La question de l'archivage des données de la recherche en SHS (Sciences Humaines et Sociales)

Michel Jacobson, Nicolas Larrousse, Marion Massol

La question de l'archivage des données de la recherche en SHS (Sciences Humaines et Sociales)

La peste de l'oubli

Peu à peu, étudiant les infinies ressources de l'oubli, il se rendit compte que le jour pourrait bien arriver où l'on reconnaîtrait chaque chose grâce à son inscription, mais où l'on ne se souviendrait plus de son usage. Il se fit alors plus explicite. L'écriteau qu'il suspendit au garrot de la vache fut un modèle de la manière dont les gens de Macondo entendaient lutter contre l'oubli : Voici la vache, il faut la traire tous les matins pour qu'elle produise du lait et le lait, il faut le faire bouillir pour le mélanger avec du café et obtenir du café au lait.

Extrait de « Cent ans de solitudes » Gabriel Garcia Marquez

Aujourd'hui, les données de la recherche sont produites nativement sous forme numérique ou proviennent de la numérisation de données analogiques. Le passage au numérique apporte un gain évident pour la transmission, la diffusion de ces informations et le travail collaboratif qui peut être effectué sur ces données. Une illusion répandue est que ces données sont éternelles, ne serait que par la facilité de les dupliquer. Paradoxalement, un objet numérique peut être plus fragile que son homologue du monde réel : en effet, une photo très abimée peut encore fournir de nombreuses informations, alors qu'un fichier informatique est totalement inutilisable à la moindre altération.

Le contexte général de production de données numériques

La place occupée par l'information numérique n'a fait que gagner du terrain ces dernières décennies. On observe cette augmentation en valeur absolue dans les estimations de volumes stockés ainsi qu'en valeur relative lorsqu'on regarde les secteurs d'activité touchés ou le temps passé dans des activités qui dépendent directement du numérique. Pour illustrer ce propos de manière caricaturale, on pourrait dire que le temps consacré à la gestion de notre messagerie électronique augmente, que le volume de messages que l'on stocke augmente aussi, que le temps passé entre deux messages diminue et qu'enfin notre tolérance à une panne de messagerie diminue également.
L'information est globalement de plus en plus numérique : qu'elle soit produite nativement dans cette forme ou qu'elle provienne de la numérisation de données analogiques. Le monde scientifique n'échappe pas à cette règle. Le numérique est une des formes privilégiées de l'information, car elle est adaptée à l'environnement actuel du chercheur. Son outil d'accès à l'information, son outil de communication avec sa communauté scientifique et son outil de manipulation de l'information est, pour une grande partie de son activité, son ordinateur, ou autre dispositif, connecté au réseau Internet. Dans les sciences humaines et sociales, la typologie des informations numériques manipulées est assez vaste sans pour autant présenter de réelles particularités que l'on ne trouverait pas aussi dans d'autres champs disciplinaires. On trouvera donc comme ailleurs : de la production bureautique, de l'édition, de la correspondance, de l'image, de la modélisation, de la mesure, etc. Les outils utilisés pour

manipuler ces informations sont variés tout comme les formats de représentation utilisés pour leurs codages.

Le numérique offre d'autres avantages que d'être simplement adapté aux outils de notre époque. Les premières utilisations du numérique en sciences humaines ont par exemple permis de réaliser des opérations de calcul trop coûteuses à effectuer par des humains. C'est ainsi que naissent les premiers logiciels statistiques d'analyses factorielles. Le numérique a aussi permis de faire des traitements sur de vastes ensembles de données permettant ainsi d'exploiter des enquêtes à grande échelle. À ces utilisations classiques de l'informatique (le traitement de données volumineuses, les calculs complexes, l'automatisation) viennent s'ajouter de nouvelles utilisations comme la modélisation et la simulation.
La facilité d'accès aux technologies numériques est aussi un facteur de leur succès. Il est par exemple à ce jour, nettement plus compliqué et coûteux de faire de la photo argentique que de la photo numérique. Dans les aspects facilitateurs on peut au moins mentionner la duplication à l'identique et le partage de l'information. Alors que dans le monde des supports analogiques (essentiellement papier) la copie se fait avec perte de qualité et le partage de l'information est fortement limité puisqu'un support ne peut être physiquement qu'à un endroit à un moment donné.
Les principales différences entre les informations sur supports traditionnels et les informations numériques sont la simplicité du numérique (basé sur un système binaire où l'atome d'information n'a que deux valeurs possibles) et l'indépendance du numérique par rapport à son support de stockage. De ces différences découlent les avantages que l'on prête au numérique. Ce sont ces mêmes aspects qui posent de nouveaux problèmes dès lors que l'on envisage de conserver cette information.

