L'archivage pérenne des données scientifiques en SHS

Laurence Rageot, Richard Walter

Du bon usage d'Adonis

L'archivage pérenne des données scientifiques en SHS

Aujourd'hui, les différentes disciplines des sciences humaines et sociales, à travers leurs objets et problématiques spécifiques, produisent massivement des données numériques. Les outils numériques font désormais partie intégrante de l'environnement de travail. Si ces instruments offrent de nouvelles perspectives de mise en réseau et d'exploitation des données, ils génèrent également de nombreuses incertitudes quant à leur sauvegarde sur le long terme et aux moyens de faire face aux adaptations constantes engendrées par l'évolution, la prolifération et l'obsolescence rapide des technologies informatiques.
Simultanément à la croissance quasiment exponentielle des données numériques, les modifications du monde de la recherche ont profondément changé le rythme de vie des projets de recherche. Ainsi, les données de la recherche sont confrontées à une transition inéluctable et massive vers le numérique mais aussi au risque de perte de données, si l'on ne structure pas correctement l'information à archiver. L'archivage des données numériques devient donc un enjeu primordial pour la transmission des savoirs.
L'archivage à long terme des documents électroniques permet de conserver le document et l'information qu'il contient dans son aspect physique comme dans son aspect intellectuel sur le très long terme, soit 30 ans et au-delà, de manière à pouvoir le rendre accessible et compréhensible. Il doit faire en sorte que le document reste compréhensible par ses utilisateurs potentiels à travers le temps. Il ne s'agit pas d'une simple sauvegarde d'un fichier sur un disque dur quelconque qui risque de ne plus fonctionner dans quelques années. Aujourd'hui, la plupart des fichiers informatiques de plus de 10 ans sont illisibles à cause de la perte d'intelligibilité ou du format des fichiers, d'un support physique détérioré ou d'un logiciel ou matériel de lecture disparu.
L'archivage à long terme d'un document numérique a donc trois objectifs principaux : conserver le document, lui garantir son accessibilité et en préserver l'intelligibilité.
Pour répondre à cette problématique, le TGE Adonis a lancé un projet d'archivage à long terme des données numériques de la recherche en SHS. Pour cela, il s'est associé au CINES (Centre informatique national de l'Enseignement supérieur) afin d'offrir à la communauté SHS une infrastructure souple et fiable de pérennisation et d'accès des données produites par la recherche.

Historique et acteurs du projet d'archivage à long terme du TGE Adonis

Dès 2008, le TGE Adonis et le CINES ont mutualisé leurs compétences et lancé un projet pilote sur l'archivage à long terme des données orales. Il a été mené avec la collaboration du CRDO (Centre de ressources pour la description de l'oral), des laboratoires LPL (Laboratoire parole et langage), Lacito (Langues et civilisations à tradition orale) et RISC (Relais d'informations sur les sciences de la cognition). Cette initiative a permis de mettre en place une nouvelle infrastructure mutualisable et pérenne et de définir les responsabilités de chacun (producteur, service versant et service d'archives) autour des différentes actions d'un projet d'archivage que sont la préparation et le versement des données à archiver, l'archivage des données, la gestion de l'évolution des formats des données archivées et l'accès aux données. Ce projet pilote est désormais terminé et l'archivage est effectif depuis septembre 2010.
Le CINES est l'opérateur qui, pour le compte du TGE Adonis, assure la conservation, la pérennisation et les migrations futures. Il contrôle en entrée les versements, et en particulier, la conformité des fichiers à des formats de conservation définis à l'avance.

Le TGE Adonis est l'interlocuteur des laboratoires et des structures demandant à bénéficier de l'archivage à long terme. Il remplit le rôle de service versant et coordonne les projets de conservation à long terme ouverts au CINES. De plus, le TGE Adonis assurera prochainement l'accès aux données, ce qui favorisera leur diffusion au sein de la communauté de la recherche en SHS.

