Colloque *Les Archives de la Recherche.*
Problèmes et enjeux de la construction du savoir scientifique.
18 au 20 janvier 2012, Université Paris-Sud 11 et CNAM, Paris

Mutualisation et archivage pérenne des données orales : un nouveau cadre technique et juridique au service de la recherche en linguistique

Bernard Bel
Laboratoire Parole et Langage (LPL)
CNRS - Aix-Marseille Université
http://lpl-aix.fr
bernard.bel@lpl-aix.fr

Le Laboratoire Parole et Langage est chargé du développement et de la gestion d'un service versant d'archivage : la « Banque de données parole et langage » (*Speech & Language Data Repository*, SLDR). L'objectif de ce service est de regrouper les ressources associées à la communication orale. Une telle centralisation permet tout d'abord d'éviter que les laboratoires travaillant dans ce domaine ne soient conduits à recréer sans cesse des données. Elle répond par ailleurs au besoin stratégique de rapprocher les connaissances aujourd'hui dispersées dans des domaines variés comme la linguistique descriptive, formelle et computationnelle, la sociolinguistique, la littérature, la traductologie, les neurosciences, la psycholinguistique etc.

Le SLDR est ouvert aux producteurs et utilisateurs du monde entier grâce à son interface multilingue : anglais, espagnol, français et chinois. Pleinement opérationnel depuis 2008, il héberge fin 2012 environ 262 000 documents pour 340 personnes inscrites provenant de 46 pays. Il est associé au CNRTL (www.cnrtl.fr) pour les données écrites dans l'équipement d'excellence ORTOLANG (www.ortolang.fr).

Sur le plan technique, le service versant est destiné à traiter n'importe quel objet numérique comme un « paquet d'informations » (*information package*) ; la typologie des objets repose en effet sur leurs métadonnées descriptives. L'archivage et la mutualisation des ressources s'appuient sur deux grands centres de calcul : le CINES (www.cines.fr) et le CC-IN2P3 (cc.in2p3.fr), sous l'égide du TGE Adonis (www.tge-adonis.fr), coordinateur français du réseau européen DARIAH (www.dariah.eu).

Le téléchargement d'un objet peut être réservé à certaines catégories d'utilisateurs ou aux membres d'une institution bénéficiant d'une licence partagée. Les restrictions d'accès aux documents déposés en archive publique s'appuient sur les dérogations au principe de libre communicabilité prévues par le Code du patrimoine (art. L213-2, loi du 15 juillet 2008).

Le SLDR conserve la trace des téléchargements d'objets effectués après acceptation des licences. Cette trace est rendue visible à toute personne concernée pour faciliter la création de communautés d'utilisateurs.

Étapes de la réalisation

Issu d'une initiative conjointe, en 2006, de la *Direction de l'Information Scientifique* et du *Département scientifique Homme et Société* du CNRS, le *Centre de ressources pour la description de l'oral* (CRDO) était un dispositif de préservation des données orales et linguistiques ouvert à l'ensemble de la communauté scientifique. Les composantes CRDO-Aix et CRDO-Paris étaient portées respectivement par le LPL (www.lpl-aix.fr) et le LACITO (lacito.vjf.cnrs.fr) [1, 2]. Entre 2008 et 2010, ces deux composantes ont participé au projet de stockage, d'archivage pérenne et d'accès mutualisé aux corpus oraux piloté par le TGE Adonis [3] en collaboration avec le *Centre informatique de l'enseignement supérieur* (CINES) et le *Centre de calcul de l'Institut national de physique nucléaire et de physique des particules* (CC-IN2P3). Prenant exemple sur les astrophysiciens, les acteurs du projet ont fait appel à l'*Open Archival Information System* (OAIS) promu par le *Consultative Committee for Space Data Systems* et normalisé en 2003 [4] .

Le service versant CRDO-Aix est passé en production pour l'archivage pérenne en juillet 2010. Courant 2011, les appellations CRDO-Aix et CRDO-Paris ont été abandonnées à la demande de la Direction de l'INSHS. CRDO-Aix a été rebaptisé *Speech and Language Data Repository* (SLDR) et CRDO-Paris *CoCoon*.

Dans le cadre du programme ORTOLANG (www.ortolang.fr) en lien avec la TGIR CORPUS (www.corpus-ir.fr), le SLDR pour les données orales et le CNRTL (www.cnrtl.fr) pour les données écrites sont les centres de ressources à partir desquels un sous-réseau de CLARIN (www.clarin.eu) est en construction en France. ORTOLANG accordera un intérêt particulier à l'interopérabilité des bases de données en s'appuyant sur les directives de CLARIN pour ce qui concerne les métadonnées (DC OLAC, CMDI…), les identifiants pérennes (PID) et les vocabulaires contrôlés (ISOcat).

Les enjeux de l'archivage pérenne

L'archivage de données des sciences humaines et sociales est confronté à deux malentendus qui freinent son accès au monde des « humanités numériques » (*Digital Humanities*). Le premier consiste à croire qu'il pourrait se réduire au stockage de documents informatiques sur une plateforme reliée à des serveurs distants garantissant leur restitution en cas de défaillance technique. Or un tel dispositif pose problème dès qu'il est question de conservation à long terme — typiquement plus de 30 ans. En effet, rien ne garantit que l'organisme financeur sera disposé à négocier une prolongation du stockage de données dont on a perdu la trace des créateurs et/ou des institutions productrices. La remise en cause de la valeur patrimoniale ou scientifique d'une ressource en l'absence de ses producteurs risque d'entraîner sa disparition dans un contexte de rigueur budgétaire. Il est donc nécessaire que l'évaluation du contenu soit effectuée en amont, en accord avec les auteurs des projets, et que sa préservation soit assurée par une archive institutionnelle plutôt qu'un consortium de centres informatiques.

Le second malentendu est illustré par cette remarque entendue d'un professeur : « Tout laisse à penser que dans cinquante ans je ne serai plus de ce monde : à quoi bon préserver mes données pour l'éternité ? » Cette boutade contient l'affirmation implicite que les chercheurs du service public seraient propriétaires des données produites dans le cadre de leur activité professionnelle… Mais elle fait surtout

l'impasse sur l'objectif premier de l'archivage pérenne qui est la réutilisation des données et des résultats de leurs traitements par d'autres équipes et d'autres laboratoires. Laura Campbell [5] (Library of Congress) cite ce proverbe en exergue de son plaidoyer pour la World Digital Library : *"A society grows great when old men plant trees in whose shade they will never sit."* L'archivage numérique perd tout intérêt s'il ne s'inscrit pas dans un processus de mutualisation des ressources, ce qui sous-entend l'existence de service(s) de diffusion et l'accomplissement d'un travail (*Digital Curation*) que l'on peut définir comme « l'ensemble des activités qui ajoutent de la valeur et de la connaissance aux collections » — tout particulièrement les métadonnées descriptives et les identifiants pérennes.

