La préservation à long terme des données de la recherche en Sciences Humaines et Sociales : un retour d'expérience

Nicolas Larrousse
TGIR Huma-Num / UMS 3598
190-198 avenue de France
75648 PARIS CEDEX 13

Marion Massol
CINES (Centre Informatique National de l'Enseignement Supérieur)
950 rue de Saint-Priest
34097 MONTPELLIER CEDEX 5

Résumé

Peu de structures ont aujourd'hui l'expérience et le recul sur ce type nouveau de préoccupation qu'est l'archivage de données numériques sur le long terme. Cette activité très particulière doit en effet se projeter sur des temporalités bien différentes du cycle de vie actuel, très court, des évolutions dans le domaine du numérique en général. La pérennisation de données sur le long terme ne saurait donc se limiter à choisir une « bonne » technologie de stockage, il s'agit plutôt de mettre en œuvre un ensemble de technologies associées à des procédures adaptées mais aussi d'acquérir et de développer des compétences nouvelles, comme par exemple une expertise sur les formats et leur évolution. Nous exposerons la problématique et les solutions qui sont apportées aujourd'hui à ce véritable besoin en nous basant sur notre expérience de la préservation de données en Sciences Humaines et Sociales. Le sujet sera abordé selon différents angles : l'infrastructure technique à mettre en place, l'organisation à prévoir basée sur les recommandations du modèle OAIS (Open Archival Information System), et également la prise en compte des aspects réglementaires issus du monde des archives. Nous terminerons par une mise en perspective sur la réflexion autour des besoins en préservation de données numériques aujourd'hui.

Mots-clefs

Préservation à long terme, Sciences Humaines et Sociales, Formats, Archives, Labellisation, Digital Management Plan, Plan de gestion des données

1 Contexte général

Les données de la recherche en SHS (Sciences Humaines et Sociales), comme pour nombre d'autres disciplines, sont produites nativement sous forme numérique ou proviennent de la numérisation de données analogiques. Le passage au numérique apporte un gain évident pour la diffusion et les traitements qui peuvent être effectués sur ces données. Cependant, contrairement à ce qu'il est communément admis, la possibilité de dupliquer facilement ces informations ne les rendent pas éternelles. En effet, un objet numérique peut être paradoxalement considéré comme plus fragile que son homologue du monde réel, essentiellement par le fait que l'information représentée est dissociée du support utilisé.

Aujourd'hui, les SHS prennent massivement le tournant du numérique, constituant un mouvement fréquemment désigné par le terme d'« Humanités Numériques ». Le volume de données numériques issues de la recherche en SHS a ainsi considérablement augmenté au cours de ces dernières années. Si l'on ne s'en préoccupe pas, ces données souvent très coûteuses à produire et à gérer sont menacées à moyen terme de disparition. Les agences de financement ne se souciaient guère du devenir de ces données jusqu'alors, mais certaines commencent à prendre conscience de cette question. On peut citer comme exemple récent, l'Union européenne, qui exige l'écriture d'un plan de gestion des données (Data Management Plan) pour les réponses aux appels à projets lancés dans le cadre du programme « Horizon2020 ».

2 Quelques spécificités des données en Sciences Humaines et Sociales

Si on les compare à celles des sciences dites « dures », les données manipulées en SHS sont généralement très hétérogènes. Elles comportent des objets aussi différents que des textes, des images, des enregistrements sonores, des cartes associées aujourd'hui à des SIG (Systèmes d'Information Géographique), des vidéos, des modèles en 3D sans oublier les données physiologiques. Cet inventaire « à la Prévert », qui est pourtant loin d'être exhaustif, montre la diversité et le foisonnement du matériel scientifique qu'il est nécessaire de préserver.

La pérennisation, l'interopérabilité et la diffusion des données numériques requièrent l'adoption de formats ouverts, spécifiés, largement utilisés, si possible normalisés. Si des formats « pivots » se sont imposés pour la production de certaines données, comme par exemple le format Tiff pour les images, les SHS continuent à recourir souvent à des formats « métiers » ou même propriétaires qui rendent leur préservation plus complexe.

Contrairement à d'autres domaines scientifiques tel que le spatial, la culture de la standardisation, et l'adoption de « bons » formats qui en découle, est encore marginale dans les SHS, même si l'on note une nette amélioration des pratiques ces dernières années.

