L’archivage électronique dans l’enseignement supérieur et la recherche : problématique et solutions

Lorène Béchard

L'archivage électronique dans l'enseignement supérieur et la recherche : problématique et solutions

Lorène BÉCHARD

Introduction

Confronté à la production massive de documents numériques, le monde de l'enseignement supérieur et de la recherche (ESR) connaît un véritable bouleversement dans les pratiques traditionnelles de la recherche. De la sélection des sources à la diffusion des résultats, l'information est de plus en plus accessible sous forme numérique, qu'elle le soit nativement ou à la suite d'une numérisation. Alors que des structures « archives » voient le jour au sein des établissements et se constituent en réseau afin de mutualiser des pratiques et des outils, l'ampleur de la tâche peut sembler considérable : plusieurs centaines de mètres linéaires d'arriérés papier mais aussi des serveurs informatiques, disques durs, voire des cédéroms ou des clés USB sur lesquels s'accumulent des fichiers, dont la gestion est dans la plupart des cas seulement confiée aux services informatiques. Face à ces « nouvelles » archives, l'archiviste doit alors faire reconnaître ses compétences de gestionnaire de l'information et doit pouvoir en acquérir de nouvelles en systèmes d'information afin de travailler en collaboration avec les services informatiques à la prise en charge des documents. Au-delà de la problématique classique liée à la conservation des fichiers électroniques, l'archivage de la production scientifique numérique représente un défi supplémentaire dans le sens où les données produites sont relativement complexes tant sur la forme que sur le fond. Aussi, le ministère de l'enseignement supérieur et de la recherche (MESR) a mandaté le Centre informatique national de l'enseignement supérieur (CINES) pour mettre en place une solution d'archivage électronique mutualisée pour la communauté ESR.

Le patrimoine numérique de l'enseignement supérieur et la recherche : des documents administratifs aux données scientifiques, considérations techniques

0100001101110101011100100110100101100101011101010111100000100000001110110010110100101001

À la source, un document créé sous une forme informatique n'est qu'une suite de deux états représentés par des 0 et des 1, à l'image du sous-titre de cette partie[1]. Aussi, à moins de parler couramment le binaire, on comprend dès lors l'ampleur de notre dépendance vis-à-vis des outils informatiques, capables d'interpréter ce langage et de nous en restituer un contenu intelligible.

L'objectif premier de la mise en place de procédures d'archivage électronique sera donc de ne surtout pas perdre le lien entre les différents maillons de la chaîne informatique qui relient la source binaire au rendu visible et compréhensible sur un écran d'ordinateur. Pour cela, il faut bien identifier en amont l'ensemble des maillons de cette chaîne afin de mettre en place des actions spécifiques pour maintenir ce lien sémantique et technique ou pouvoir le recréer en cas de besoin.

Les maillons de la chaîne informatique © CINES

[1] La minute du *geek* : à vous de décoder le train de bits !

Ces actions à mettre en place relèvent de la gestion du risque. Elles pourront prendre la forme :

- de la mise en place de jeux de métadonnées pour accompagner les archives. Elles devront être définies en fonction des besoins de description de l'information à archiver, en essayant le plus possible de respecter les standards existants pour faciliter l'interopérabilité. Cette notion de métadonnées n'est pas nouvelle pour les archivistes mais elle revêt un aspect plus technique dans le contexte numérique à la fois à cause de la nature même des objets à décrire (format du fichier informatique, encodage, compression, etc.) mais aussi par la manière d'exprimer ces métadonnées (schémas XML, RDF, etc.) ;

- d'une surveillance du vieillissement des supports de stockage ainsi que d'une veille pour anticiper les changements de technologie (tant sur les supports de stockage que sur les appareils de lecture de ces supports) ; ces actions sont également mises en œuvre en général dans le cadre d'une sauvegarde sécurisée ;

- d'une sélection des formats de fichiers pouvant être archivés. S'engager sur la lisibilité d'un fichier dans le temps nécessite dans la plupart des cas[1] d'être en mesure de pouvoir le convertir lorsque ce sera nécessaire. Pour cela, il faut *a minima* avoir accès aux spécifications de ce format (c'est-à-dire au document expliquant la manière dont ce format est composé), ce qui est le cas notamment du format PDF. En effectuant une veille sur l'utilisation de ces formats de fichiers et de leurs outils de lecture, on pourra alors détecter les signes d'obsolescence et entreprendre le cas échéant des conversions après avoir identifié le format cible.

