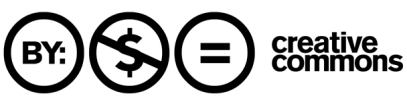
Sélectionner les données pour l'archivage : réflexions et pratiques au sein de la plate-forme PAC du CINES

Alexia de Casanove



Sélectionner les données pour l'archivage : réflexions et pratiques au sein de la plate-forme PAC du CINES

Alexia de CASANOVE

Introduction

Le Centre informatique national de l'Enseignement supérieur (CINES) est un établissement public créé en 1999 comme successeur du Centre national universitaire sud de calcul (CNUSC, 1980). Placé sous la tutelle du ministère de l'Enseignement supérieur et de la Recherche (MESR), il a trois missions principales : le calcul intensif, mission historique du centre, l'archivage pérenne de données électroniques et l'hébergement de matériel informatique à vocation nationale[1].

C'est en 2004 et sur mandat du MESR que débutent les réflexions sur l'archivage électronique, réflexions qui aboutissent en 2006 au développement d'une première version de la plate-forme PAC. Depuis, le CINES propose des solutions de conservation pour les données et documents sur support numérique issus de la communauté de l'Enseignement supérieur et de la Recherche.

Dès l'origine, les questions liées à la sélection de ces données et documents ont été intégrées dans les réflexions et les réponses apportées se sont enrichies au fil du temps et des évolutions de la plate-forme.

Le propos ici est de présenter le cycle de vie d'une donnée dans la plate-forme PAC, de sa sélection (quels critères ?) à son élimination (quelle méthode ?), en consacrant un point à sa gestion dans le temps, plus spécifiquement à la manière de décrire les durées de conservation et le sort final dans les différents bordereaux de versement utilisés.

[1] Pour plus d'informations : http://www.cines.fr/

Sélectionner les données

Avant-projet et sélection des données en entrée

La sélection des données en amont du projet respecte trois principes, chacun d'égale importance :

- identification des objets à archiver. La première étape consiste à identifier le type de données que le service versant souhaite conserver : thèses, mémoires, documents patrimoniaux numérisés, photographies, données administratives, etc., avec une préférence pour les données brutes. Dans un fonds composé de PDF bruts et de leurs OCR, on choisira généralement de ne conserver que la donnée brute (PDF). En effet, l'évolution technologique de ces dernières décennies peut laisser penser que les futurs traitements automatiques effectués sur le PDF seront plus performants, de meilleure qualité et moins coûteux que l'archivage des résultats actuellement produits (OCR), d'où cette politique de sélection. Autre question essentielle : quelle sera la « granularité » du paquet, c'est-à-dire quelle unité d'archivage choisir ? Par exemple, pour un fonds composés d'ouvrages numérisés, le paquet sera-t-il le titre (donc un ou plusieurs tomes) ? Le tome ? Le chapitre ? Ce choix de l'unité d'archivage doit être le plus pertinent possible et dépend de plusieurs facteurs. On peut citer entre autres : la qualité des métadonnées ; les besoins du service versant, en termes de recherche notamment ; la préservation de l'unité intellectuelle ; les questions de durée d'utilité administrative (DUA) dans le cadre de données publiques, etc. ;

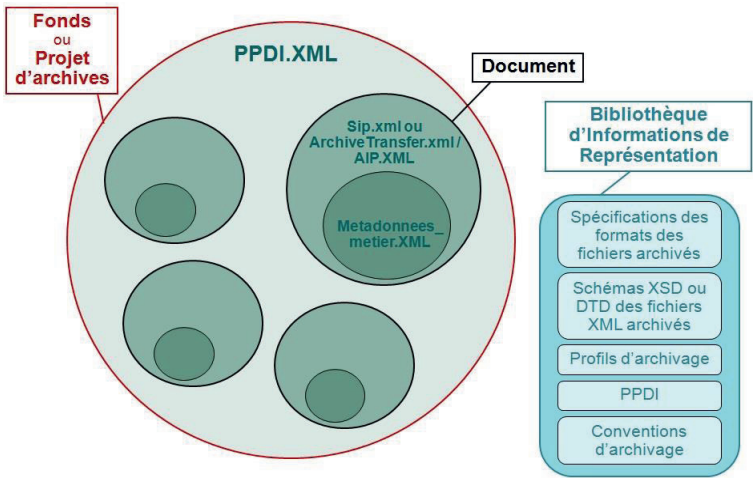
- identification des formats d'archivage. Ce paramètre doit impérativement être pris en compte dans un contexte numérique. Pour être archivable, un format doit être exploitable sur une durée indéterminée. C'est pourquoi le CINES privilégie les formats publiés (WAVE, SVG, etc.), largement utilisés (XML, MPEG4, etc.), et si possible normalisés, comme le PDF (ISO 32000-1:2008) ou le PNG (ISO 15948:2004). Les spécifications de ces formats doivent en outre être ouvertes, spécifications que les fichiers doivent respecter à la lettre. Ces conditions sont nécessaires pour permettre l'identification d'un fichier et en contrôler la qualité avant archivage, pour sa lecture et sa compréhension dans le temps, mais aussi en cas de migration (transformation vers un autre format en cas d'obsolescence par exemple). Une liste restreinte de formats respectant ces critères est acceptée, et ce afin d'en avoir une gestion plus aisée[1]. Elle reste cependant ouverte pour répondre aux besoins de nouveaux services versants ;