Une implémentation efficace du modèle OAIS

L'*Open Archival Information System* (OAIS) est décrit par la norme ISO 14721. Il a fourni le modèle de référence du projet pilote TGE Adonis officialisé par une convention entre le CINES et le CNRS régissant un service de préservation à long terme de documents numériques (25 mai 2010). Le CINES a par ailleurs reçu l'agrément du Service interministériel des archives de France (SIAF) pour la conservation d'archives publiques courantes et intermédiaires, délivré pour 3 ans à compter du 14 décembre 2010. Indice de qualité, le CINES bénéficie du *Data Seal of Approval* (www.sldr.org/wiki/DSA).

L'engagement du CINES quant à l'archivage pérenne est de triple nature : (1) conserver les données et métadonnées associées ; (2) conserver les informations permettant de déterminer à tout instant les droits d'accès aux données ; (3) préserver la lisibilité des données, ce qui implique la migration des formats de fichiers — sans dégradation de leur contenu — avant que ceux-ci ne deviennent obsolètes.

La diffusion des données des sciences humaines et sociales confiées au CINES dans le cadre du programme d'archivage du TGE Adonis est assurée sur la grille Adonis au CC-IN2P3 (Lyon).

L'articulation entre le service versant (SLDR), le service d'archivage et le service de diffusion est régie par l'OAIS. Son implémentation a nécessité une concertation entre de nombreux acteurs en raison des contraintes associées aux données orales/linguistiques et plus généralement à celles des sciences humaines et sociales :

- La diversité des formats de fichiers : son/vidéo et tous signaux associés à la production de parole ou de chant, ainsi que textes, images, tableaux etc.
- La versatilité des données secondaires (annotations etc.) et des métadonnées descriptives.
- Une approche multilingue des métadonnées descriptives avec un usage fréquent de graphies extra-européennes ou symboliques : alphabet phonétique etc.

La procédure de dépôt des données confiées au SLDR est illustrée sur la figure 1. Les « paquets à verser » (*Submission Information Package*, SIP) sont mis en forme par le service versant et déposés au CINES. Ils sont contrôlés par la plateforme d'archivage (PAC) : conformité de la description, formats de fichiers acceptés pour l'archivage pérenne et cohérence numérique de chaque fichier. Si le paquet est validé, il est enregistré sur la PAC en tant qu'*Archival Information Package* (AIP). Un certificat d'archivage est alors expédié au service versant et le paquet immédiatement transféré

au service de diffusion complété par des données qui lui permettent de créer un « paquet à diffuser » (*Dissemination Information Package*, DIP) dans l'environnement *Fedora Commons* (www.fedora-commons.org).

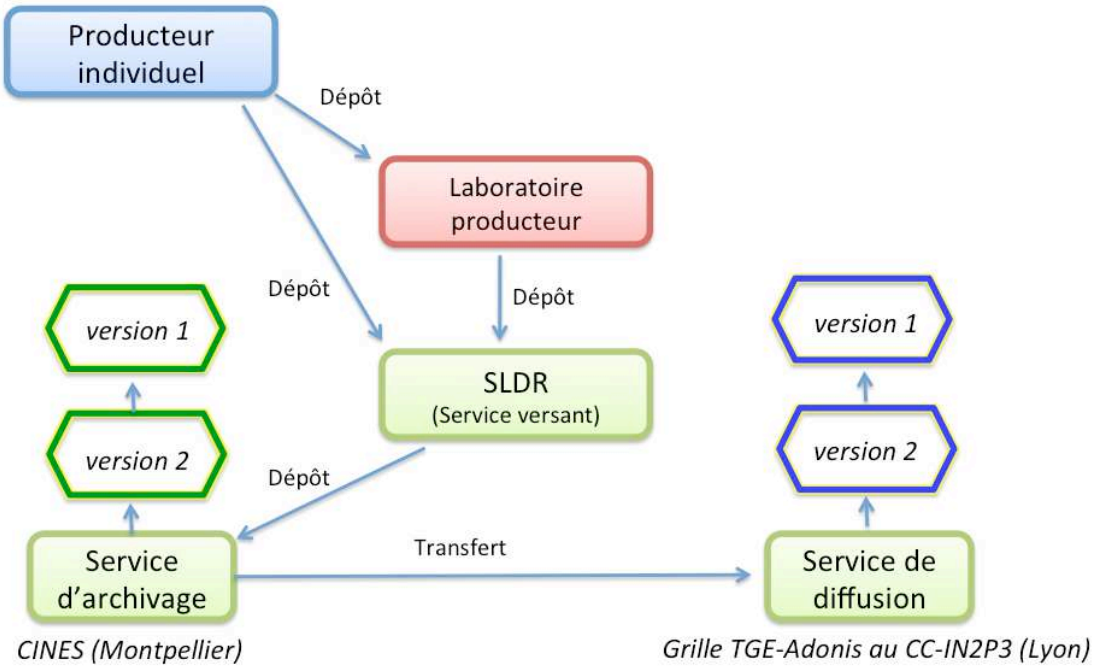

Fig. 1: Le dépôt d'un objet au SLDR

L'accès aux données est illustré sur la figure 2. Les *datastreams* du service de diffusion peuvent être lus directement s'ils sont en libre accès. Dans le cas contraire, l'utilisateur (ou le site client) doit s'acquitter au préalable d'une transaction (non-commerciale) avec le SLDR : identification du demandeur, vérification de son statut dans la base d'utilisateurs inscrits et, si son statut autorise l'accès à la ressource, présentation de la licence SLDR ainsi que, le cas échéant, d'une licence spécifique de l'objet. En cas de succès, la lecture du *datastream* est autorisée. Cette transaction est mise en mémoire de telle sorte que l'accès à des *datastreams* de statut identique sera immédiatement possible pour toute requête lancée au cours de la même session. Cette mémorisation permet l'exécution de requêtes complexes qui combinent plusieurs objets ou documents. Il sera par ailleurs possible d'étendre cette interopérabilité à d'autres sites lorsqu'un mécanisme de transfert d'identité (*Single Sign-On*, SSO) aura été installé entre sites partenaires — une priorité des réseaux CLARIN et DARIAH.