En effet, alors qu'une grande partie de l'activité de conservation de l'information sur les supports traditionnels passe par la surveillance de leurs conditions de conservation (conditions d'hygrométrie, de température, d'exposition à la lumière) et par des opérations de restauration en cas de dégradation, les supports du numérique sont par construction fragiles et ne sont pas fait pour durer. Souvent consommateurs d'énergie, ils vieillissent en quelques années de manière dramatique au point de ne plus garantir d'accès à l'information qu'ils détiennent. Les progrès technologiques et industriels les rendent aussi très rapidement obsolètes ou économiquement non rentables en quelques années. Le support du numérique devient alors un simple vecteur de l'information en changement permanent et les choix en la matière sont opportunistes (recherche permanente de réduction des coûts de consommation d'énergie, d'encombrement, de maintenance des équipements, etc.). Les supports étant par nature fragiles, embarquent de plus en plus souvent, même pour leur courte période de vie, des systèmes logiciels de gestion de redondance et de restauration en cas de défaillance. Les efforts pour conserver l'information vont donc se déplacer de la conservation du support d'origine à la gestion des migrations successives de supports. Pour ne pas perdre l'information numérique, on sera amené à la relire et la réécrire fréquemment – ce qui est sans conséquence sur la qualité des données puisque la copie n'entraîne pas de perte – alors que pour des informations sur supports traditionnels on évitera le plus possible le recours à ces opérations.

L'autre difficulté pour la conservation de l'information numérique vient de son aspect codé. La représentation binaire ne pouvant distinguer que deux valeurs, il faut

s'inventer, pour coder les informations dont nous avons besoin, de nouvelles structures enchaînant de longues suites d'unités binaires dans des syntaxes complexes. À un niveau assez « bas », il faut coder la notion de document qui a une réalité physique beaucoup plus répartie que ce que l'utilisateur peut s'imaginer. Les données d'un fichier, sont souvent réparties sur un même support dans plusieurs zones distinctes. Une même donnée peut être codée plusieurs fois afin d'entretenir de la redondance. Enfin, certains systèmes préfèrent éclater les fichiers en une multitude de petits morceaux afin de paralléliser plus facilement des opérations. A un niveau plus « haut », il faut coder les types d'information que manipulent l'utilisateur (du texte, de l'image, du son, des quantités, etc.). Ce codage est en général mieux connu de l'utilisateur, car il en a besoin pour choisir le logiciel adapté à son édition, sa visualisation ou toute autre utilisation. C'est aussi à travers les échecs de partage qu'il va prendre conscience de son importance (échec de partage avec un collègue qui n'a pas le même système ou le même logiciel, ou échec de partage dans le temps quand il tente d'accéder à un fichier créé quelques années plus tôt). Ces codages sont connus sous l'appellation de format de fichier.