Depuis quelques années déjà, ces différentes actions se généralisent dans les organismes réalisant de l'archivage électronique et les recommandations à ce sujet sont de plus en plus nombreuses[2]. Dans la majorité des cas, les documents produits par une administration sont des fichiers bureautiques, ce qui ne représente pas de difficultés particulières en termes d'archivage numérique. Mais devant l'importance croissante des applications reposant sur des bases de données, l'extraction de l'information pertinente à archiver devient plus complexe. Les systèmes d'information intègrent de plus en plus

[1] Seule la stratégie d'émulation repose sur le maintien des outils de lecture du fichier de manière à ce qu'il puisse s'exécuter dans les mêmes conditions que celles de son environnement de création.

[2] À l'image du guide des bonnes pratiques réalisé en 2012 dans le cadre du mandat archivage électronique de la DISIC et disponible en ligne : http://references.modernisation.gouv.fr/sites/default/files/DISIC_AE_Guide_bonnes_pratiques_0.pdf

des architectures multicouches dans lesquelles les données, les traitements effectués et les formes de présentation sont fortement imbriqués. Dans ces cas-là, les données initiales ne sont peut-être qu'une partie de ce qu'il faut archiver. La question essentielle étant : quelle utilisation sera faite de ces données ? La stratégie d'archivage à mettre en œuvre dépend en grande partie de la réponse à cette question[1].

La particularité des données scientifiques

Outre la production documentaire bureautique, les établissements d'enseignement supérieur et de recherche doivent également prévoir l'archivage des données et documents produits dans le cadre des activités de recherche. Or, l'arrivée de l'informatique, tout en démultipliant le champ des possibles pour la recherche, a aussi considérablement accru la complexité des données à traiter pour les archivistes, d'abord parce que le domaine de la recherche scientifique est à la fois vaste et très spécialisé. Il s'agit principalement de pouvoir prendre en compte de fortes volumétries de fichiers (aspect quantitatif) dans des formats de données variés et complexes, et souvent sous-documentés (aspect qualitatif).

Cet aspect quantitatif des données scientifiques est étroitement lié à l'augmentation continue de la puissance des outils utilisés pour créer les données. Qu'ils s'agissent de supercalculateurs capables d'aboutir à des modèles de plus en plus fins ou d'appareils de mesure de haute précision, l'impact sur la quantité et le volume des informations produites est considérable. Alors que l'on parle en Go, voire en To, pour la production bureautique d'une administration, un fonds de données scientifiques, quant à lui, se mesure généralement en Po[2]. Or, on constate un risque de perte des données non négligeable car, faute de moyens et d'expertise, les systèmes de stockage mis en place dans les laboratoires sont en fait assez sous-évalués[3].

Sur le plan qualitatif, du fait de la spécificité de chaque discipline, la structure ou le format d'une donnée scientifique n'est souvent pleinement exploitable que par le chercheur ou le laboratoire qui l'a produite. Bien que certains formats soient largement utilisés tels que HDF ou NetCDF, il reste très

1 Pour plus d'informations, voir le *Guide méthodologique pour l'archivage des bases de données*, CINES, mars 2013.
2 1 Pétaoctet = 1000 To.
3 Résultats d'une enquête réalisée par le CINES en 2011 auprès de 155 laboratoires français.