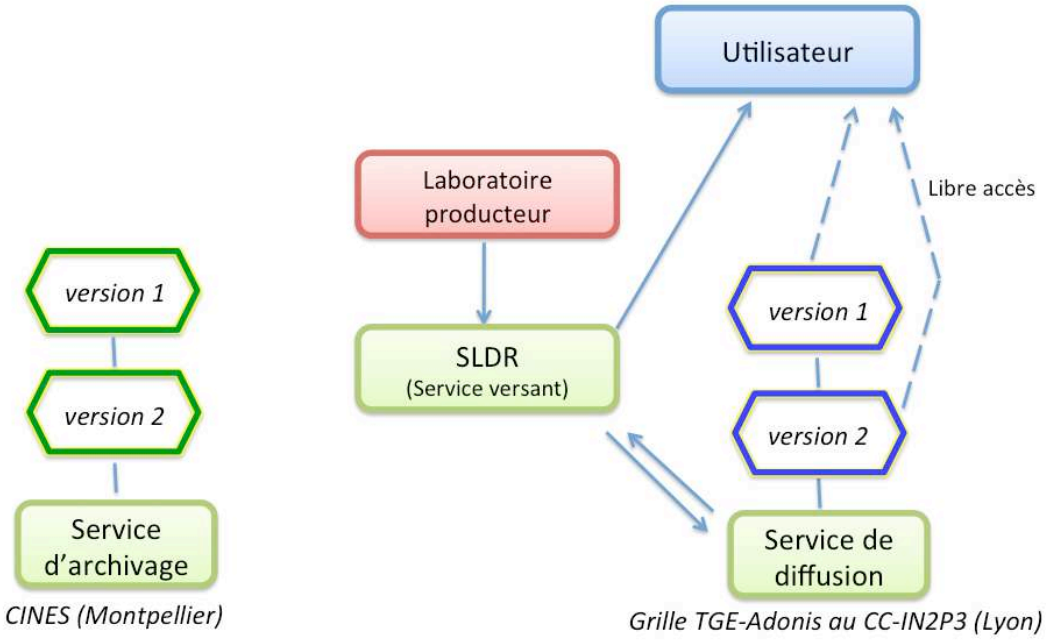

Fig. 2 : Accès aux documents. Noter que le service d'archivage n'est pas impliqué dans ce processus.

Le SLDR est capable de reconstruire entièrement un objet identique à la source à partir des *datastreams* disponibles sur le service de diffusion (voir figure 3). Cette procédure de récupération autorise la suppression sur le service versant des objets déposés en archivage pérenne ou intermédiaire.

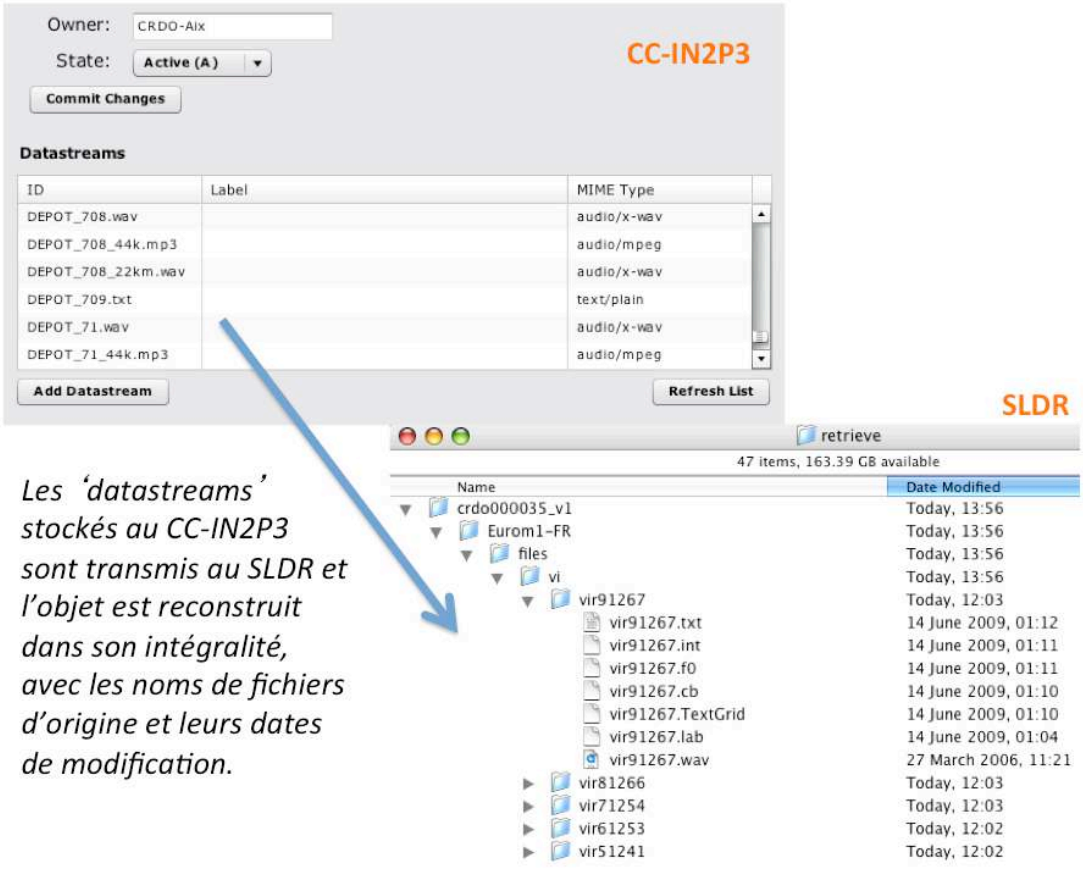

Fig. 3 : Récupération d'un objet archivé à partir du service de diffusion

Peut-on modifier un objet archivé ?

La plupart des objets numériques contiennent à la fois des données figées (ex. les données primaires son/vidéo) et des données documentaires (métadonnées descriptives, annotations, conditions d'accès…) modifiables à tout moment. Sachant qu'une archive institutionnelle doit préserver toutes les versions des paquets d'information qui lui ont été confiés, il serait regrettable de redéposer l'intégralité d'un objet après chaque modification de ses données documentaires.

La solution mise au point avec le CINES consiste à scinder le paquet à verser (*Submission Information Package*, SIP) en deux répertoires de données archivables : l'un (DEPOT) contenant les données « figées » et le second (DEPOT/DESC) les données documentaires. Les mises à jour peuvent alors porter sur l'objet entier ou se limiter au contenu de la partie documentaire, ce qu'on désigne comme « mise à jour de métadonnées ». La structure du SIP est présentée plus bas.

Intégrer la pratique de l'archivage aux projets de recherche

Le SLDR utilise les plateformes de « test » du CINES et du CC-IN2P3 pour pratiquer l'archivage intermédiaire. Les mécanismes de dépôt et d'accès aux données y sont identiques à ceux décrits sur les figures 1 et 2, avec pour seule exception que le CINES ne conserve pas les AIP sur sa plateforme d'archivage. Lorsqu'un objet déposé en archivage intermédiaire est devenu « stable » — par exemple, aucune nouvelle donnée primaire ne sera ajoutée à un corpus de parole —, il est possible de le basculer en archivage pérenne sans modifier les identifiants d'accès. Toutes les versions déposées précédemment en archivage intermédiaire sont alors éliminées du service de diffusion.