A ces difficultés de préservation de l'objet numérique en lui-même et de son intelligibilité s'ajoute le besoin, classique dans le monde des archives et des bibliothèques, d'une description.
Cette description, autrement nommée métadonnée, doit servir à remplir au minimum deux objectifs.
Le premier est que cette information puisse être retrouvée. L'histoire humaine nous a enseigné que le meilleur moyen de perdre une donnée est de ne pas savoir qu'elle existe.
Le deuxième objectif de cette documentation est de rendre l'objet archivé compréhensible pour des utilisateurs qui n'ont pas participé à sa production. Cette compréhension de l'objet numérique nécessite d'appréhender le contexte dans lequel il a été produit, mais également en utilisant quels moyens et avec quels objectifs.
Ce besoin de documentation renvoie à la citation de « 100 ans de solitude » qui débute cet article. On y décrit l'objet, mais également sa fonction et même bien plus dans le cas des habitants de Macondo : on décrit de manière précise comment utiliser les produits « dérivés » et ainsi de suite.
De plus cette métadonnée pose un problème complémentaire dans le contexte de préservation, car elle est elle même exprimée sous forme « numérique » et par conséquent soumise aux même impératifs de conservation que l'objet qu'elle décrit. Ce problème se retrouve dans le village de Macondo puisque les habitants finissent par oublier la signification de l'écriture, et leurs écriteaux deviennent inutiles.

Compte tenu de ces différents dangers qui guettent les objets numériques, on peut estimer leur durée de vie entre cinq et dix ans si l'on ne s'en préoccupe pas.
De plus, la conservation des objets numériques a un coût non négligeable qui jusqu'à lors n'a que peu été pris en compte dans les projets de recherche.
On se rend ainsi compte que la communauté scientifique, par ses pratiques actuelles liées à l'utilisation d'objets numériques, organise sa propre amnésie.

Le cas des données de la recherche en SHS

Les SHS sont très hétérogènes en France (cf. rapport O. Bärring commandé par le TGE-Adonis 2008) avec des diversités de pratiques importantes.

De même, si on les compare aux sciences dites « dures », les objets manipulés peuvent être de nature très variée. Les SHS utilisent des textes bien sûr, mais qui peuvent être exprimés dans des formats différents en fonction de leur utilisation. On peut citer à titre d'exemple le format XML/TEI issu des travaux de la « Text Encoding Initiative ». Les SHS utilisent aussi des enregistrements sonores pour la linguistique et la musicologie, des cartes et des données de SIG (Système d'Information Géographique) pour l'histoire et la géographie, des vidéos pour toutes les disciplines sans oublier les images. Les données en trois dimensions sont essentiellement fournies par l'archéologie mais aujourd'hui les sciences cognitives en font un usage important. On peut évoquer également des données plus spécifiques de type physiologique, comme celles générées par un « eye-tracker », utilisées pour les études comportementales.
Cet inventaire à la Prévert, qui n'a pas la prétention d'être exhaustif, montre le foisonnement des données produites par les SHS.

De même, la typologie des données produites par la recherche en SHS est très étendue : des données brutes à des articles en passant par les différents rapports, sans oublier les courriels.

Pour les données primaires qui sont produites en nombre, on ne dispose en général que de peu de textes réglementaires régissant leur archivage. En effet, seules les données de type administratif entrent dans un cadre précis de préservation qui est en général assurée par l'établissement de recherche.

Alors que les publications sont prises en charge par des circuits d'archivage bien connus des chercheurs, à travers les collectes des bibliothèques ou le dépôt légal, le cycle de vie des données primaires, qui sont créées et utilisées par les chercheurs, n'est souvent, tout simplement, pas pris en compte.

Ces données sont potentiellement réutilisables par d'autres chercheurs, éventuellement issus d'autres communautés. Une carte crée par un géographe pourra servir à un historien qui pourra l'enrichir de données nouvelles qui seront à leur tour utiles pour un archéologue et ainsi de suite.

Les données des SHS ne peuvent pas toujours être régénérées et par voie de conséquence présentent souvent un intérêt patrimonial. Un enregistrement du dernier locuteur s'exprimant dans une langue rare, un relevé en trois dimensions d'un bâtiment qui n'existe plus ou la version numérique d'un manuscrit qui n'est plus consultable en sont autant d'exemples.

