant en grande quantité des documents électroniques dont le contenu possède une valeur patrimoniale avérée pour la communauté de l'Information Scientifique et Technique, d'initier un projet d'archives, dans le respect du contexte législatif « archivistique » français.

Tout d'abord, une équipe dédiée à la plate-forme d'archivage a été constituée : elle est chargée de couvrir les aspects organisationnels (définition et expertise des processus métiers et des méthodes), et culturels (renforcement des collaborations entre informaticiens, archivistes et bibliothécaires), en plus de l'aspect technique. Ensuite, les aspects fonctionnels et technologiques ont été analysés, à la fois sur le plan théorique et sur le plan pratique, avec notamment les retours d'expérience d'autres organismes travaillant sur des projets similaires. La future plate-forme a commencé à

Article disponible sur le site http://www.medecinesciences.org ou http://dx.doi.org/10.1051/medsci/200824121099

se dessiner, en suivant des contours donnés par les normes internationales en vigueur.
Quelques exemples : pour éviter la perte d'informations relatives au document électronique et à son contenu, des métadonnées génériques décrivant les propriétés du document (auteur, titre, résumé, mots-clés, etc.) ont été utilisées afin de le replacer dans son contexte et en préserver le sens. Il a également été décidé d'attribuer un identifiant unique et pérenne aux documents au moment de leur archivage pour permettre, notamment, de les retrouver et de les référencer. Les formats de fichiers durables ont été privilégiés afin d'éviter une obsolescence trop rapide des formats de fichiers acceptés par la plateforme. Tout document versé est vérifié pour s'assurer que les formats de fichiers qu'il contient sont conformes. En outre, des procédures de veille technologique et de migration logique ont été élaborées pour identifier les formats émergeants ou obsolètes, et transférer les fichiers d'un format désuet vers un format pérenne. Des outils de gestion du vieillissement des supports utilisés pour conserver les documents, ainsi que des procédures de migration physique ont été définis, accompagnés d'un effort de veille et d'anticipation sur les technologies émergentes en termes de médias de stockage.

Figure 1. ***Le Centre Informatique National de l'Enseignement Supérieur.***

Après trois années de conception et de développement, une première version du système PAC (plate-forme d'archivage au Cines) [1] a été mise en service au printemps 2007 avec comme axe initial une intégration avec l'application STAR (signalement des thèses, archivage et recherche) [2] (développée sous la responsabilité de l'ABES, Agence bibliographique de l'enseignement supérieur) pour le dépôt, la diffusion, le référencement et l'archivage des thèses électroniques.
Toutefois, l'infrastructure matérielle ne permettant pas de gérer le volume de données anticipé pour de nouveaux projets d'archives, il a été décidé de procéder à un appel d'offres pour l'acquisition, au printemps 2008, d'une plate-forme capable de gérer de larges volumes (plus de 40 Téra-octets). C'est donc une deuxième version du système PAC qui est actuellement exploitée, privilégiant toujours la même approche généraliste, évitant ainsi le traitement des projets au cas par cas, et permettant de mutualiser la plateforme pour tous les projets d'archives.

L'architecture de la plate-forme d'archivage

Le système PAC a été conçu comme un ensemble de trois serveurs logiques, s'inspirant du modèle proposé par la norme ISO 14721 (OAIS, *open archival information system)* [3], mettant à disposition des différents acteurs impliqués les principales fonctionnalités du processus d'archivage *(Figure 2)*. La plateforme se compose d'un serveur de transfert auquel le service versant, qui collecte les documents auprès de sa communauté de producteurs, pourra transmettre ses archives, d'un serveur de stockage où sont conservés les documents sous la responsabilité du service d'archives qui en assure l'administration, et d'un serveur d'accès où le service versant et éventuellement les utilisateurs autorisés à consulter ses archives pourront rechercher et obtenir une copie des documents archivés. L'ensemble est supervisé par un service de contrôle, qui s'assure que les échanges respectent la codification en vigueur (Code du patrimoine en matière de communicabilité des archives, législation relative au droit intellectuel, au droit des auteurs, au droit de reproduction et au droit de représentation des œuvres de l'esprit).
L'implémentation des processus d'échanges s'appuie (très) largement sur le Standard d'échanges de données pour l'archivage [4] défini par la DGME (direction générale de la modernisation de l'État) [5] et la DAF (direction des archives de France) [6], qui est en cours de normalisation au niveau européen, et qui a déjà été utilisé pour l'implémentation du projet PILA@E des Archives de France.