La pratique de l'archivage intermédiaire permet d'encourager les producteurs à déposer leurs données dès le commencement d'un programme de recherche ou de documentation. Ils peuvent ainsi s'assurer de la compatibilité de leurs dépôts avec le dispositif d'archivage, notamment pour ce qui concerne les formats de fichiers. L'archivage n'est donc plus traité comme une activité annexe, trop souvent reportée en fin de projet alors que l'équipe de recherche est dispersée ou mobilisée sur d'autres programmes. Cet accompagnement des projets est en accord avec les préconisations de Habert et Huc [6] :

The production of "sustainable data" will get more attention from researchers and from laboratories as well only if these archived data are evaluated as such for individuals and laboratories, just like papers, if they are really made part of the scientific production. Researchers will then be in a better position to plan the archiving process, to decide what is precious and how to document it.

Structure du SIP

La structure du paquet à verser (SIP) est illustrée sur la figure 4. Les données archivables sont déposées dans les répertoires DEPOT et DEPOT/DESC comme précédemment expliqué. Chaque SIP comprend aussi un répertoire DIFFUSION dont le contenu est directement transmis au service de diffusion. Ces documents qui échappent à l'archivage utilisent des formats réservés à la diffusion, comme par exemple FLV et MP3 pour les *streamings* vidéo et audio.

Un fichier *sip.xml* est inclus à chaque versement. Il contient les métadonnées d'archivage : identification minimale de l'objet, structure des répertoires, catalogue des fichiers incluant leurs empreintes numériques pour vérifier l'intégrité du transfert.

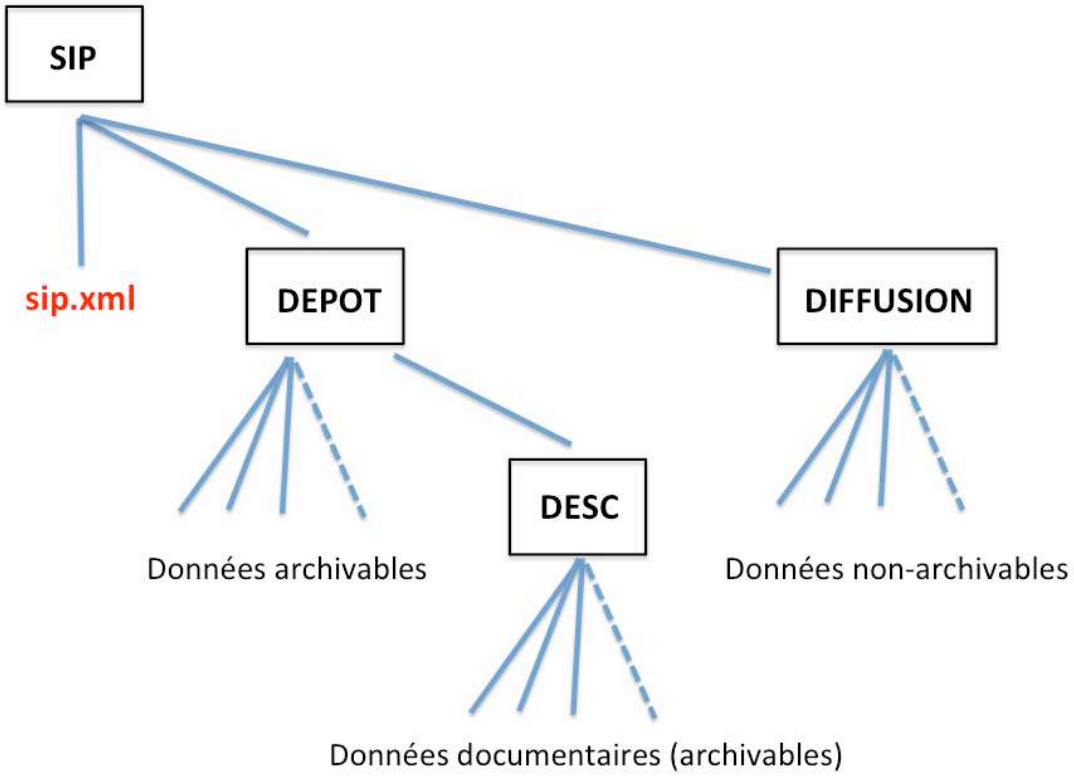

Fig. 4 : Structure du *Submission Information Package* (SIP) au CINES

La création du SIP est une étape critique du travail de curation des données après leur dépôt sur le service versant d'archivage. La curation consiste ici à modifier la structure du dépôt et en extraire les propriétés fondamentales, ainsi que générer les métadonnées qui assureront son archivage, sa localisation et sa réutilisation. Pour les objets de la recherche, la réutilisation débouche potentiellement sur la création et le dépôt de nouvelles données ; ce cycle de vie est plus court que pour les ressources de valeur exclusivement patrimoniale. Cette fréquence accrue de réutilisation oblige les services versants à automatiser autant que possible le travail de mise en forme et d'extraction de propriétés (*packaging*). L'algorithme utilisé par le SLDR fait l'objet de fréquentes mises à jour à l'occasion du dépôt d'objets dont la structure et le contenu nécessitent une attention particulière (voir www.sldr.org/wiki/Packaging-fr).

Segmentation des objets, identifiants ARK

Des contraintes techniques sur le site d'archivage (CINES) et le service de diffusion (CC-IN2P3) entraînent des limites de taille des *Archival Information Packages* (AIP). La limite courante a été fixée à 40 Go/10 000 fichiers. Les objets de plus grande taille doivent donc être segmentés en plusieurs AIP, cette segmentation demeurant invisible aux utilisateurs.

Chaque AIP reçoit un identifiant ARK (*Archival Resource Key*) pour chacune de ses versions. Le CINES préserve des liens pour reconstituer la chaîne de versionnage de sorte que l'identifiant de la première version suffit à caractériser la chaîne.

Un exemple de segmentation d'objet est donné figure 5 pour *The Open ANC* (sldr000770) segmenté en sept AIP car il contient plus de 60 000 fichiers.