Il ne faut pas oublier que le coût de production de ces données représente une part importante d'un projet de recherche.

Tous ces éléments conduisent à envisager de pérenniser ces précieuses données. Cependant, on ne dispose aujourd'hui d'aucune organisation générale pour cela. De même, les agences de financement et les établissements scientifiques ne sont pas encore fortement sensibilisés à ce problème.

On note cependant que le programme de financement Européen « H2020 » demande dorénavant à chaque projet de décrire le devenir des données produites durant son

déroulement. Ce plan de gestion de données (« Data Management Plan ») n'impose aucune contrainte, mais il s'agit d'un début de prise de conscience de l'importance des données par les institutions qui, on l'espère, se poursuivra dans le futur.

Le projet d'archivage pilote initié par le TGE-Adonis

Pour répondre à ces besoins, le TGE-Adonis (Très Grand Equipement Adonis – CNRS UPS 2916) a lancé en 2008 un projet d'archivage pilote pour les données produites et utilisées par les SHS.

Une étude préliminaire avait été commandée au CERN (Laboratoire européen pour la physique des particules). Cette étude s'est basée sur une enquête effectuée auprès des différents centres existants en France et a été réalisée par Olof Bärring.
Le rapport résultant de cette étude, « Hosting of IT services and data for Human and Social Sciences in France », énonçait des recommandations générales pour la mise en place de services numériques pour les SHS en se basant sur l'expérience et l'expertise acquises dans le domaine de la physique des particules. L'une des principales recommandations était de ne pas recréer de nouveaux centres spécifiques pour les SHS, mais plutôt de s'appuyer sur l'existant : c'est-à-dire d'utiliser le centre de calcul de l'IN2P3 (Institut national de physique nucléaire et de physique des particules) pour la fourniture de services et en ce qui concerne l'archivage à long terme de s'appuyer sur l'infrastructure et l'expertise du CINES (Centre Informatique National de l'Enseignement Supérieur).

Le CINES a été chargé en 2004 par le Ministère de l'Enseignement Supérieur et de la Recherche de la mission d'assurer l'archivage pérenne des thèses électroniques, ainsi que des revues numérisées en Sciences Humaines et Sociales du portail Persée. Ces projets ont amené le CINES à concevoir puis à mettre en œuvre une solution générique d'archivage pour ce type de documents numériques. Ce type de documents numériques, qui ne sont pas des archives publiques, n'ont pas vocation à être pris en charge par les Archives Nationales et/ou départementales. La plate-forme ainsi construite au CINES a donc été conçue pour prendre en charge l'archivage définitif de ce type de documents. Par ailleurs, les coûts induits par l'archivage électronique (coûts matériels et humains notamment) justifiaient une stratégie du CINES basée sur l'ouverture de sa plateforme d'archivage à d'autres types de données numériques telles que les archives scientifiques intermédiaires.

Compte-tenu de la diversité, évoquée plus haut, des données produites par les SHS, il était clair qu'il serait nécessaire d'adapter la plate-forme générique du CINES développée à l'origine pour archiver des documents numériques essentiellement exprimés en format PDF (Portable Document Format) et en formats d'image (TIFF, JPEG).
L'objectif du projet pilote était d'évaluer sur un ensemble de données réelles les changements nécessaires à la prise en charge de nouveaux types de données mais également de pouvoir prendre en compte les pratiques des différents acteurs des SHS.

Pour les données, le choix s'est porté sur les ressources orales car la typologie des données est assez variée : des enregistrements sonores ou des vidéos associées à des

transcriptions, des annotations, des enregistrements physiologiques etc. De plus, la communauté de l'oral s'est structurée autour d'un format de métadonnées commun : le format OLAC (Open Language Archives Community) qui permet une description standardisée des objets considérés. Enfin, un guide de bonnes pratiques pour les corpus oraux avait été publié (« Corpus oraux – Guide des bonnes pratiques » - CNRS Editions 2006) et pouvait ainsi servir de référence.