Le fonctionnement type d'un projet d'archivage à long terme

Il convient de faire la différence entre producteur de données, service versant (TGE) et service d'archives (CINES). Le producteur dépose les données dans une « collection » du service versant. La notion de collection permet de cloisonner les dépôts ; un producteur peut déposer plusieurs collections différentes. Le service versant supervise et assure le contrôle qualité. Le service d'archives effectue une validation immédiate du document archivé (format, métadonnées, etc.).
Le projet d'archivage implique le respect d'une procédure précise. En amont, les trois parties (producteur, TGE Adonis et CINES) doivent se mettre d'accord sur le plan d'archivage et la structure de la collection. Pour cela, le corpus à archiver doit être « clos », mais l'archivage numérique permet le dépôt de versions successives de celui-ci, toutes horodatées. Le producteur doit effectuer une sélection intelligente des documents à archiver : ce document doit-il être archivé et pourquoi ? Cela nécessite une stratégie pour éliminer les données redondantes. Il faut de préférence choisir des documents « bruts » : il n'est pas nécessaire d'archiver des documents qui peuvent être reconstitués à partir de données déjà présentes dans le système d'archivage.
La discussion porte également sur les formats acceptés et les métadonnées. En effet, cet archivage à long terme n'est basé que sur un certain nombre de formats dont l'accessibilité est garantie. Pour qu'un format soit archivable, il doit être exploitable dans son intégralité sur une durée indéterminée et posséder une spécification accessible qui décrit ses caractéristiques. Le CINES garanti l'accessibilité d'un certain nombre de formats dont la liste est disponible sur son site. Cette liste évolue en fonction des demandes et de l'évolution des formats et des usages numériques. Un producteur d'archive peut effectuer une demande de prise en compte de nouveaux formats. Le CINES effectuera une étude de faisabilité en interrogeant les spécialistes du format ou du domaine concerné et fournira une

réponse argumentée.
Si le format retenu lors de l'archivage devient obsolète, le CINES fera une demande d'autorisation de changement de format. En cas d'évolution des supports de stockage, le CINES avise le producteur d'un changement de support.

Les métadonnées sont la carte d'identité d'un document. Elles permettent de l'identifier, de le décrire, d'expliquer l'origine de sa création, son utilité et ses destinataires. Sans ces éléments, un document peut vite devenir incompréhensible et donc inexploitable. Le projet d'archivage demande de manipuler deux types de métadonnées : celles spécifiques à l'archivage et celles propres au domaine du producteur (métadonnées dites « métiers »). La liste de ces dernières est actée par le producteur, le TGE Adonis et le CINES. Chaque ressource à archiver est alors accompagnée d'un fichier décrivant les métadonnées et la structure du fichier à archiver.
À chaque dépôt, le CINES vérifie la conformité des fichiers à archiver et ajoute les métadonnées d'archivage (identifiant unique et pérenne dans la plateforme d'archivage du CINES, date d'archivage, empreinte numérique des fichiers et identifiant du projet d'archives). Si le fichier est conforme, il est archivé et le CINES envoie au TGE Adonis un certificat d'archivage comportant l'identifiant unique et pérenne de l'archive transférée ainsi que ses métadonnées.

Les données et les métadonnées sont aussi transférées aux serveurs de diffusion du TGE Adonis, hébergés au Centre de calcul de l'IN2P3, où elles seront consultables par l'intermédiaire de simples adresses URL qui pourront être intégrées dans un site internet. Éventuellement, de nouveaux formats de fichiers peuvent être ajoutés afin de faciliter la consultation. Par exemple, pour chaque fichier "son" conservé, peuvent être fabriqués un fichier MP3 et un fichier wav dégradé, pour chaque fichier " image", un fichier JPG.
Un système d'authentification et de restriction d'accès aux documents (avec login et mot de passe) est en cours de développement par le TGE Adonis. Sur la plateforme d'archivage à long terme, toutes les données seront par défaut accessibles à tout le monde. Il sera possible cependant de définir des règles de restriction d'accès pour une collection complète ou partie de cette collection. Ces règles devront être discutées en amont du projet avec les responsables du TGE Adonis et en conformité avec les règles de diffusion définies par le Service interministériel des Archives de France.