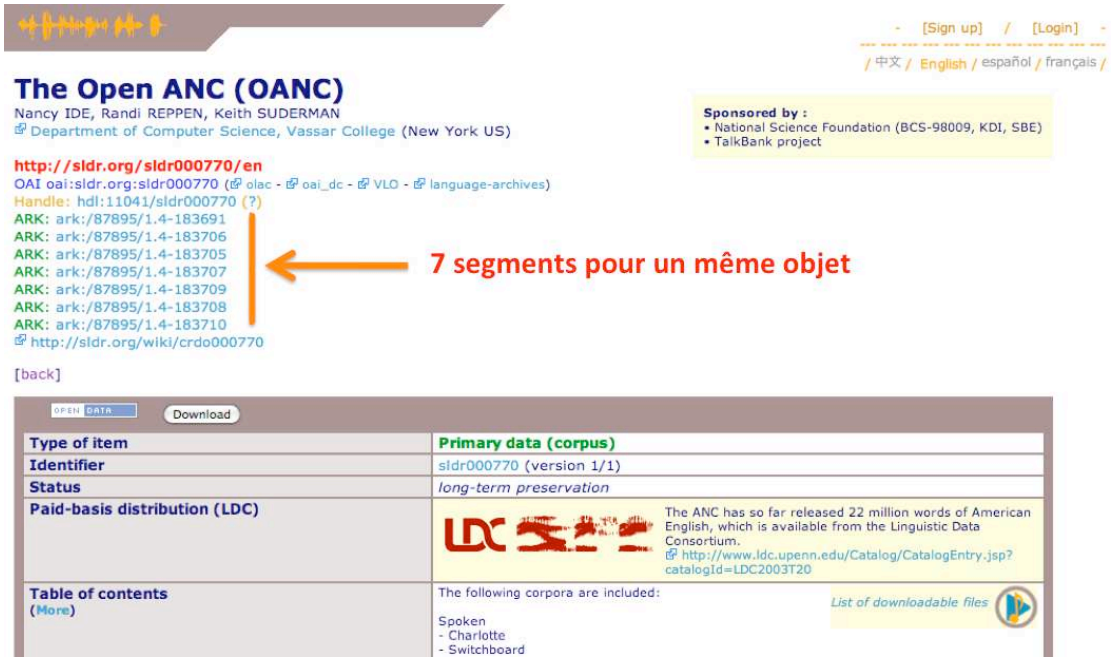

Fig. 5 : Segmentation de l'objet sldr000770

Identifiants pérennes (PID)

Il n'existe pas de mécanisme de résolution global pour les identifiants ARK, ce qui revient à dire qu'aucun dispositif ne permet de localiser directement un paquet d'informations à partir de son identifiant, par exemple ark:/87895/1.4-183706. Un utilisateur averti peut savoir que '87895' désigne le CINES, mais cet organisme n'assurant pas la diffusion des objets, la seule démarche envisageable serait une restitution d'archive au demandeur légalement autorisé.

Prenant exemple sur la Bibliothèque nationale de France (www.bnf.fr/fr/professionnels/s_informer_autres_numeros/a.ark_autres_numeros.html), nous avons implémenté localement une autorité nommante (*name mapping authority*). Par exemple, www.sldr.org/ark:/87895/1.4-183706 pointe vers le segment 2 de *The Open ANC*. On a toutefois besoin d'une granularité plus fine d'identifiants et ceux-ci devraient être indépendants des AIP. Des identifiants pérennes (*Persistent IDentifiers*, PID) sont donc attribués à chaque objet et aux documents qu'il contient, mais pas à chaque AIP d'un objet segmenté.

Le *Handle System* est utilisé pour construire ces identifiants avec une syntaxe compréhensible, par exemple hdl:11041/BeQuali-000532 pointe vers l'objet *BeQuali-000532* dont l'URL est www.sldr.org/BeQuali-000532.

La figure 6 donne le script de création du PID de *The Open ANC* dans le Handle system. Les champs DESC sont (provisoirement) utilisés pour indiquer le service versant d'archivage, le nom de l'objet dans les quatre langues de navigation, ainsi que les identifiants ARK des paquets archivés.

```
AUTHENTICATE PUBKEY:300:0.NA/11041/hs/svr_1/admpriv.bin|[password]
CREATE 11041/sldr000770
100 HS_ADMIN 86400 1110 ADMIN 300:110011111110:0.NA/11041
7 EMAIL 86400 1110 UTF8 webmaster@sldr.org
8 URL 86400 1110 UTF8 http://sldr.org/sldr000770
9 DESC 86400 1110 UTF8 Speech & Language Data Repository (SLDR)
10 DESC 86400 1110 UTF8 The Open ANC (OANC)
11 DESC 86400 1110 UTF8 Corpus abierto del lenguaje americano
12 DESC 86400 1110 UTF8 Corpus ouvert de l'américain
13 DESC 86400 1110 UTF8 打开语料库的美国语言
14 DESC 86400 1110 UTF8 ark:/87895/1.4-183691
15 DESC 86400 1110 UTF8 ark:/87895/1.4-183706
16 DESC 86400 1110 UTF8 ark:/87895/1.4-183705
17 DESC 86400 1110 UTF8 ark:/87895/1.4-183707
18 DESC 86400 1110 UTF8 ark:/87895/1.4-183709
19 DESC 86400 1110 UTF8 ark:/87895/1.4-183708
20 DESC 86400 1110 UTF8 ark:/87895/1.4-183710
```

Fig. 6 : Script de création de l'identifiant pérenne (PID) de l'objet sldr000770

Les PID sont attribués à chaque version d'un objet, ainsi qu'à chaque document contenu dans cette version. Par exemple, hdl:11041/swedia-000788_v1_f388 pointe vers le fichier numéro 388 de la version 1 de l'objet *swedia-000788*. Si le numéro de version n'est pas spécifié, le PID pointe vers la version la plus récente du document. Nous avons pour projet la gestion de PID dans lesquels l'index du fichier pourrait être remplacé par un code dérivé de son nom afin qu'il reste localisable lorsque cet index varie en fonction des versions de l'objet auquel il appartient.

Certains identifiants pérennes pointent directement vers un contenu en vue d'un téléchargement ou du traitement d'un document. D'autres pointent vers la description d'un (ensemble de) document(s) dans un format destiné à un lecteur humain (page web) ou mécanique (métadonnées dans un format XML). Actuellement, au SLDR, un PID désignant un objet pointe vers sa page descriptive, mais des suffixes seront utilisés pour récupérer ses métadonnées aux formats OAI_DC, OLAC, CMDI etc.

L'utilisation systématique des PID pour assurer l'interopérabilité des bases de données est un thème transversal débattu dans les *Virtual Competency Centres* VCC1 et VCC3 du réseau DARIAH, ainsi que dans les travaux de CLARIN et d'ORTOLANG en liaison avec APARSEN (www.aparsen.eu). C'est pourquoi nous évitons l'implémentation d'options qui pourraient devenir inopérantes une fois que des règles de bonne pratique auront été validées par ces groupes de travail.

Gestion systématique des droits d'accès

Parmi les réticences des chercheurs à archiver/mutualiser leurs ressources orales/linguistiques figure presque toujours l'exigence de maîtrise des droits d'accès. Cette demande légitime met en exergue une apparente contradiction entre l'obligation de donner accès à toutes les données issues de la recherche publique et les limites imposées par la législation : propriété intellectuelle, droit à l'image, protection de la vie privée, confidentialité d'informations à caractère médical etc. Or ce n'est pas la

législation qui fait obstacle, mais plutôt l'absence de solutions techniques adaptées qui oblige les collecteurs à des simplifications incompatibles avec la réalité du terrain.

Un modèle générique de gestion des droits d'accès doit prendre en compte :

- les catégories d'utilisateurs définies dans le profil de l'organisme producteur (par défaut, le profil SLDR) ;
- le statut juridique des données imposé par le cadre légal de leur dépôt déterminé par l'organisme producteur (par défaut, le cadre SLDR).