Les principes de fonctionnement

Le service versant transmet par réseau *via* un protocole sécurisé, ou sur support amovible (suivant le protocole de transfert préalablement établi), les documents à verser au système PAC *(Figure 3)*. Le transfert est repéré par l'application, qui envoie par courriel un accusé de réception du transfert et procède alors à une série de contrôles de validité technique des données transférées, en s'assurant que les objets transférés respectent les conditions définies par le protocole de transfert : conformité des formats, structure de versement...
Si le résultat du contrôle est positif, le système PAC enverra un certificat d'archivage au service versant, signalant que l'objet à archiver est accepté et transféré sur le serveur de stockage. Cette notification comporte l'identifiant unique et pérenne qui permettra au service versant ou à des utilisateurs de retrouver l'archive. Dans le cas contraire, il enverra une notification de rejet, accompagné de la raison du refus.

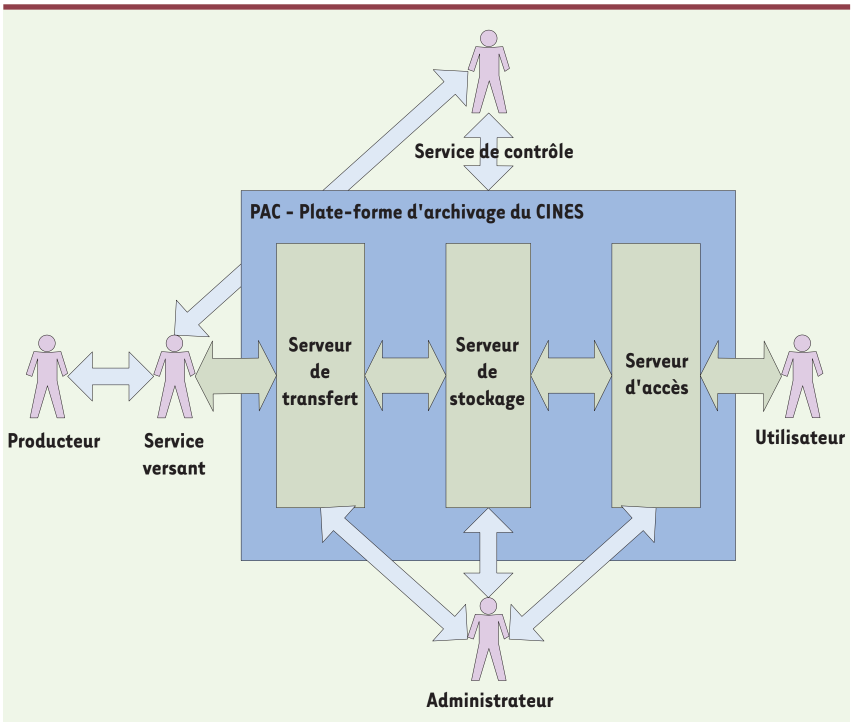


Figure 2. ***Architecture logique de PAC.***

vue d'un doctorat [7]. Cette disposition prévoit qu'un doctorant qui doit déposer son mémoire de thèse dans la Bibliothèque de l'Université dans laquelle il va effectuer sa soutenance, peut le faire soit sous forme papier, soit sous forme électronique. Dans ce deuxième cas, la Bibliothèque va déposer les documents numériques constituant le mémoire de thèse à l'ABES grâce à l'application STAR, puis ajouter des informations qualifiant la thèse, sous forme de métadonnées. Après plusieurs phases de validation, la thèse sera archivée au CINES. L'intégration de la plate-forme PAC à STAR a été réalisée, et le projet est maintenant en exploitation. Après avoir démontré la faisabilité, et les choix retenus pour la première version de la plateforme, il est en phase de montée en charge, qui se fait au rythme de l'implantation de l'application STAR dans les universités.