fréquent d'avoir des données dans un format « maison » binaire nécessitant alors une forte connaissance de leur organisation interne. Cette connaissance peut être décrite à l'aide de fichiers annexes, de métadonnées, de rapport ou dans une thèse. Dans le cas extrême où elle ne serait comprise que par le producteur de la donnée, il sera alors le seul garant de la réutilisabilité de cette donnée. Au-delà de l'aspect syntaxique, une des autres difficultés est d'appréhender le contenu informationnel de ces fichiers. Il ne s'agit pas pour l'archiviste d'arriver à comprendre ces données (elles sont généralement trop spécifiques et complexes) mais de s'assurer qu'elle sera comprise par les utilisateurs finaux. De fait, encore plus que dans n'importe quel autre contexte, une étroite collaboration avec les producteurs ou une communauté structurée autour de sa thématique est cruciale, et ce dès la création des données. Elle permet d'assurer correctement l'identification de la typologie des données, leur évaluation pour fixer la durée d'utilité et le sort final, ainsi que le renseignement des métadonnées.

L'offre d'archivage mutualisé proposée par le ministère de l'Enseignement supérieur et de la recherche

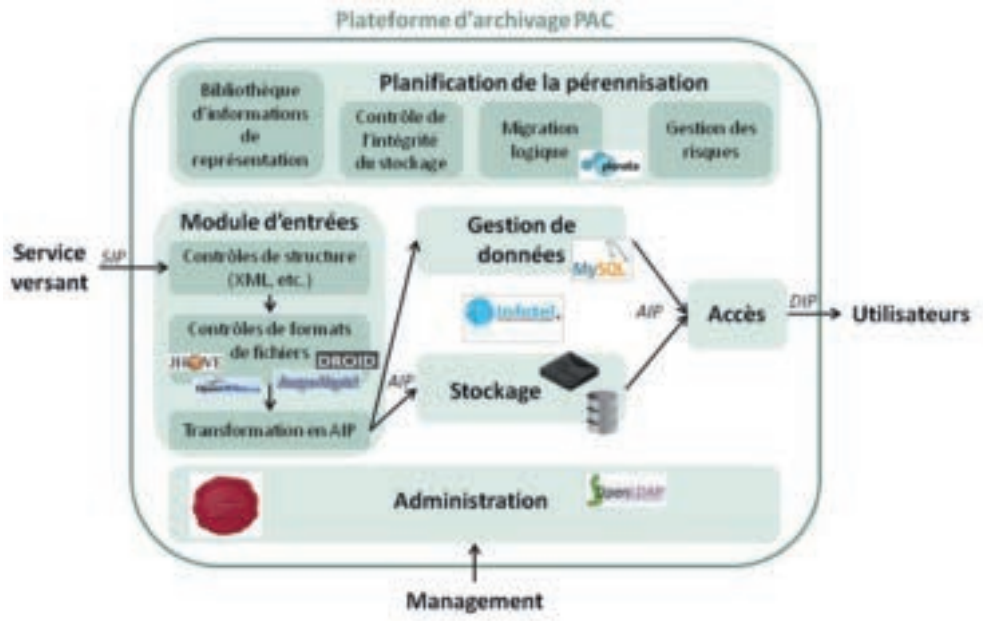
Devant les enjeux soulevés, le ministère de l'enseignement supérieur et de la recherche a été un des premiers ministères à s'intéresser de près à la question afin de proposer des solutions mutualisées permettant de préserver le patrimoine documentaire numérique produit par la communauté ESR. À ce titre, dès 2004, il a mandaté le Centre informatique national de l'enseignement supérieur (CINES) qui, en tant que centre de calcul, disposait d'infrastructures et d'une expertise informatique. La mise en œuvre d'une compétence archivistique début 2006 (*via* un recrutement) a permis d'apporter l'expertise en gestion documentaire qui manquait et par là même, de positionner, sur le plan national, le CINES comme centre de compétences sur le sujet et centre d'archivage pour la communauté ESR.