Les objets soumis à l'archivage dans une archive institutionnelle (le CINES) doivent respecter les dispositions du Code du patrimoine français pour ce qui concerne les archives publiques. Ce cadre légal actualisé en 2008 (15 juillet, articles L213 1-5) constitue une avancée considérable de par l'existence de règles formelles établissant les conditions d'accès aux archives publiques. Cette formalisation permet en effet d'intéger au processus d'archivage les procédures de gestion des droits d'accès aux documents.

Le Code version 2008 instaure un changement radical dans les pratiques de l'archivage numérique puisque l'article L213-1 stipule que *« les archives publiques […] sont communicables que plein droit »*. L'article L213-2 introduit cependant 24 dérogations au principe de libre-communicabilité qui autorisent les services d'archivage à réserver l'accès à certains documents pendant une période déterminée.

Ces dérogations ont été codifiées par le Service interministériel des Archives de France (SIAF, voir www.sldr.org/wiki/table_derogations_fr). Ainsi, par exemple (code AR048) : *« 50 ans* […] *pour les documents dont la communication porte atteinte* […] *à la protection de la vie privée* […] *(ou les) documents qui portent une appréciation ou un jugement de valeur sur une personne physique, nommément désignée ou facilement identifiable, ou qui font apparaître le comportement d'une personne dans des conditions susceptibles de lui porter préjudice »*. De tels objets peuvent donc, pendant 50 ans, faire l'objet d'une diffusion réservée à certains utilisateurs sous le couvert d'autorisations signées par les informateurs et producteurs.

Fig. 7 : Signature d'autorisations de diffusion par Victorine Dumas et Claude Jouval au Mas de la Pyramide, St-Rémy-de-Provence, le 22 avril 2011

La dérogation AR048 (50 ans) est la plus fréquente pour ce qui concerne les corpus de parole. Il faut néanmoins noter qu'elle ne peut pas se substituer aux dispositions du Code de propriété intellectuelle puisque les droits patrimoniaux s'exercent jusque 70 ans après le décès de l'auteur d'une œuvre. Il convient donc de s'assurer qu'un enregistrement de parole (dans le cadre d'un travail linguistique) ne constitue pas une

œuvre susceptible de conférer un droit d'auteur aux locuteurs ; dans le cas contraire, il faut veiller à ce que l'autorisation de diffusion soit équivalente à une cession des droits patrimoniaux.

Une autre obligation nouvelle du Code du patrimoine est celle (art. L213-5) d'obliger *« toute administration détentrice d'archives publiques ou privées* [à] *motiver tout refus qu'elle oppose à une demande de communication de documents d'archives »*. L'utilisateur qui tente de télécharger un document en accès protégé doit être informé des raisons de l'interdiction d'accès ainsi que de la date à laquelle ce document deviendra librement accessible. Le dispositif implémenté au SLDR est illustré sur la figure 8 (voir http://sldr.org/sldr000019/toc). Ce corpus de parole est publiquement accessible au format AAC (*Advanced Audio Coding*) qui utilise une compression avec perte de données. Ce format est ici qualifié de « basse résolution ». Par contre, les fichiers sources en « haute résolution » (format WAV) sont réservés aux utilisateurs identifiés comme chercheurs, enseignants ou étudiants en sciences du langage. Cette double procédure de diffusion est satisfaisante pour des locuteurs qui craindraient qu'un document en haute résolution soit utilisé malintentionnellement (falsification, appropriation etc.) par une personne non-identifiée. Le fait que le même corpus soit entièrement audible en basse résolution suffit à respecter le principe de libre communicabilité imposé par le Code du patrimoine.

Fig. 8 : Partage de données en basse et haute résolutions

Le passage de la souris au-dessus d'un lien en accès réservé fait apparaître le texte de l'article concernant la dérogation AR048 qui s'applique ici. Ce texte est assorti du commentaire *« Fichiers haute résolution, distribution restreinte »* ainsi que de la date limite de cette restriction : 16 mars 2060. Toutes ces informations sont affichées dans la langue de navigation choisie par l'utilisateur.

En cliquant sur un lien en accès réservé, l'utilisateur s'engage dans le processus d'identification, d'examen des droits et d'acceptation des licences décrit plus haut. Le lien peut directement aboutir à la ressource si l'autorisation a déjà été accordée.

Le réglage des droits d'accès est par défaut identique pour l'ensemble des documents contenus dans un objet. Toutefois il peut être modifié au niveau de chaque répertoire par le biais d'un fichier XML accessible au producteur de l'objet (voir figure 9), ou même au niveau de chaque document par un autre dispositif.