Une autre caractéristique des SHS réside dans le fait que ces données sont potentiellement réutilisables par d'autres disciplines avec d'autres modalités. Une carte pourra être utilisée de manière très différente par un historien ou un géographe de même qu'un enregistrement sonore par un musicien ou un linguiste. Pour que cette réutilisation soit possible, il sera bien sûr nécessaire de préserver la donnée mais également d'être à même de savoir qu'elle existe et de la retrouver. Ce dernier aspect implique que l'on ait correctement effectué la description de la donnée dans son contexte de production et que l'on soit capable de maintenir des référentiels (ou catalogues) de description sur le long terme. En effet, une donnée produite trente ans auparavant peut parfois se révéler très utile pour la compréhension d'un phénomène actuel.

Enfin, ces données ont parfois un caractère patrimonial, qui ne fait que renforcer encore la nécessité de les préserver. On peut citer l'exemple d'enregistrements de langues dont le dernier locuteur a disparu ou encore la numérisation de manuscrits trop fragiles pour être consultables aujourd'hui. Les événements récents démontrent l'intérêt de préserver la modélisation en trois dimensions de constructions qui peuvent disparaître. Ce dernier cas est typique de la problématique de l'archéologie qui souvent détruit l'objet qu'elle étudie.

Cet aspect patrimonial nous rappelle que les données de la recherche sont aussi, en tant que « productions d'agents de l'état dans le cadre de leurs missions », des archives dites « publiques ». Cela signifie qu'il faudra tenir compte des aspects réglementaires qui régissent les archives, en particulier du code du patrimoine, lors de la mise en œuvre du processus de préservation de ces données.

Aujourd'hui, le statut des données de la recherche ne fait pas l'objet d'une définition précise ni, par conséquent, d'un traitement spécifique d'un point de vue archivistique. Elles ont donc vocation, comme pour toutes les archives publiques, à être finalement prises en charge par un service d'archive compétent (par exemple les Archives nationales).

3 Les différents défis à relever

Quelle organisation doit-on mettre en place pour être capable de lire, de comprendre, de retrouver mais également de réutiliser ces données dans vingt ou trente ans ? Ces préoccupations, qui sont plus habituellement celles des archivistes et des professionnels de l'IST, prennent une tournure différente pour le cas des données de la recherche, et ce d'autant plus, qu'elles sont exprimées sous forme numérique. On ne préservera plus l'objet original à l'identique, comme on le faisait pour une archive classique, mais plutôt l'information qu'il contient.

On peut regrouper ainsi en trois grandes catégories les conditions nécessaires à la préservation des données numériques :

- Le support sur lequel sont stockées les données

Faute d'action spécifique, l'irrémédiable dégradation ou vieillissement des supports de stockage (dû à des effets chimiques, démagnétisation, etc.) va participer à la perte de l'information. Une multiplication des exemplaires sur des technologies variées ainsi qu'un contrôle régulier de l'intégrité (checksum + CRC…) avec éventuelle recopie de la donnée archivée sont donc nécessaires. Par ailleurs, l'obsolescence rapide des technologies de stockage (lecteurs, médias mais aussi logiciels de pilotages des dispositifs) impose une migration régulière de l'information numérique vers des technologies tout à la fois récentes, éprouvées et robustes.

- Le format informatique utilisé pour exprimer leur codage

Une donnée ne sera intelligible que si son codage peut être décrypté. Il existe aujourd'hui une multitude de formats de fichiers qui ne sont guère pérennes. Un suivi de l'évolution de ces formats est indispensable pour garantir la compréhension d'une donnée. Une autre facette de cette activité est d'être capable d'évaluer les risques de la conversion d'un format vers un autre en cas d'obsolescence due au manque d'outils ou même aux pratiques commerciales des éditeurs ;

- La description des données

Le meilleur moyen de perdre une donnée est d'ignorer son existence. Il sera donc nécessaire de décrire le mieux possible cette donnée afin de pouvoir constituer des répertoires de signalement, si possibles redondants et interdisciplinaires. Par ailleurs, dans l'optique de renforcer l'intelligibilité de ces données, la documentation doit également servir à rendre le contexte dans lequel elles ont été produites.