Le TGE met à disposition des producteurs de données différents outils de gestion pour les demandes d'amélioration ou de prise en charge de nouveaux formats, les statistiques d'utilisation, etc. Par ailleurs, pour aider à la réalisation d'un projet d'archivage des données sonores et visuelles, un guide méthodologique a été réalisé, en collaboration avec le CINES, sur le choix de formats numériques pérennes. D'autres guides sont en cours de réalisation (sur le format PDF) ou à l'étude (les formats images). Enfin, le CINES met à disposition un outil permettant de vérifier l'éligibilité des formats à l'archivage à long terme sur sa plateforme : FACILE – validation du Format d'Archivage du CInes par anaLyse et Expertise.

Les conditions et procédures de soumission d'un projet d'archivage à long terme au TGE Adonis

Les producteurs de données doivent remplir certaines conditions essentielles, parmi lesquelles :
▶ appartenir à un organisme français dont la mission est liée à l'Enseignement supérieur et/ou à la recherche en SHS ;
▶ collecter et/ou produire une collection d'objets numériques présentant un intérêt scientifique ou pédagogique ;
▶ définir une stratégie et caractériser le besoin d'un archivage patrimonial (définitif) pour cette collection d'objets numériques ;
▶ mettre en place un plan d'archivage et une sélection intelligente des données ;
▶ vérifier le contexte légal de production de ces objets numériques : droits de propriété intellectuelle et droits d'auteur sur les documents à archiver ;
▶ produire des métadonnées décrivant la collection d'objets numériques auxquelles s'ajouteront les informations produites par les CINES lors de l'archivage ;
▶ choisir des formats de fichiers éligibles à l'archivage (le CINES et le TGE peuvent être sollicités pour étudier tout nouveau format de fichier non pris en charge actuellement sur leur plateforme) ;
▶ disposer de moyens humains pour la gestion et/ou la participation au projet d'archivage électronique.
L'organisme souhaitant initier un tel projet effectue une demande dans ce sens qui sera examinée par la direction du TGE Adonis, en collaboration avec le CINES. Si l'organisme est d'accord avec le fonctionnement du projet d'archives, celui-ci pourra se mettre en place. Pour tout autre mode de production, l'organisme devra prendre directement contact avec le CINES.

Les données numériques ne sont pas d'emblée pérennes. Il ne suffit pas de les stocker à un endroit pour se dire qu'elles seront accessibles dans la durée. Un projet d'archivage à long terme a pour but de maintenir l'intégrité, l'authenticité et l'intelligibilité des objets numériques produits par une communauté scientifique donnée, de permettre l'accès et de garantir le cas échéant la confidentialité de ces données.
L'archivage numérique à long terme est une opération spécifique qui nécessite gestion, surveillance, renouvellement des supports d'enregistrement mais aussi absence de formats propriétaires et bon codage initial des données. Le TGE Adonis, en partenariat avec le CINES, a mis en place une infrastructure pour assurer cet archivage pour la communauté SHS, avec une adaptation aux métadonnées métier des producteurs et aux contraintes d'accès spécifiques.
La capacité de stockage des archives augmente continuellement et rapidement. Mais il est primordial de veiller à ce que les données soient archivées selon des formats et des descriptions pérennes. L'usage de ces bonnes pratiques doit être largement diffusé au sein de la communauté SHS. D'où l'importance des guides méthodologiques ou des actions de formation proposés par le TGE Adonis.

Richard Walter & **Laurence Rageot**, TGE Adonis

contact&info
▶ Richard Walter
Directeur-adjoint TGE Adonis
richard.walter@tge-adonis.fr
▶ Laurence Rageot
Responsable du pôle Digital Humanities
laurence.rageot@tge-adonis.fr
▶ Pour en savoir plus
www.tge-adonis.fr