

Leur quantité croît de façon exponentielle, comme les volumes de données produites.

Des centres d'archivage ont alors été développés dans plusieurs pays depuis une quinzaine d'années. Ils ont d'abord été l'œuvre d'organismes publics, tels des agences spatiales, des instituts de recherche, des bibliothèques nationales ou des services administratifs. Un marché privé s'est aussi mis en place, mais avec des qualités de service inégales et des garanties diverses sur la pérennité des données, assurée plus souvent à court ou moyen terme qu'à long terme.

En outre, de grandes infrastructures ont été créées, à des échelles nationales ou internationales. Par exemple, les projets européens PRACE (*Partnership for advanced computing in Europe*, ou Partenariat pour le calcul avancé en Europe) et Eudat (*European data infrastructure*, ou Infrastructure de données européenne) visent à se doter respectivement d'importants moyens de calcul (et de stockage des données associées) et d'une plate-forme collaborative de partage et de conservation des données scientifiques.

Plus de huit millions de fichiers archivés

En France, le pionnier de l'archivage numérique est le CNES (Centre national d'études spatiales), qui archive depuis 1998 des images prises par des satellites et de grands observatoires, ainsi que diverses données de calcul. Le CINES, quant à lui, dispose d'un service d'archivage numérique opérationnel depuis 2006. Il a une capacité de stockage de plusieurs pétaoctets (10^{15} octets), soit l'équivalent d'une pile de CD d'environ dix kilomètres de hauteur. Plus de huit millions de fichiers sont aujourd'hui archivés, occupant une centaine de téraoctets (10^{12} octets). Le CINES a reçu un agrément (une certification) délivré par le SIAF (Service interministériel des archives de France) pour les archives dites intermédiaires, c'est-à-dire les documents qui ne sont pas nécessaires à l'activité quotidienne des services, mais qui doivent être conservés un certain temps pour des raisons administratives ou juridiques. La Commission européenne lui a aussi délivré une accréditation validant la qualité de son système d'archivage électronique.

Le CINES est chargé d'archiver toutes les thèses de doctorat françaises au format numérique. Il participe au projet de Bibliothèque scientifique numérique nationale, ou BSN (voir l'article de Frédéric Blin dans

L'ESSENTIEL

■ D'énormes quantités de données numériques sont stockées dans de grands centres d'archivage dotés d'infrastructures sécurisées.

■ Leur préservation à long terme pose plusieurs difficultés : détérioration et évolution des supports, formats de fichiers qui se périment, etc.

■ On y répond le plus souvent par une gestion active des archives, incluant la migration des documents vers de nouveaux supports et l'actualisation de leur format.

ce numéro), où il pilote les questions liées à l'archivage du patrimoine scientifique public. Il héberge une partie des données de nombreux projets et organismes : la Très grande infrastructure de recherche (TGIR) HumaNum du CNRS (composée de moyens humains et numériques destinés aux sciences humaines), plusieurs bibliothèques universitaires, divers organismes de recherche (INSERM, IRSTEA, etc.), la cour des comptes... Il héberge aussi la première plate-forme numérique française de cours en ligne ouverts et massifs (CLOM, ou MOOC en anglais), dont le lancement a été annoncé début octobre 2013. En outre, il mène des actions de sensibilisation et de formation aux problématiques de l'archivage auprès de la communauté scientifique.

Outre l'archivage, le CINES propose des moyens de calcul numérique intensif à l'ensemble de la communauté scientifique. Le centre héberge notamment l'un des plus puissants supercalculateurs européens, financé par la société civile GENCI (Grand équipement national de calcul intensif) et utilisé pour des simulations complexes dans de nombreux domaines (climatologie, mécanique des fluides, physique des plasmas, astrophysique, géophysique, biologie, etc.). Archivage et calcul sont complémentaires, car les données servent d'entrées aux calculs qui, eux-mêmes, produisent de grandes quantités de données, dont une partie doit être conservée. Cependant, les volumes de données étant colossaux, on ne leur applique pas certaines procédures d'archivage à long terme (telle la copie du document en cinq exemplaires, sur laquelle nous reviendrons).