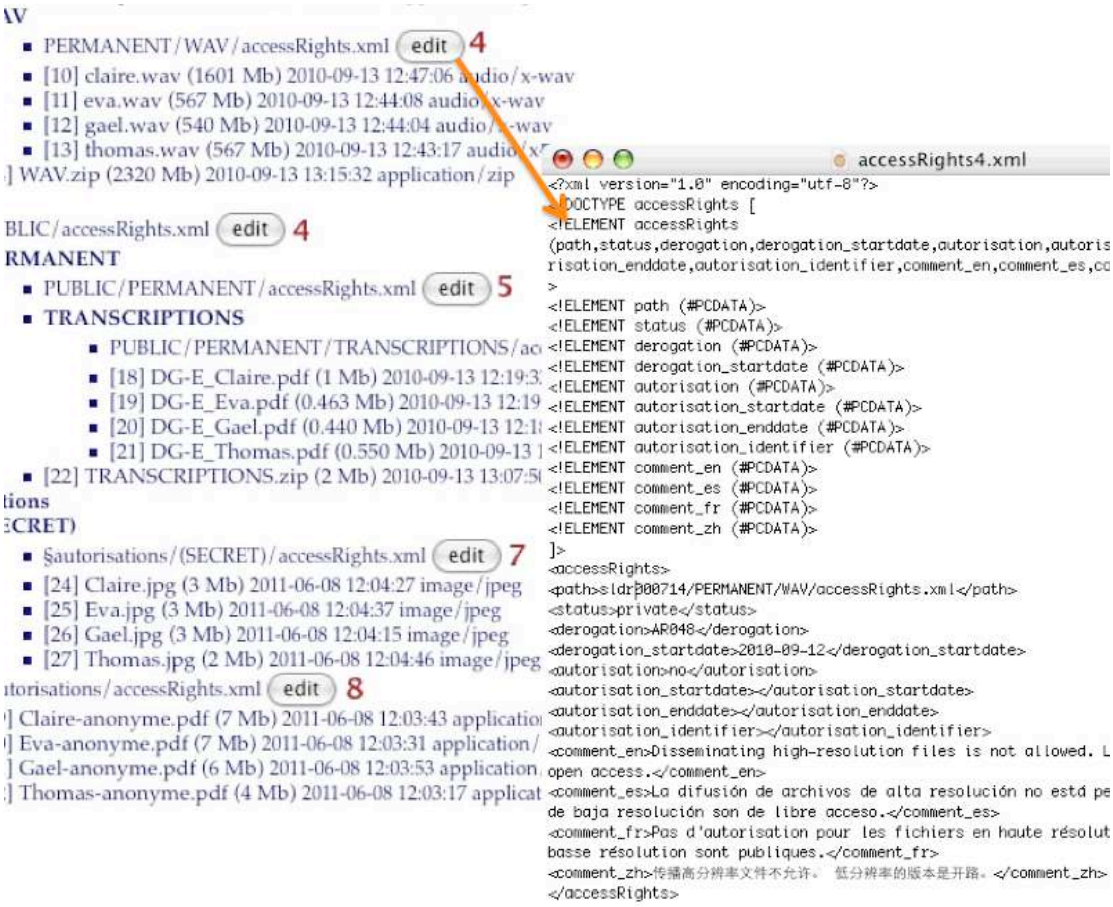

Fig. 9 : Contrôle des droits d'accès au niveau d'un répertoire

Le SLDR supervise quotidiennement les statuts des droits d'accès aux objets et documents afin de prévenir les administrateurs de l'obligation de les basculer en accès public — ou restreint si une autorisation est parvenue à expiration.

Les conditions d'accès aux objets numériques doivent être rendues visibles dans les métadonnées distribuées par le serveur OAI-PMH associé au SLDR, comme illustré sur la figure 10. L'utilisation d'un vocabulaire contrôlé emprunté à l'espace de noms *info:eu-repo* répond aux spécifications des portails DRIVER (www.driver-repository.eu) et OpenAIRE (www.openaire.eu) dédiés à la recherche européenne.

<dc:type>info:eu-repo/semantics/dataset</dc:type>
<dc:rights>info:eu-repo/date/submitted/2008-05-02</dc:rights>
<dc:rights>info:eu-repo/semantics/embargoedAccess</dc:rights>
<dc:rights>info:eu-repo/date/embargoEnd/2058-05-02</dc:rights>
- <dcterms:accessRights xml:lang="en">
SLDR licence; rightsHolder = Laboratoire parole et langage - UMR 7309 (LPL, Aix-en-Provence FR)
</dcterms:accessRights>
<dcterms:accessRights xml:lang="en">Privileged user: CID-user</dcterms:accessRights>
<dcterms:license xsi:type="dcterms:URI">http://sldr.org/licence_v1/en</dcterms:license>
<dcterms:license xsi:type="dcterms:URI">http://sldr.org/licence_v1/es</dcterms:license>
<dcterms:license xsi:type="dcterms:URI">http://sldr.org/licence_v1/fr</dcterms:license>
<dcterms:license xsi:type="dcterms:URI">http://sldr.org/licence_v1/zh</dcterms:license>
<dcterms:provenance xml:lang="en">long-term preservation</dcterms:provenance>
<dcterms:provenance xml:lang="es">archivo</dcterms:provenance>
<dcterms:provenance xml:lang="fr">archive pérenne</dcterms:provenance>
<dcterms:provenance xml:lang="zh">长期档案</dcterms:provenance>
- <dcterms:accessRights xml:lang="fr">
Restriction AR048 (50 ans à partir de 2008-05-02) - Documents dont la communication porte atteinte à la protection de la vie privée ou portant appréciation ou jugement de valeur sur une personne physique nommément désignée, ou facilement identifiable, ou qui font apparaître le comportement d'une personne dans des conditions susceptibles de lui porter préjudice. (Code du Patrimoine, art. L. 213-2, I, 3)
</dcterms:accessRights>
- <dcterms:accessRights xml:lang="en">
Restriction AR048 (50 years from 2008-05-02) - Documents disclosure of which undermines the protection of privacy or for appreciation or value judgments about a person named or easily identifiable, or which reveal the behavior of a person under circumstances which might cause him/her prejudice. (Code du Patrimoine, art. L. 213-2, I, 3)
</dcterms:accessRights>
- <dcterms:accessRights xml:lang="zh">
制约 AR048 (从2008-05-02准入限制组50年) - 提供破坏隐私保护或欣赏或关于容易辨认的人的价值判断的命名或, 或者在情况也许带来他或她的伤害下显露人行为的透露。 (Code du Patrimoine, 艺术。 L. 213-2, I, 3)
</dcterms:accessRights>
- <dcterms:accessRights xml:lang="es">
Restriccion AR048 (50 years from 2008-05-02) - Documentos de divulgación de lo que perjudica la protección de la intimidad o de los juicios de valor acerca de apreciación o una persona con nombre o fácilmente identificables, o que revelan el comportamiento de una persona en circunstancias que podrían llevarle lesión. (Code du Patrimoine, art. L. 213-2, I, 3)
</dcterms:accessRights>

Fig. 10 : Mention des droits d'accès dans les métadonnées OLAC de l'objet sldr000027

Licences partagées, licences commerciales

Les objets déposés au SLDR peuvent être partagés entre les membres d'une institution lorsque celle-ci bénéficie d'une licence partagée, non-commerciale ou parfois commerciale dans le cas d'un achat collectif. Un exemple (non-commercial) est celui du *Buckeye Corpus of Conversational Speech* distribué par Ohio State University (sldr000776). Pour partager une licence les conditions suivantes doivent être réalisées :

- L'institution est inscrite sur le site du SLDR (voir www.sldr.org/labs) ;
- L'utilisateur est inscrit au SLDR et identifié comme membre de cette institution ;
- Une copie de l'objet est disponible au SLDR ;
- Le SLDR a reçu un document certifiant que la licence est accordée à cette institution.

Bien que les objets déposés en archive publique ne puissent faire l'objet de tractations commerciales, la possibilité de restreindre leur accès à des catégories d'utilisateurs

permet de réaliser une double distribution de certaines ressources. L'accès est alors gratuit pour les chercheurs, enseignants et étudiants, tout autre utilisateur étant dirigé vers des services qui utilisent un modèle économique différent comme le *Language Data Consortium* (LDC, www.ldc.upenn.edu) et *European Language Resources Association* (ELRA, www.elra.info). Ce cas de figure est illustré par les objets sldr000770 pour LDC et sldr000035 pour ELRA.

Valorisation des dépôts

La mutualisation des ressources orales et linguistiques se concrétise au minimum par un partage de données. Ce besoin est ressenti en priorité pour les enregistrements de parole ayant une valeur patrimoniale avec pour cible le grand public, les artistes ou les médias. En ciblant la communauté scientifique, la mutualisation encourage aussi leur réutilisation, qu'il s'agisse de données primaires (enregistrements bruts) ou de données secondaires produites par des chercheurs : transcriptions, traductions, annotations, analyses formelles etc.