L'archivage intermédiaire et pérenne des documents numériques : la plateforme PAC

Les services proposés dans le cadre de la plateforme PAC s'adressent principalement à l'ensemble des organismes publics dépendant du MESR ayant des données ou des documents à conserver sur le moyen et long terme dès lors que le *corpus* à conserver est reconnu d'intérêt national. Ces archives peuvent être des documents administratifs[1], des publications (thèses, ouvrages numérisés, périodiques, etc.) ou encore des données scientifiques (relevés d'observations, résultats de simulations, etc.). Pour être éligibles à un archivage dans PAC, elles doivent respecter certaines exigences comme ne plus subir de modifications, être accompagnées d'une description ou encore être encodées dans un format de fichiers spécifié, ouvert et largement utilisé. Il faudra s'assurer par ailleurs que le contexte légal de production du document permet son archivage (droits de propriété intellectuelle). Afin d'éviter les redondances et limiter les coûts, seules des informations « brutes » ou ayant une plus-value informationnelle sont archivées. Il n'est en effet pas nécessaire de conserver des documents pouvant être reconstitués à partir de données déjà présentes dans le système d'archivage.

Après deux années d'études et de développements informatiques, une première solution d'archivage électronique (PAC v1) a vu le jour pour archiver les thèses de doctorat soutenues en France et déposées au format numérique. Face à l'arrivée de nouveaux projets d'archivage (Persée, plateforme HAL, partenariat avec le TGE Adonis, etc.), le besoin de recourir à une solution du marché, capable de prendre en charge une plus forte volumétrie, s'est exprimé. En 2008, une deuxième version de la plateforme (PAC v2) a donc été mise en place. Elle repose sur le progiciel Arcsys (société Infotel) accompagné de modules complémentaires métier développés en interne : module d'entrées, module de contrôle des formats de fichiers, attribution d'un identifiant unique et pérenne à chaque archive (de type ARK), module de contrôle de l'intégrité du stockage, bibliothèque d'informations de représentation, etc. La plateforme est mutualisée entre tous les projets d'archives afin de diminuer les coûts : partage des infrastructures techniques, des procédures d'archivage, des contrôles, etc.

[1] Depuis fin 2010, le CINES est agréé par les Archives de France pour la conservation d'archives publiques courantes et intermédiaires au format électronique.



Les fonctionnalités de la plateforme PAC © CINES

La solution proposée est basée sur les normes du domaine. Le socle conceptuel est issu de l'OAIS[1]. L'implémentation technique respecte notamment les exigences de la NF Z 42-013[2] et est conforme au Standard d'échange de données pour l'archivage (SEDA). La description des archives s'effectue à plusieurs niveaux grâce aux jeux de métadonnées du Dublin Core non qualifié ou du SEDA (niveau document), un ensemble de métadonnées techniques et de gestion spécifiques (niveau document et fichier), ainsi que des métadonnées tirées d'ISAD/G et d'ISAAR/CPF (niveau projet d'archives/fonds).

Une équipe de douze personnes composée d'ingénieurs informatiques et d'un archiviste a été mise en place pour assurer le fonctionnement de cette plateforme et des activités afférentes. L'expertise apportée couvre les formats de fichiers (aide au choix, identification et contrôle, migration logique), les métadonnées (sélection, organisation, mappage, contrôle) et de manière plus large toutes les informations permettant de comprendre les documents archivés (bibliothèque d'informations de représentation). Par exemple, dans le cadre de l'archivage des

[1] ISO 14721 - *Open Archival Information System*.

[2] 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.

revues numérisées par le programme Persée, la cellule d'expertise formats a conseillé l'archivage des documents au format PNG plutôt qu'au format TIFF, ce dernier étant considérablement plus volumineux à qualité égale. Cela a donc entraîné une modification de la chaîne de numérisation des revues.