La mise en œuvre du projet pilote s'est faite en associant des personnels du CINES, du CC-IN2P3 (Centre de Calcul de l'IN2P3), du TGE Adonis associés aux producteurs de données représentés par deux centres de ressources numériques labellisés par le CNRS (CRDO Paris & Aix - Centre de Ressource pour les Données de l'Oral) et du SIAF (Service Interministériel des Archives de France).
Un consultant extérieur, Claude Huc, expert du modèle de référence OAIS (Open Archival Information System - ISO 14 721), a assuré la coordination et le suivi du projet. Le projet a fonctionné sur la base d'une formation initiale pour mettre tous les acteurs à niveau sur le modèle OAIS et de réunions très régulières.

Ce projet pilote a permis de mettre rapidement en évidence, outre la prise en compte de nouveaux objets numériques sur la plate-forme du CINES, le besoin de disposer de différentes versions du même objet numérique. En effet, tout comme d'autres données, les données de la recherche en SHS ne sont pas figées et sont susceptibles d'évoluer au cours du temps. De plus, il est intéressant de conserver l'évolution de ces données. En effet, il est utile scientifiquement de pouvoir disposer par exemple de plusieurs versions d'une transcription ou d'une traduction. La plate-forme du CINES, développée à l'origine pour archiver des thèses, ne prévoyait pas ce cas de figure. Il a donc été nécessaire de l'adapter.

La possibilité de créer des liens entre des objets se trouvant sur la plate-forme du CINES a permis de prendre en compte le besoin de « versionnage » en effectuant une relation entre les différentes versions de l'objet. Un effet secondaire de cette nouvelle fonctionnalité a été de l'exploiter pour créer des objets numériques qui relient, au sens propre en créant les liens nécessaires, un ensemble de données entre elles. On retrouve ici, la classique notion de « collection » qui permet de représenter l'organisation intellectuelle des données. Les collections sont plus utilisées par certaines disciplines des SHS que d'autres, mais il s'agissait d'un besoin générique identifié lors du projet pilote. Les métadonnées descriptives fournies au CINES ont été modifiées pour tenir compte de ces développements.

Par ailleurs, la communauté de l'oral utilise des métadonnées spécifiques plus précises pour décrire les objets qu'elle manipule et qui sont complémentaires de celles plus générales demandées par le CINES.
Ces données sont exprimées en format OLAC et sont susceptibles d'évoluer dans le temps alors que l'objet numérique lui même reste identique. Il a donc été introduit la possibilité de mettre à jour uniquement ces métadonnées, dites « métiers » pour prendre en compte les pratiques de cette communauté. Ce dispositif permettra de mettre à jour tout autre type de métadonnées métiers utilisées par la communauté SHS.

En parallèle, le projet pilote a introduit le prise en compte de nouveaux formats sur la plate-forme du CINES, en particulier des formats audio (e.g. Wave) et vidéo. Le choix des

formats a été fait conjointement par les membres du projet pilote pour répondre aux besoins des communautés scientifiques, mais aussi satisfaire aux prérequis du CINES qui en assumera la responsabilité. Du côté du CINES, les formats des données doivent être largement utilisés, leur description doit être publiée, mais il est également important de disposer d'outils libres pour en vérifier la conformité. Du côté des producteurs, il est nécessaire de disposer d'outils qui permettent de migrer les données vers ces formats sans perte de qualité ou avec une perte de qualité acceptable. On constate donc que la prise en compte de nouveaux formats est une « coproduction » entre les différents acteurs du processus d'archivage. Une procédure générale a donc été mise au point durant le projet pour proposer la prise en charge de nouveaux formats au CINES : outre les accords préalables évoqués plus haut, cela nécessite une mise à jour de la plate-forme et une adaptation des procédures de contrôle de la qualité des données versées au CINES.
Cette première étude pour la prise en compte des formats utilisée par la communauté de l'oral a donné lieu à un guide présentant, d'une part la méthodologie employée qui est reproductible pour d'autres types de données, et d'autre part les résultats obtenus pour l'oral (incluant la liste des formats étudiés associés aux critères de choix et à leurs évaluations). La méthodologie de ce guide a déjà été reprise pour d'autres typologies de données comme celle des documents de PDF ou encore pour des plans d'architectes.