[1] http://facile.cines.fr/

▪ identification du jeu de métadonnées. Les métadonnées sont nécessaires pour décrire le document et l'identifier. Sans elles, ce dernier peut se révéler inexploitable, d'où l'importance de réaliser un bon *mapping* entre les métadonnées « métier » du service versant (par exemple du TEF[1]) et celles utilisées dans le bordereau de versement. Le *mapping* consiste à sélectionner, parmi les métadonnées métier, les plus pertinentes pour la compréhension du paquet, puis à les faire correspondre aux métadonnées requises par le CINES. Le bordereau de versement du CINES (sip.xml) est volontairement générique pour être compatible avec tout projet d'archives. Le choix s'est ainsi porté sur le format Dublin Core, enrichi de quelques métadonnées techniques et de gestion propres à PAC. Un bordereau SEDA (archiveTransfer.xml) a aussi été mis en place pour traiter les archives publiques. Cependant, conscient du caractère générique du sip.xml et de la spécificité de chaque service versant, ceux-ci sont encouragés à archiver aussi leurs métadonnées « métier », car ce sont elles qui renferment la description la plus précise et la plus précieuse pour la compréhension des données conservées.

Les différents niveaux de métadonnées dans PAC

Un fonds se présente ainsi :



Les niveaux de métadonnées dans PAC © CINES

[1] TEF : Thèse Électronique Française. Ce jeu de métadonnées est employé pour les thèses électroniques soutenues en France (http://www.abes.fr/abes/documents/tef/).

Il contient plusieurs documents (les paquets) décrits chacun dans un bordereau de versement sip.xml ou archiveTransfer.xml et qui peuvent être accompagnés de métadonnées « métier ». Le fonds est présenté dans le ppdi.xml (Project Preservation Description Information), fichier qui détaille l'historique du fonds, le service versant, le producteur… et est conservé dans la Bibliothèque d'Informations de Représentation (BIR). Elle réunit quant à elle toutes les informations nécessaires à la compréhension du fonds mais aussi à la lisibilité des fichiers : l'ensemble des spécifications relatives à chaque format archivé, les profils d'archivage dans le cas du SEDA, les conventions etc.

La structure du document à archiver

Un paquet d'archives se décompose ainsi :

- la description de l'archive, avec deux possibilités :
  - fichier sip.xml (format « maison » basé sur du Dublin Core) ;
  - fichier archiveTransfer.xml (issu du SEDA 1.0 ; contexte des archives publiques ; description à plusieurs niveaux) ;
- le dossier contenant les documents électroniques à archiver : répertoire « DEPOT »
  - une sous-arborescence est autorisée ;
  - tout fichier présent doit être décrit dans le fichier sip.xml ou archiveTransfer.xml ;
  - si l'archivage est au format PAC : un sous-répertoire DESC facultatif peut être utilisé pour joindre le(s) fichier(s) de métadonnées métier.

Un exemple : l'archivage des données satellitaires Géosud

L'exemple qui suit illustre la mise en application des critères de sélection. Il arrive qu'ils subissent quelques infléchissements suivant les spécificités des fonds présentés, afin de répondre au mieux à la demande, tout en essayant de rester le plus possible proche des exigences définies.

Géosud est un Équipex[1] dont l'objectif est de fournir un accès pérenne à l'information spatiale sur les écosystèmes et les territoires par acquisition annuelle d'images satellitaires haute résolution. L'intérêt de ce projet est de pouvoir suivre année après année l'évolution des territoires.

[1] Équipex : « Équipement d'excellence ». Les Équipex font partie du programme « investissements d'avenir » mené par l'Agence nationale de la Recherche (ANR) et le Commissariat général à l'investissement.