La mise en œuvre d'un projet d'archivage au CINES est encadrée par un référent informaticien et un archiviste qui travaillent en collaboration avec les équipes de l'organisme souhaitant archiver ses données (service versant). Cela prend la forme d'un accompagnement archivistique personnalisé pour décider notamment des termes de la convention, de la granularité de l'archivage, des profils de données et du contenu des métadonnées. En parallèle, le référent sert d'interlocuteur privilégié pour l'ensemble des aspects techniques : connexions des machines, sélection des formats de fichiers, conseils d'implémentation, assistance lors des phases de tests, etc. Les choix, qu'ils soient techniques ou archivistiques, sont pris d'un commun accord. Côté service versant, l'équipe projet mise en place allie généralement des compétences informatiques et archivistiques ou documentaires. Elle se charge notamment de réaliser l'outil informatique qui met en forme les données telles qu'attendues et les envoie à la plateforme d'archivage. En moyenne, la durée de mise en place d'un projet est de 6 à 18 mois. Cela varie en fonction de la complexité des données à archiver, de la disponibilité des acteurs impliqués ou encore de l'état d'avancement dans l'identification de ce qu'il faut archiver (documents et métadonnées).

Début 2013, la plateforme d'archivage PAC comptait onze services versants, parmi lesquels plusieurs bibliothèques universitaires ayant numérisé leurs fonds d'ouvrages anciens (Cujas, BUPMC, BIU Santé, Bibliothèque Sainte-Geneviève) ou encore plus récemment la Cour des comptes dans le cadre de l'archivage intermédiaire des documents produits par les juridictions financières. Capable d'accueillir 40 To de données dans sa configuration actuelle (prochainement étendue à 80 To), PAC conserve près de 25 To utiles[1], soit plus de 330 000 documents.

Sur la volonté du ministère, plusieurs solutions sont en train de voir le jour pour faciliter la prise en charge du coût lié à ce service[2]. Parmi elles, les offres du TGE Adonis et de BSN6[3] vont permettre de financer tout ou partie de cet archivage sur la base d'appels à projets fondés sur l'intérêt scientifique et la qualité technique du *corpus* numérique.

---

1 Au 30 avril 2013.

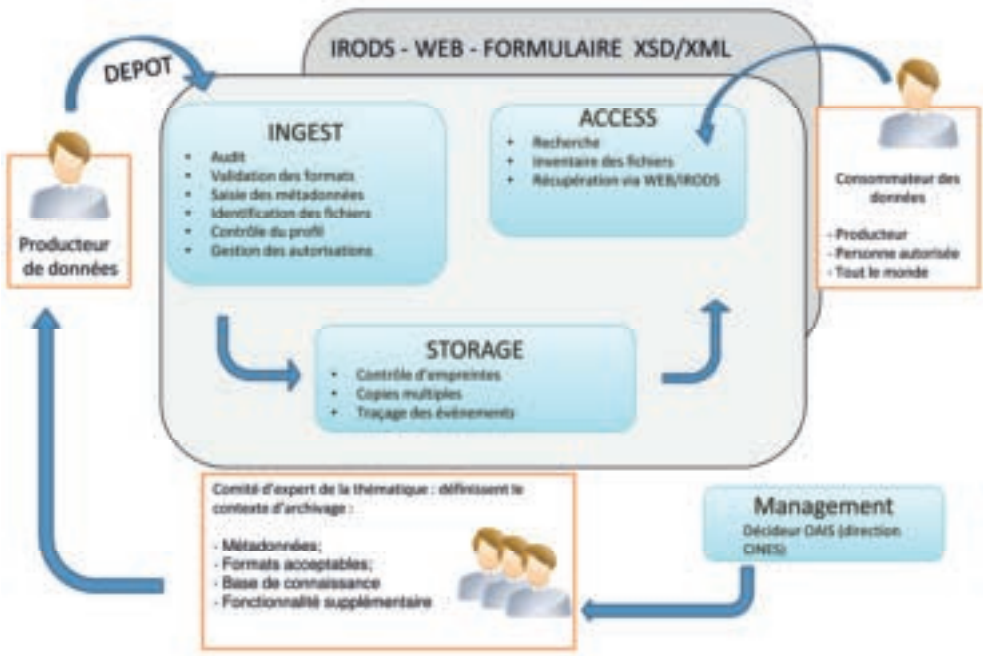
2 Ce qui représente, en mai 2013, 5 000 € TTC/To utile archivé.