Pour réaliser ces adaptations nécessaires, le TGE-Adonis a mis à disposition un ingénieur durant un an qui a travaillé sous la direction du CINES en coordination avec les participants au projet.

Dans sa définition initiale, le projet pilote avait prévu de retourner la donnée archivée ainsi que ses métadonnées vers le Centre de Calcul de l'IN2P3. L'idée était ainsi de mettre en œuvre au centre de calcul l'entité fonctionnelle « accès » prévue dans le modèle de référence OAIS, car le CINES ne donne l'accès à la donnée qu'au service versant (i.e. dans le cas du projet pilote, les deux centres de ressources numériques labellisés par le CNRS : CRDO Paris & Aix). Un effet secondaire était que cette donnée mise à disposition aurait ainsi une garantie de « bonne qualité » puisqu'elle aurait satisfait aux contrôles du CINES pour son format et ses métadonnées.
Cependant certaines données ne sont pas librement communicables ; il est donc nécessaire de pouvoir gérer les droits d'accès pour ce dispositif de diffusion.
Une maquette, se basant sur le service de fédération d'identité fourni par RENATER (Réseau National de télécommunications pour la Technologie l'Enseignement et la Recherche), avait ainsi été réalisée.
Ce « circuit » de mise à disposition avait nécessité une modification du flux de données transitant vers le CINES. Il a été depuis abandonné car, entre autres raisons, cela pouvait donner l'impression que l'on ne pouvait partager que des données préalablement archivées au CINES. De plus, le format d'archivage est souvent différent de celui de diffusion car les objectifs et les utilisations ne sont pas identiques.

En résumé, le projet pilote a permis de valider la possibilité d'archiver des données issues de la recherche en SHS sur la plate-forme du CINES. Les modifications apportées lors du projet-pilote sur la plate-forme du CINES en font l'une des seules en Europe à supporter par exemple la gestion de versions. C'est une fonctionnalité utilisée dorénavant par d'autres utilisateurs de la plate-forme du CINES, par exemple pour

l'archivage des articles de l'archive ouverte HAL (Hyper Articles en Ligne) effectué par le CCSD (Centre pour la Communication Scientifique Directe - CNRS).

Le projet pilote a élaboré également la procédure de prise en compte de nouveaux formats par le CINES pour le domaine des SHS. Une communauté repère un besoin, propose au CINES des formats candidats, puis la décision est prise en commun en collaboration avec le TGE-Adonis.

La difficulté de faire dialoguer les différents métiers des participants du projet pilote a créé des échanges riches et intéressants qui ont fait évoluer fortement le projet, mais aussi les participants. La bonne évaluation du projet par un expert indépendant a été un motif de satisfaction pour tous.

Comme indiqué précédemment, le projet pilote a soulevé un certain nombre de questions comme le besoin de disposer pour ce type d'archives d'informations complémentaires permettant de connaître les conditions de diffusion de la donnée lors de sa sortie du CINES. Le besoin n'existait pas auparavant car, on le rappelle, la plate-forme du CINES a été développée pour archiver des thèses qui avaient vocation à rester définitivement au CINES, la diffusion étant gérée par l'ABES (Agence Bibliographique de l'Enseignement Supérieur). Tant que les données se trouvent au CINES, le problème ne se pose pas car le CINES ne restitue une archive qu'au service qui la lui a versé. En revanche si l'on se place dans l'optique d'un versement futur aux archives nationales, il sera utile de disposer de ces informations qui relèvent du code du patrimoine comme la durée pendant laquelle le CINES doit conserver cette donnée, son sort final et sa communicabilité.