4 Les réponses apportées par le CINES (Centre Informatique National de l'Enseignement Supérieur)

Conscient de ces enjeux, le ministère de l'Enseignement supérieur et de la Recherche a missionné le CINES (Centre Informatique National de l'Enseignement Supérieur) pour la préservation des données de la recherche. L'objectif initial était la préservation des thèses. Rapidement, le dispositif mis en place a été adapté à d'autres types de données plus complexes, en particulier celles des disciplines des SHS. Le CINES, engagé dans la préservation de l'information numérique depuis plus de dix ans, offre des solutions économiques, fiables, sécurisées, mutualisées, certifiées et personnalisées pour l'archivage des données produites par la communauté Enseignement supérieur et Recherche. Le dispositif est basé sur la mise en œuvre d'un modèle conceptuel d'organisation : le modèle OAIS (Open Archival Information System – ISO 14721) qui a été élaboré par les acteurs du domaine spatial.

- Une plate-forme technologique

Le CINES a mis en place une infrastructure dédiée à la préservation qui repose sur un stockage de la donnée en plusieurs exemplaires sur différents médias (disque et bandes) dans différentes salles machines situées dans différents bâtiments, mais elle repose surtout sur une chaîne de traitements. Celle-ci assure, entre autres, le contrôle des données et des métadonnées associées, leur intégration dans le système,

l'affectation d'identifiants pérennes, une préservation de la documentation du système et des protocoles de versement négociés avec les producteurs de données, etc.

Par ailleurs, pour limiter le risque de perte de données en cas d'incident majeur sur le site principal, une réplication du système et des données archivées est assurée sur l'infrastructure du CC-IN2P3, distant de plusieurs centaines de kilomètres du CINES.

On notera également que le système PAC (Plate-forme d'Archivage du CINES) effectue très régulièrement des contrôles de la qualité de l'ensemble des données préservées, conformément aux normes et standards de la pérennisation de données numériques (contrôle de checksum des fichiers archivés, validation de la non-corruption des référentiels de catalogage, présence et disponibilité de tous les exemplaires de chaque donnée, validation des « anciennes » données et métadonnées selon les processus et référentiels « du moment », etc.).

Le CINES prend également en charge les processus de conversion de formats de fichiers, lorsque ces derniers sont déclarés obsolètes.

- Une organisation humaine

Si l'aspect technologique est important pour la préservation, le dispositif humain est lui-aussi crucial pour mettre en œuvre toutes les opérations d'archivage. En effet, le CINES prend la responsabilité de pérenniser l'information qui lui est confiée et il ne s'agit donc pas d'un stockage « classique ».

D'un point de vue pratique, il gère les données, les métadonnées, les informations sur le projet mais également d'autres niveaux d'informations comme la documentation sur les formats, les outils de contrôle et de manipulation des formats, etc.

Des compétences spécifiques sont ainsi nécessaires comme la veille technique sur les formats, mais aussi des compétences archivistiques pour prendre en compte les aspects réglementaires, et garantir l'intelligibilité et la réutilisation des données.

- La certification du dispositif

Comme nous l'avons déjà évoqué, en France, la gestion des archives est assurée par l'Etat. De ce fait, pour pouvoir effectuer l'archivage de données, même de manière temporaire, il est indispensable d'obtenir un agrément des Archives de France. Cet agrément, d'une durée limitée, est basé sur un audit général. Les normes imposées par ce service ont nécessité des aménagements spécifiques dans la salle machine ou encore l'implémentation du protocole spécifique standardisé d'échange de données entre services d'archives (SEDA - standard d'échange de données pour l'archivage).

Par ailleurs, la communauté de l'archivage numérique s'est organisée pour mettre en place une certification de systèmes d'archivage électronique internationale et indépendante. Les services du CINES ont ainsi reçu l'accréditation « Data Seal of Approval » qui est reconnue par la Commission européenne. Le CINES représente également la communauté francophone au sein du bureau du DSA.

Enfin, il est à noter que le CINES a activement travaillé à démontrer la qualité de ses processus et à développer une chaîne de confiance autour de ses services d'archivage en testant et/ou obtenant les principales certifications du domaine : ISO 16363, DSA, DRAMBORA, TRAC, etc.

5 Le rôle de l'infrastructure Huma-Num

Le TGE (Très Grand Equipement) ADONIS, prédécesseur de la TGIR (Très Grande Infrastructure de Recherche) Huma-Num s'est préoccupé très tôt de la préservation des données en SHS. On rappelle que l'un des principaux objectifs de la préservation à long terme est de pouvoir fournir l'accès dans le futur à des données lisibles techniquement et intelligibles par une personne n'ayant pas participé à leur création. Il est donc indispensable, outre les aspects purement techniques de préservation, d'associer aux données leur contexte de production ainsi que d'autres informations complémentaires qui faciliteront leur

compréhension générale. La TGIR Huma-Num accompagne ainsi les producteurs tout au long de leur projet d'archivage d'un point de vue technique, organisationnel, mais aussi documentaire.