Le deuxième projet, actuellement en phase de test de recette, concerne l'archivage de revues en Sciences Humaines et Sociales (SHS), dans le cadre du projet Persée. Cette initiative vise à valoriser et à préserver des collections rétrospectives originales au format papier dont certaines ont près de cent ans. L'université Lumière de Lyon 2 y est chargée de la digitalisation de masse des revues, de la centralisation et de la robotisation des traitements, de la description des collections et de leur mise en ligne *via* le portail Persée. La chaîne de numérisation est intégrée à la plateforme PAC, et les documents électroniques créés sont déposés au CINES pour une préservation à long terme.

D'autres projets d'archives sont en cours de discussion, parmi lesquels l'archivage des documents versés dans la plate-forme HAL (hyper article en ligne) [8, 13] ou encore l'archivage de *corpus* sonores pour le CRDO (centre CNRS pour la conservation et la diffusion de corpus oraux) [9], dans le cadre d'un projet pilote pour le TGE-Adonis (TGE = très grand équipement).

En parallèle à ces projets, le CINES est membre de plusieurs initiatives ou groupes de travail au plan national ou européen : il participe depuis 2004 au groupe PIN (Préservation de l'information numérique) de l'association ARISTOTE, dont les principaux acteurs français du domaine de l'archivage public (CNES, centre national d'études spatiales, BnF, Bibliothèque nationale de France, DAF, direction des archives de France) font également partie. Au niveau européen, le CINES est

Tout document archivé sur le serveur de stockage fait l'objet d'une copie multiple : deux exemplaires du document sont conservés sur des baies de disques[1] situées dans des salles machines différentes, et un troisième exemplaire du document est sauvegardé sur bande magnétique. Tout événement, action ou manipulation des données ou métadonnées sous la responsabilité du service d'archives est horodaté, historisé et traçable.

La communauté des utilisateurs ayant accès aux archives est définie par le service versant, en collaboration avec le service d'archives, lors de la phase préalable du projet d'archives. L'utilisateur authentifié par le système PAC pourra consulter le catalogue d'un projet d'archives particulier et faire une demande de communication d'archives. Si l'archive est communicable, le système PAC pourra mettre une copie de l'archive à disposition de l'utilisateur, avec éventuellement l'accord du service versant.

Les projets d'archives au CINES

Actuellement, deux projets sont en cours de réalisation. Le premier concerne l'archivage pérenne de thèses collectées par l'ABES - Agence Bibliographique de l'Enseignement Supérieur - par l'intermédiaire de l'outil STAR, initié à la suite de l'arrêté du 7 août 2006 relatif aux modalités de dépôt, de signalement, de reproduction, de diffusion et de conservation des thèses ou des travaux présentés en soutenance en vue d'un doctorat.

[1] En informatique, espace d'un boîtier d'ordinateur, où l'on place les périphériques internes (disques durs, lecteurs, etc.).