3 Bibliothèque scientifique numérique, section Archivage pérenne. Pour plus d'informations : http://www.bibliothequescientifiquenumerique.fr/?BSN-6-Archivage-perenne

L'archivage intermédiaire des données scientifiques : le projet-pilote ISAAC

Étant l'un des plus grands centres de calcul intensif français et européen, les équipes du CINES sont au cœur des préoccupations de la communauté scientifique quant au devenir de leurs données : besoins de réutilisation et d'annotation des données, d'échanges entre équipes scientifiques ou encore importance des coûts de production des données. Fort de ses relations avec les membres de la communauté ESR et de l'expérience acquise avec PAC depuis dix ans, le CINES a pu saisir l'importance, la complexité et l'hétérogénéité de leurs besoins ainsi que la difficulté pour les enseignants-chercheurs de documenter l'information scientifique et technique produite. C'est donc à partir de ce constat que le CINES a lancé il y a deux ans une solution d'archivage intermédiaire recentrée sur les données de la recherche : le projet ISAAC (Information scientifique archivée au CINES). Au-delà des fonctionnalités « classiques » d'un système d'archivage électronique (métadonnées, identifiant unique et pérenne, contrôles d'accès, etc.), ISAAC définit une véritable organisation pour l'archivage des données de la recherche. La création de comités thématiques d'archivage (CTA), sur le même modèle que les comités thématiques pour le calcul intensif, permet d'encadrer l'archivage des données en définissant des règles par communauté scientifique ou sous-ensemble. Ces règles portent sur l'étude et le choix des formats de fichiers acceptés pour l'archivage, sur les métadonnées à utiliser ou encore sur la sélection des projets d'archivage.

Du point de vue technique, l'enjeu est d'offrir un système paramétrable, capable de s'adapter à de grandes volumétries et de gérer les nombreux accès, notamment le partage des données. Le choix de la technologie « *open source iRods* » correspond à cette volonté de performance pour la gestion de gros volumes de données réparties, tout en bénéficiant d'une interface Web conviviale et simple d'utilisation. Les laboratoires n'ayant pas toujours les moyens techniques et humains pour générer eux-mêmes des fichiers de description des données en XML, l'interface encapsule un système de génération automatique de formulaires à partir de schémas XML.



Les fonctionnalités de la plateforme ISAAC © CINES

En moyenne, la durée de conservation des données est de 3 à 5 ans avant de décider de les archiver sur le long terme dans PAC (éventuellement après enrichissement de la description) ou de les supprimer si elles n'ont plus d'intérêt.

Depuis le début de l'année 2013, le projet ISAAC est en phase de tests en partenariat avec le Complexe de recherche interprofessionnel en aérothermochimie (CORIA) sur des jeux de données de simulation du brûleur semi-industriel PRECCINSTA. D'autres collaborations devraient permettre de valider les développements réalisés.

Conclusion : vers une mutualisation européenne…

Depuis plusieurs années, la recherche scientifique s'oriente vers une structuration des communautés scientifiques à l'échelle européenne. Cette évolution facilite le partage des données tout en permettant des économies d'échelle par la mutualisation des outils et des compétences notamment. Dans cette optique, penser l'archivage électronique non seulement à l'échelle nationale mais aussi au niveau européen devient envisageable : c'est le projet européen EUDAT. L'objectif est de mettre en place une grille européenne de préservation des données scientifiques dans laquelle les grands centres de calcul européens tels que le CSC en Finlande, le CINES en France, ou le RZG du *Max Planck Institute* en Allemagne, se positionnent en tant que fournisseurs d'infrastructures. Cinq communautés scientifiques ont été désignées comme pilotes pour ce projet : CLARIN (linguistique), ENES (climatologie), EPOS (sismologie), *LifeWatch* (écologie) et VPH (physiologie humaine). Après un an et demi d'analyse des besoins, de définition de l'architecture et de mise en place, les premiers services commencent à entrer en phase de production. Le « déluge » de données numériques (*Big Data*) peut arriver : nous sommes prêts… !

Lorène BÉCHARD
Archiviste-expert en archivage électronique
CINES