Le fonds à archiver est composé d'images satellitaires brutes et orthorectifiées[1], couverture annuelle nationale et régionale de la France. Le niveau de service proposé est un archivage pérenne, avec conservation sur le long terme. Les métadonnées à verser sont formalisées dans le bordereau sip.xml, donc en Dublin Core, et accompagnées des métadonnées métier conformes à la norme ISO 19115 (norme sur les métadonnées géographiques). Quant aux formats de fichiers, il s'agit du GeoTIFF, du JPEG2000, du PDF et du XML.

La mise en place de ce projet n'a pas été sans complication. À l'origine, seules les données orthorectifiées, et non les brutes, devaient être archivées, décision en contradiction avec les propos tenus précédemment sur la préférence accordée aux données brutes. Pour quelle raison ? L'orthorectification est obtenue suite à des traitements coûteux que les producteurs n'envisagent pas de reproduire régulièrement. De plus, cette modification apporte une vraie plus-value à l'image, qu'il serait donc regrettable de ne pas conserver. Dans la majorité des cas, la donnée brute présente un moindre intérêt. Le problème est que le GeoTIFF orthorectifié proposé n'est pas conforme à la norme GeoTIFF (dans le détail, le système de projection Lambert 93 utilisé pour les images orthorectifiées est absent de la norme officielle). Le CINES ne peut s'engager à assurer le même niveau de service sur une donnée au format non conforme, notamment en cas de migration. Donc, étant donné qu'il était nécessaire d'archiver les images traitées pour les raisons exposées, décision fut prise de conserver également les images brutes, conformes à leur norme, à partir desquelles l'orthorectification pourra être appliquée en cas de problème.

Gérer les durées de conservation (données publiques)

Contexte de réflexion

Historiquement, la plate-forme PAC était destinée à l'archivage définitif de thèses (programme STAR en partenariat avec l'ABES) et de revues numérisées (programme Persée). La question des durées de conservation n'était donc pas une priorité et était traitée par une métadonnée, <dureeConservation>, au

[1] L'orthorectification est un ensemble de traitements pour rectifier géométriquement et égaliser radiométriquement les images.

contenu quelque peu artificiel puisqu'elle proposait par défaut « P10000Y » (10 000 ans). Par la suite, de premiers projets d'archivage de données publiques scientifiques furent menés. Leur sort final étant la conservation définitive et s'agissant de données scientifiques pour lesquelles la pertinence d'une durée d'archivage intermédiaire peut se poser, dans un premier temps la même métadonnée que pour les fonds non publics leur fut appliquée, sans mention particulière de DUA.

En 2012 la Cour des Comptes entame avec le CINES un projet pilote d'archivage numérique des données produites par les juridictions financières. Données administratives publiques, leur gestion nécessite de traiter de manière idoine la question des durées de conservation et particulièrement celle des DUA. De plus, en période de renouvellement de l'agrément de tiers-archivage électronique de données publiques délivré par le Service interministériel des Archives de France, il devenait impératif pour le CINES de trouver une solution plus appropriée que celle employée jusqu'alors, condition essentielle à sa prolongation.

Mise en œuvre

Deux changements ont eu lieu afin de répondre au mieux à ces problématiques :

- implémentation intégrale du SEDA dans sa version 1.0, avec le développement d'un profil propre au CINES (présence de quelques restrictions supplémentaires spécifiques à l'implémentation CINES) ;
- modification du bordereau de métadonnées « maison » (ou bordereau PAC), afin qu'il puisse être également utilisé dans le cadre de données publiques (type archives photographiques, pour lesquelles une description en Dublin Core est plus pertinente). Les nouvelles métadonnées ajoutées sont directement inspirées du SEDA.

Les DUA sont donc ainsi renseignées :

- dans le bordereau archiveTransfer.xml du SEDA, un bloc <AppraisalRule> constitué des trois métadonnées <Code>, qui indique le sort final, <Duration> pour la durée de la DUA et <StartDate> qui précise la date à partir de laquelle se calcule cette DUA ;

```
<AppraisalRule>
    <Code listVersionID="edition 2009">detruire</Code>
    <Duration>P15Y</Duration>
    <StartDate>2012-09-07</StartDate>
</AppraisalRule>
```

Gestion des DUA : bordereau SEDA © CINES

- dans le bordereau PAC, un bloc semblable au précédent, dont seuls les intitulés changent.

```
<evaluation>
    <DUA>P15Y</DUA>
    <traitement language="fra">conservation définitive</traitement>
    <dateDebut>2013-10-09</dateDebut>
</evaluation>
```

Gestion des DUA : bordereau CINES © CINES

Deux contenus sont possibles pour <code> et <traitement>, soit « conservation définitive », soit « détruire ».