En 2009, un projet pilote, basé sur l'archivage de données orales, a été initié avec le CINES. Ce projet a permis de mettre en évidence des besoins spécifiques aux données produites par les SHS.

Le dispositif d'archivage du CINES, conçu à l'origine pour la préservation des thèses , a dû évoluer pour prendre en compte :

- - le besoin de regrouper des objets dans des « collections » et de les organiser ;
- - de nouveaux formats de données, comme les enregistrements sonores ;
- - l'ajout de métadonnées spécifiques à certaines disciplines scientifique.

Enfin, d'un point plus général, la démarche scientifique fonctionne par accumulation (nouvelle interprétation ou modèle, critique, autre matériau etc.) et il est utile d'en conserver la trace par la gestion de versions tant des données elles-mêmes que des métadonnées. Là encore, des modifications de la plate-forme ont été effectuées.

Pour réaliser ces adaptations, le TGE-Adonis a mis à disposition un ingénieur au CINES durant un an. Il a travaillé sous la direction du service d'archivage et en coordination avec les participants au projet.

Lors du bilan du projet pilote, certains choix on été remis en cause, comme par exemple la nécessité de transiter par l'infrastructure d'archivage du CINES pour la mise à disposition des données sur l'infrastructure de la TGIR Huma-Num.. Ce choix initial d'accès n'était pas toujours pertinent pour les producteurs de données, entre autres pour des questions liées aux droits de diffusion.

Aujourd'hui, la TGIR Huma-Num poursuit l'action initiée par le TGE-Adonis pour la préservation des données en SHS sur le long terme en suivant plusieurs axes :

La TGIR assure le financement de l'archivage pour les projets en SHS. Le coût de l'archivage est assez important, surtout comparé aux financements des projets. Jusqu'à présent, ce besoin n'était pas ou rarement pris en compte.

- L'identification de besoins

La TGIR effectue un repérage de besoins dans les communautés et leur prise en compte au CINES. On peut citer comme exemple l'archivage des données en 3D. La TGIR a initié le processus avec comme partenaire le conservatoire de données 3D en archéologie, ArcheoVision. L'idée était de déterminer et parfois définir les formats de données ainsi que des métadonnées pertinents pour l'archivage. Ce travail se poursuit aujourd'hui dans le cadre d'un consortium créé et soutenu par la TGIR qui a permis d'associer au projet d'autres utilisateurs de données en 3D et ainsi de l'enrichir.

- L'accompagnement des projets d'archivage.

La TGIR, outre le soutien technique de préparation des données, aide les projets à identifier les données à préserver. Une autre phase importante, à laquelle les archivistes du CINES sont totalement associées, est de repenser l'organisation intellectuelle des données pour l'archivage par rapport à l'organisation initiale prévue pour la recherche.

- Le lien avec les autres organismes

La TGIR, en coordination avec le CINES, effectue le lien avec les services d'archives concernés par la préservation, ceux du CNRS et du ministère de l'Enseignement supérieur et de la Recherche. Il est également nécessaire de se concerter avec les Archives de France et d'informer les services compétents (e.g. les Archives nationales) des actions de préservation en cours ainsi que des différents choix effectués.

- La sensibilisation des agences de financement aux problématiques de l'archivage
- Le financement rarement pris en compte jusqu'à présent dans le montage des projets

Le fait de proposer ce service de préservation sur le long terme permet à la TGIR d'informer les communautés, et ce dès le début du projet, de l'utilité de documenter suffisamment leurs données ainsi que de la nécessité de choisir correctement les formats de données et de métadonnées : cela représente un pas important vers l'interopérabilité qui peut être considérée comme une première forme de préservation. Outre son objectif premier de préservation, le service d'archivage à long terme constitue aussi pour la TGIR Huma-Num un dispositif d'incitation à la production de « meilleures » données : c'est-à-dire exprimées dans des formats pérennes et mieux documentées.