L'archivage des données des SHS proposé par la TGIR Huma-Num

Le TGE-Adonis étant devenu la TGIR (Très Grande Infrastructure de Recherche) Huma-Num à la suite d'une fusion d'unités, le service d'archivage défini par le projet pilote est poursuivi par cette nouvelle unité en suivant les mêmes objectifs.

Le cycle de vie de ces données utilisées comme support à la recherche peut se décomposer en quatre grandes étapes :

- Phase de recueil des données
- Sauvegarde et travail de mise en forme et d'enrichissement
- Dépôt dans un service d'archivage intermédiaire (e.g. le CINES)
- Transfert (ou non) aux archives nationales pour un archivage définitif

En ce qui concerne l'archivage sur le long terme des données de la recherche en SHS, la TGIR Hum-Num accompagne la phase de dépôt au CINES, considéré alors comme un service d'archive intermédiaire.

Le rôle de la TGIR est de repérer les besoins des communautés SHS, de les accompagner dans leur démarche d'archivage et de financer cet archivage. Outre la pérennisation des données produites par la recherche, cette démarche permet de faire prendre conscience aux acteurs d'un projet de recherche de la nécessité d'utiliser de « bons » formats et de

bien documenter les données dès le début du projet. On engendre ainsi une montée en compétence des communautés et une amélioration de la qualité des données produites.

Actuellement, le processus de repérage de nouveaux formats et de leur prise en compte se poursuit. Un travail est en cours sur les données textuelles exprimées en TEI ainsi que sur les données en trois dimensions utilisées en archéologie.

Enfin, le besoin exprimé par le projet pilote de disposer d'informations complémentaires de type archivistique (cf. paragraphe précédent) a été traité en collaboration avec le CINES et les archives de France (SIAF – Service Interministériel des Archives de France), ce qui a conduit à modifier le format des métadonnées fournies au CINES pour tenir compte de ces nouvelles informations.
Ce dernier point a ouvert un chantier plus général dont le but est de disposer de critères d'évaluation pour être à même de fournir ces nouvelles informations en fonction du type de données considéré. On le rappelle, peu de textes réglementaires ne régissent ces données. Il est nécessaire d'inventer les outils d'évaluation de ces données, en collaboration avec les communautés scientifiques qui les produisent, mais aussi avec les archivistes qui en auront la responsabilité.

En complément de son rôle à caractère plus technique indiqué plus haut, la TGIR Huma-Num assure le lien avec les professionnels des d'archives (e.g. CNRS, SIAF, Réseau Aurore, BSN 6, etc.) et les autres partenaires (e.g. BNF) pour préciser la notion de donnée scientifique ainsi que sa prise en compte. De même, la TGIR effectue des actions de sensibilisation sur le sujet de la préservation auprès des différents acteurs de la recherche.

Conclusion

Dans cette époque de frénésie de production de l'information, la fragilité intrinsèque des objets numériques pose, dans le cas des données de la recherche, de nouvelles questions épistémologiques. Comment effectuer le choix des données à conserver, combien de temps doit-on les conserver, qui peut en décider… et bien d'autres encore.

On constate que nous sommes dans une phase de défrichage pour ce vaste sujet et qu'il est nécessaire de se coordonner entre tous les intervenants du processus d'archivage : des producteurs de données jusqu'aux services d'archives.

Enfin, il est nécessaire de sensibiliser les acteurs de recherche à ces besoins nouveaux pour éviter une perte de données massive. Ce serait certainement préjudiciable à la recherche et acterait une mauvaise gestion des ressources produites de manière plus générale.
Cette perte annoncée se produira de manière irrémédiable si l'on ne s'en préoccupe pas à tous les niveaux dès à présent.

Michel Jacobson
Fonction : Chef de projet sur l'archivage électronique
Affiliation : Service interministériel des archives de France

Nicolas Larrousse
Fonction : Responsable de l'archivage à long terme
Affiliation : TGIR Huma-Num

Marion Massol
Fonction : Responsable du département Archivage et diffusion
Affiliation : CINES