Notons que les DUA exploitées par l'administration de PAC sont celles situées au niveau du paquet et non au niveau des fichiers. Un contrôle strict des dates renseignées dans le bordereau SEDA est effectué, et ce afin de s'assurer qu'aucune contradiction n'est présente entre différents niveaux. Chaque date de niveau inférieur doit être cohérente avec celle de niveau supérieur et ainsi de suite.

À l'expiration de la DUA, une alerte est remontée. Il ne reste donc « qu'à » appliquer le sort final. Mais sa mise en œuvre se révèle plus complexe dans le cadre numérique, particulièrement s'il faut éliminer la donnée.

Éliminer les données

Contexte de réflexion

En fin de DUA, comment procéder à l'élimination d'une archive numérique publique quand le sort final est « détruire » ? Si l'élimination d'un document papier est assez aisée, avec destruction physique par broyage ou incinération, il est difficile d'appliquer des méthodes identiques dans le contexte numérique. La donnée ne peut être simplement jetée à la corbeille.

En 2012, le CINES est confronté à de nouveaux besoins. Dans un premier temps, dans le cadre de l'archivage des thèses électroniques, une demande d'élimination pour plagiat émane de l'ABES. Ensuite, l'archivage de données publiques intermédiaires telles que celles de la Cour des Comptes implique de prendre en compte les fins de DUA et le sort final des archives. Enfin, il était nécessaire d'être en accord avec les exigences du Service interministériel des Archives de France en vue d'obtenir le renouvellement de l'agrément de tiers-archivage électronique, une des conditions essentielles à ce renouvellement étant la mise en place d'une procédure d'élimination, en conformité avec les normes en vigueur, notamment la NF Z42-013.

Ce que disent les normes

- NF Z42-013 : « les archives ayant atteint la durée maximale de conservation doivent être supprimées sous la responsabilité d'un opérateur habilité et dans le respect des normes en vigueur. Cette opération a pour effet de rendre totalement et définitivement inaccessibles les documents éliminés ».

- GA Z42-019 : « la destruction doit être complète et irréversible et ce caractère définitif doit être contrôlable […] Au sein d'un SAE, la destruction doit pouvoir être effectuée au niveau unitaire, archive par archive. La destruction d'une archive (document ou lot de documents) dans un SAE suppose l'effacement ou la disparition des fichiers de ce document ».

Suivent ensuite des propositions de méthodes : selon le type de support utilisé (WORM physiques ou logiques) il faudrait soit le détruire physiquement, soit effacer la donnée par l'application de séquences d'écriture successives.

Problèmes rencontrés et solutions adoptées

La mise en pratique de ces recommandations pose quelques difficultés. D'un point de vue financier d'abord, détruire les supports à chaque élimination de données n'est pas toujours envisageable. Cette solution n'est pas économiquement viable. D'un point de vue technique, si la réécriture peut être efficace sur disque, elle l'est moins sur bande. De plus, l'effacement d'une zone précise est délicat : comment être sûr que la réécriture s'effectue au bon endroit et que d'autres données ne sont pas affectées ?

Épineux problème que celui de concilier les exigences normatives et la réalité technique ! Ces écueils identifiés, et après l'organisation d'une concertation avec le Service interministériel des Archives de France et d'autres acteurs français confrontés à cette même problématique, le CINES a proposé et opté pour un déréférencement de la donnée dans un premier temps, la rendant ainsi inaccessible, dans l'attente de la fin de vie du support, avant effacement complet par séquences de réécritures successives puis destruction physique.

L'opération de déréférencement, exécutée manuellement pour le moment, n'a jusqu'ici été utilisée que trois fois pour « supprimer » des thèses. Elle n'a pas encore été testée sur de plus gros volumes de données : les archives publiques conservées dans PAC n'étant pas encore arrivées en fin de DUA, l'impératif d'une élimination simultanée de grandes quantités de données ne s'est pas encore fait sentir.

Ces quelques années à venir accordent un délai suffisant pour tenter de développer une technique de déréférencement automatique de volumes plus conséquents, voire pour qu'un éventuel changement technologique apporte une réponse plus satisfaisante, en permettant de réaliser un effacement précis et réel des données, sans avoir à attendre la fin de vie du support.

Alexia de CASANOVE
Archiviste
CINES
casanove@cines.fr