FORUM DÉBAT

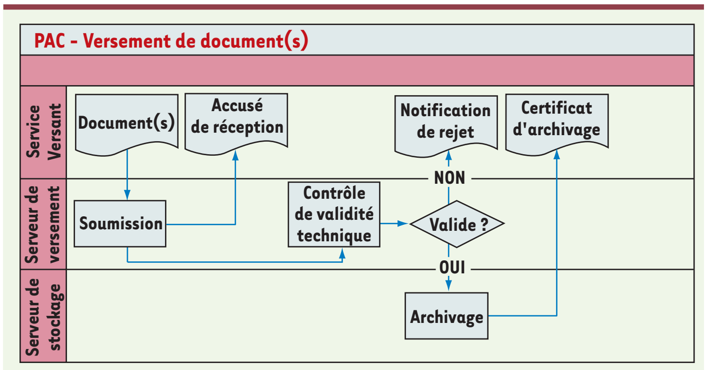


Figure 3. ***Le versement de document dans PAC.***

membre depuis 2007 de l'Alliance pour la préservation de l'information scientifique [10]. Cette initiative a pour objectif la coordination de la mise en place d'une infrastructure européenne dans le domaine de l'archivage pérenne. À cette même date, le CINES a été référencé centre de formation sur l'archivage par la DPE (*digital preservation Europe*) [11].

Le CINES est maintenant un acteur reconnu du domaine de la préservation à long terme de documents numériques, et va jouer un rôle clé dans la réussite au niveau national d'une stratégie pour l'archivage pérenne de documents électroniques produits par la communauté de l'information scientifique et technique. Il reçoit d'ailleurs de nombreuses sollicitations de laboratoires et d'universités pour divers services, tels que l'aide et le conseil dans la construction de projets d'archivage à long terme, retours d'expérience, propositions de projets d'archives. ◊

SUMMARY

The long term preservation of scientific and technical information. PAC, the archive platform for digital documents at the Centre Informatique National de l'Enseignement Supérieur

During a fifteen years period, the long-term preservation of digital information has only been a matter under consideration for a few scientific or patrimonial institutions. These have played a key role in the understanding of the subsequent risks and the definition of standards in this domain. The exponential progress of the digital information in every domain, as well as the mandatory aspect of its preservation have sped up the awareness process even at the highest management level of companies or public administrations. Thus, a significant number of projects have kicked off during the last four years, with the objective of rolling out infrastructures dedicated to the long term preservation of electronic data. Among those projects, the one currently run at the CINES, which main goal is to provide the scientific and technical community in the Higher Education and Research sectors with a genuine long term preservation service for digital information, is now operational. Here is a brief outline of the PAC system... ◊

RÉFÉRENCES

1. PAC. *Plate-forme d'Archivage du CINES*. http://www.cines.fr/
2. STAR. *Signalement des Thèses, Archivage et Recherche, Agence Bibliographique de l'Enseignement Supérieur (ABES). http://www.abes.fr/*
3. OAIS. *Open Archival Information System. Modèle de référence pour un système ouvert d'archivage d'information*. http://vds.cnes.fr/pin/documents/projet_norme_oais_version_francaise.pdf
4. https://www.ateliers.modernisation.gouv.fr/ministeres/projets_adele/a103_archivage_elect/public/standard_d_echange_d/archives_echanges_v0/downloadFile/file/archives_echanges_v0-1_description_standard_v1-0.pdf?nocache=1141748589.19
5. DGME. Direction Générale de la Modernisation de l'État. http://www.modernisation.gouv.fr/
6. DAF : Direction des Archives de France. http://www.archivesdefrance.culture.gouv.fr/
7. http://www.legifrance.gouv.fr/jo_pdf.do?cidTexte=JORFTEXT000000635069
8. HAL. Hyper Article en Ligne, Archives Ouvertes. http://hal.archives-ouvertes.fr/
9. CRDO : Centre de Ressources pour la Description de l'Oral. http://crdo.risc.cnrs.fr/exist/crdo/
10. Alliance for Permanent Access to the Records of Science. http://www.alliancepermanentaccess.eu/
11. Digital Preservation Europe - http://www.digitalpreservationeurope.eu/
12. Huc C. La pérennisation des informations sous forme numérique : risques, enjeux et éléments de solution. *Med Sci (Paris)* 2008 ; 24 : 653-7.
13. Duchange N, Autard D, Pinhas N. Le libre accès : une opportunité pour la recherche biomédicale. *Med Sci (Paris)* 2008 ; 24 : 771-5.

TIRÉS À PART

O. Rouchon