Il est important de rendre compte de cette réutilisation pour alimenter ce processus de valorisation des ressources. Le SLDR propose à cet effet trois dispositifs d'enrichissement des informations :

1. L'inscription dans les métadonnées de liens entre un objet archivé et les programmes de recherche qui en tirent parti ;
2. L'inscription de publications relatives aux travaux qui font usage de l'objet ;
3. L'affichage d'une « communauté d'utilisateurs » illustré sur la figure 11. Les personnes ayant été autorisées à télécharger un objet ont accès à la liste des téléchargements effectués par d'autres utilisateurs avec mention, pour chacun, de son affiliation professionnelle et de son domaine de recherche. Un bouton « contact » permet d'envoyer un message à l'utilisateur sans révéler son adresse électronique, un peu à la manière de ce que proposent les réseaux sociaux (Web 2.0).

Téléchargé (27) données secondaires (ressource) VfrLPL - http://sldr.org/sldr000533/fr

ID	Prénom et nom	Organisme	Domaine de recherche	Droits	Date du téléchargement	Version
2003	M Stéphane RAUZY Contact	Laboratoire parole et langage - UMR 7309 (LPL, Aix-en-Provence FR)	Informatique et Linguistique	déposants du SLDR	2007-05-21 licence #0	1
15	Mme Morgane ADER Contact	Laboratoire parole et langage - UMR 7309 (LPL, Aix-en-Provence FR)	informatique appliqué à la linguistique	déposants du SLDR	2007-05-23 licence #0	1
2003	M Stéphane RAUZY Contact	Laboratoire parole et langage - UMR 7309 (LPL, Aix-en-Provence FR)	Informatique et Linguistique	déposants du SLDR	2007-05-25 licence #0	1
2018	M Cedric GENDROT Contact	Institut de linguistique et de phonétique générales et appliquées (ILPGA, Paris FR)	Phonétique	chercheurs, enseignants, étudiants	2007-06-06 licence #0	1
2019	M Bruno GUILLAUME Contact	Laboratoire lorrain de recherche en informatique et ses applications - UMR 7503 (Loria, Nancy FR)	TAL	chercheurs, enseignants, étudiants	2007-06-06 licence #0	1
2024	M Yoshiyuki FUKUSHIMA Contact	Osaka University, Japan	linguistique française, analyse de la communication, didactique du français	chercheurs, enseignants, étudiants	2007-06-17 licence #0	1
2028	M Matthieu HERMET Contact	Universite d'Ottawa - EITI http://www.site.uottawa.ca/index.shtml	CALL, Informatique	chercheurs, enseignants, étudiants	2007-07-20 licence #0	1

Fig. 11 : Communauté d'utilisateurs de l'objet sldr000533 (extrait)

Conclusion

Nous avons montré que l'implémentation d'un service versant d'archivage numérique se devait de respecter un ensemble de contraintes d'ordre technique ou juridique, mais aussi méthodologiques pour l'accompagnement de projets produisant des objets archivables. Le modèle OAIS permet un cadrage conceptuel et institutionnel prenant en compte tous les acteurs de l'archivage, condition indispensable pour assurer la pérennité du dispositif.

Les professionnels de l'archivage numérique de données scientifiques doivent par ailleurs relever un défi exposé par Margaret Hedstrom [7] : la distribution inégale des compétences sur une échelle qui s'étend de l'expertise dans le domaine à la curation de données. La plupart des professionnels se situant aujourd'hui à faible distance des extrémités de cette échelle, on déplore une carence relative de compétences « hybrides ». Hedstrom suggère que les problématiques de gestion et de curation des données soient intégrées aux curricula de formation à la recherche, et d'autre part que les curateurs soient mieux formés au traitement spécifique des objets de la recherche. C'est à ce prix que les programmes collaboratifs (comme ORTOLANG dans le domaine de la linguistique) pourront susciter une évolution des pratiques vers le paradigme de *digital scholarship* qui met l'accent sur le traitement intensif de données fortement diversifiées dans leurs sources et contenus.

Il reste à espérer que le contexte de réoganisation de la recherche, en France, reste favorable au soutien de grandes infrastructures de recherche (TGIR) en liaison étroite avec les réseaux internationaux comme DARIAH et CLARIN en Europe. Les craintes exprimées en 2010 par Habert et Huc [6] sont toujours d'actualité :

Long term preservation policy implies a minimal stability for the concerned communities. At the moment, it is not the case for SSH, which pay lip service to the global aim of digital archiving without necessarily having the strength to make the necessary decisions and to stick to it. […] Since the 17th century at least, France has been a very centralized state. The past thirty years dramatically changed this tendency, with the law on state decentralization in 1982 and the law in 2007 granting more autonomy to universities. To make a long story short, there is now a contradiction between a centralized state and centrifugal forces.

L'engagement de tous les acteurs institutionnels est vital en réponse à la demande croissante de développement coopératif et de partage de ressources dans le domaine des humanités numériques. Le financement des programmes d'archivage numérique pérenne, dont les coûts ont été minimisés par l'adoption de modèles structurels à large échelle (comme l'OAIS), devrait bénéficier d'une politique affirmée de soutien à la recherche publique et la préservation du patrimoine.

Références

[1] Bel, B.; Blache, P. (2006). Le Centre de Ressources pour la Description de l'Oral (CRDO). *Travaux interdisciplinaires du Laboratoire parole et langage d'Aix-en-*Provence (TIPA), 25, pp. 13-18. hal.archives-ouvertes.fr/hal-00142931
[2] Michailovsky, B. ; Michaud, A. ; Guillaume, S. (2011). A simple architecture for the fine-grained documentation of endangered languages: the LACITO multimedia archive. *International Conference on Speech Database and Assessments* (Oriental COCOSDA 2011), Hsinchu: Taiwan. halshs.archives-ouvertes.fr/halshs-00620893

[3] Bärring, O. (2008). *Hosting of IT services and data for Human and Social Sciences in France.* A preliminary study for TGE Adonis (Contract Nr K1432). www.sldr.org/docs/admin/RapportBarring.pdf
[4] CCSDS (2009). *Reference Model for an Open Archival Information System (OAIS)*. Draft Recommended Standard, CCSDS 650.0-P-1.1 (Pink Book) Issue 1.1 August.
[5] Campbell, L. (2012). What might the future be for international collaboration in digital scholarship and preservation? Conférence *Cultural Heritage on Line, Trusted Digital Repositories & Trusted Professionnels*. Florence, 11-12 décembre. (À paraître)
[6] Habert, B. ; Huc, C. (2010). Building together digital archives for research in social sciences and humanities. *Social Science Information* 49, 3:415-443. hal.archives-ouvertes.fr/hal-00466352_v1
[7] Hedstrom, M. (2012). Digital Data Curation – Workforce demand and educational needs for digital data curators. Conférence *Cultural Heritage on Line, Trusted Digital Repositories & Trusted Professionnels*. Florence, 11-12 décembre. (À paraître)