Une cinquantaine d'ingénieurs et de techniciens travaillent au CINES. Une douzaine d'entre eux se consacrent aux projets d'archivage, qui font aussi parfois appel à des intervenants extérieurs. Les profils sont variés : archivistes (spécialistes de la gestion des documents), informaticiens, juristes (garants du respect de la réglementation en vigueur), etc.

Des infrastructures sécurisées

Le centre s'est doté d'une infrastructure et d'une plate-forme logicielle adaptées à l'archivage. Les données y sont accessibles en permanence *via* le réseau Internet. L'alimentation électrique des équipements liés à l'archivage est redondante, afin de

prévenir les coupures de courant : elle est garantie par deux lignes électriques d'ERDF indépendantes et par des groupes électrogènes ; en outre, divers dispositifs de protection sont installés. Les conditions climatiques (température, humidité) des salles hébergeant les machines sont contrôlées. L'accès au site est sécurisé (filtrage des visiteurs, accès restreint à certaines zones), de même que les connexions *via* le réseau (grâce à des pare-feux, des protocoles sécurisés, des antivirus). En outre, toutes les données informatiques du centre (archives et systèmes) sont sauvegardées quotidiennement. Cette sauvegarde s'effectue de façon incré-

mentale, c'est-à-dire qu'on n'enregistre que les changements.

La plate-forme d'archivage du CINES, conforme aux normes internationales du domaine, est fondée sur trois « serveurs logiques », respectivement dits de transfert, de stockage et d'accès. Le premier reçoit les documents à archiver, le deuxième les conserve et le troisième permet leur consultation. Décrivons-les plus en détail.

Un paquet d'information à archiver contient l'objet à conserver et un ensemble de métadonnées (des données permettant de comprendre d'autres données). Ces dernières sont élaborées en partie par l'organisme

dit versant, qui fournit les documents à archiver (c'est en général l'organisme qui les produit) et en partie par l'archiviste. Elles décrivent le contexte (organisme producteur, organisme versant), le document lui-même (titre, auteur, format), les paramètres de gestion (localisation physique, dispositif de stockage, etc.). Un autre type de métadonnées, qui ne sont pas associées à un document précis mais à la plate-forme d'archivage, décrit les informations dites de représentations, essentielles à la lecture de l'archive : les caractéristiques des formats, les structures des métadonnées, la description de la plate-forme d'archivage et des normes utilisées...

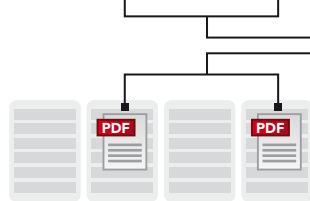
PLUSIEURS COPIES SUR DES SUPPORTS VARIÉS

Les supports physiques sont vulnérables et les estimations de leur durée de vie varient selon les études. Pour parer cette menace, cinq copies du document à archiver sont réalisées sur des types de supports différents : deux sur des bandes magnétiques et trois sur des disques durs (dont une située à plus de 300 kilomètres des autres, afin de préserver le document d'une catastrophe majeure, tel un séisme). Chaque document est accompagné d'un ensemble de métadonnées qui permettent de l'identifier, de le localiser et de faire évoluer son format.

Bandes magnétiques



Disques durs



PDF

PDF

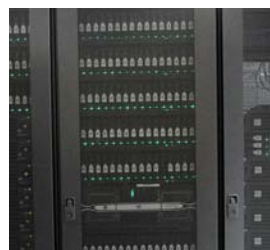
Disque dur dans un centre de données éloigné de plus de 300 km

Métadonnées :

Titre : L'archivage électronique
Auteur : Daumas, Francis
Identifiant : ark://85579/123456789
Date d'archivage : 30-09-2013
Format : PDF
Version du format : 1.7
...



Les robotiques gèrent de façon automatique un ensemble de bandes magnétiques (ci-dessus, une robotique vue de l'extérieur et, ci-contre, à gauche, vue de l'intérieur). Deux copies du document y sont enregistrées. Lorsqu'une archive est demandée, un bras robot préleve la bande où elle est stockée et la place dans un dispositif de lecture.



Les baies de stockage renferment de multiples disques durs. Deux copies du document y sont stockées, dans des salles distinctes.