6 Quelles perspectives pour les années à venir ?

Compte-tenu du réel « déluge » de production de données numériques actuel, il est clair que l'on est démuni pour assurer la préservation de ces objets sur le long terme et que l'on crée ainsi une fragilité certaine. Au mieux aujourd'hui, est-on capable de migrer les données « en l'état » sur de nouveaux supports à la faveur de nouvelles acquisitions matérielles. Au pire on les perd si le support défaille (défaillance du support ou du matériel de lecture, obsolescence des composants matériels et logiciels, incident majeur etc.) ou plus simplement si l'on oublie l'existence même de ces données. L'un des archivistes d'une grande entreprise Française a récemment fait le constat qu'ils ont quasiment perdu, sur les vingt dernières années, toutes les données numériques qu'ils utilisaient peu. Il ne s'agit malheureusement pas d'un cas isolé.

Une prise de conscience est donc nécessaire tant au niveau institutionnel qu'à celui des producteurs de données. Ce retour d'expérience de quelques années sur la préservation de données en SHS nous permet d'affirmer qu'il s'agit d'un processus complexe techniquement, humainement et donc aussi financièrement. Une première étape vers cette prise de conscience est certainement d'effectuer une bonne information auprès des chercheurs et ingénieurs producteurs de données pour les sensibiliser à ces problématiques. De bonnes pratiques implémentées en début de projet permettent de mieux préserver les données même si l'on ne met pas en œuvre immédiatement un processus complet d'archivage sur le long terme.

Par ailleurs, l'offre de plates-formes technologiques pour effectuer de la préservation à long terme est très réduite aujourd'hui. Les systèmes fonctionnels comme ceux du CNES (Centre National d'Etudes Spatiales), de la BNF (Bibliothèque Nationale de France) et du CINES sont des développements spécifiques en perpétuelle évolution. Ces institutions travaillent activement à pouvoir accueillir, à coût constant, la masse des données numériques en permanente évolution produite chaque année. De même, en ce qui concerne le devenir plus éloigné dans le temps de ces archives, le projet interministériel VITAM (Valeurs Immatérielles Transmises aux Archives pour Mémoire), qui associe les ministères de la Défense, des Affaires étrangères et de la Culture, et sur lequel sera bâtie l'infrastructure future de préservation numérique des services compétents (Archives nationales) vient à peine de débuter. Ce service ne sera donc pas pleinement opérationnel avant quelques années.

Hors les infrastructures qu'il faudra développer, de nouvelles compétences, et donc de nouveaux métiers qui mêleront les versants technologiques, documentaires et archivistiques seront à créer.

Il semble aussi avéré que l'on ne sera pas capable de préserver toute la production numérique scientifique. Il sera indispensable d'être capable d'effectuer une sélection des données à pérenniser Les critères de choix devront avant tout reposer sur des critères scientifiques. Une collaboration avec les archivistes sera nécessaire pour accompagner les scientifiques dans le processus de sélection des données. Bien qu'indispensables, il n'existe aujourd'hui peu ou pas de structures de décision, ni d'expertise réelle dans nos communautés pour adresser ce type de besoins.

Un autre point crucial, est la dimension juridique associée aux données scientifiques, sujet insuffisamment maîtrisé aujourd'hui : on se trouve confronté, selon les experts, à du « droit en train de se faire » et qui demande à se stabiliser. Cette dimension interviendra au moment du choix des données à pérenniser mais surtout dans la manière et la possibilité de les communiquer dans le futur. Dans le même ordre d'idée, Il ne faut pas négliger les aspects réglementaires associés à la pérennisation des données de

la recherche : actuellement peu de textes existent sur leur gestion, et il faudra les « inventer » dans les années à venir ou approprier ceux qui existent déjà.

On peut ainsi constater que beaucoup a été fait ces dernières années, mais aussi que beaucoup reste encore à faire dans le domaine de la préservation des données numériques scientifiques sur le long terme.

La plupart des établissements n'ont et n'auront pas les moyens, tant humains que financiers, pour gérer ces problématiques, en particulier d'être capable de réaliser leur propre système d'archivage. Par conséquent, une mutualisation nationale, voire internationale, des efforts semble indispensable. Cet effort de mutualisation doit avoir lieu au niveau de l'infrastructure mais aussi à celui de la définition et de l'utilisation de standards pour les données et les métadonnées.

Dans un autre registre, il sera utile d'entamer une réflexion sur la définition et l'interopérabilité des catalogues descriptifs des données ainsi pérennisées pour en favoriser la diffusion et la réutilisation.

La coopération entre le CINES et la TGIR Huma-Num participe pleinement au dispositif de mutualisation pour l'ESR français.