

# TGE Adonis – Projet d’archivage des données produites en France par les SHS

Projet pilote sur les données orales, novembre 2008 – avril 2009

Rapport d’expertise - Version préliminaire du résumé opérationnel  
28 mai 2009

Yves MARCOUX, Ph.D.

Professeur agrégé  
EBSI, Université de Montréal, Canada  
<<http://mapageweb.umontreal.ca/marcoux>>

Consultant principal, Marcoux Médias  
<<http://marcouxmedias.com>>

---

## Résumé opérationnel

### Introduction

La présente expertise porte sur le projet pilote sur les données orales ayant reçu l’aval du Comité de pilotage d’Adonis en mars 2008 et réalisé de novembre 2008 à avril 2009 sous la maîtrise d’ouvrage du TGE-Adonis par le CRDO-Paris, le CRDO-Aix-en-Provence, le CINES et le CC-IN2P3. Le projet pilote consiste essentiellement en la mise en place d’une infrastructure d’archivage et d’accès pérennes pour les données orales transitant par les portails métier du CRDO.

Les objectifs de l’expertise, aux termes de la lettre de mission, étaient de :

- Déterminer si le projet pilote donne les garanties nécessaires pour l’archivage pérenne des données orales, à la fois sur le plan technique et sur le plan organisationnel ;
- Déterminer si, et alors dans quelles conditions, ce dispositif peut être étendu à d’autres types de données.

La lettre de mission demandait aussi que le travail réalisé soit resitué dans le contexte de projets similaires sur le plan international et que soient indiqués les enseignements qui peuvent être tirés de ces rapprochements.

### Portée et limite de l’expertise

Bien qu’à strictement parler, l’expertise demandée ne porte que sur le volet archivage pérenne, l’*accès* est indissociable de la problématique d’archivage et nous nous permettrons donc de nous prononcer sur cet aspect également. Nous nous permettrons aussi, en plus d’émettre un avis sur les garanties fournies par l’infrastructure mise en place, de formuler des commentaires et des recommandations quant à la suite éventuelle du projet.

Étant donné les contraintes de temps, mais également la complexité et le dynamisme extrême du domaine de la préservation numérique, l'expertise réalisée ne saurait prétendre à aucune forme d'exhaustivité. Nous sommes néanmoins confiant qu'elle aidera les décideurs du TGE-Adonis à déterminer la suite à donner au projet pilote.

Comme sources d'information sur le projet pilote lui-même, le wiki mis sur pied pour le projet <<http://www.tge-adonis.fr/wiki/>> a été utilisé, complété par quelques documents (ou fichiers) communiqués sur demande par les différents intervenants, de même que plusieurs communications personnelles. Pour fins de comparaison à l'international, nous avons surtout considéré la « Syndicated Storage Platform » de la *Data Preservation Alliance for the Social Sciences* (Data-PASS) <<http://www.icpsr.umich.edu/DATAPASS/syndicated-storage.html>> et un projet d'infrastructure de préservation numérique pour le consortium Synergies <<http://synergiescanada.org>> au Canada [Beaudry-2009], mais également d'autres programmes, projets et initiatives, dont par exemple le *JISC Digital Preservation & Curation* britannique <<http://www.jisc.ac.uk/whatwedo/topics/digitalpreservation.aspx>> et, aux États-Unis, *Lots of Copies Keep Stuff Safe* (LOCKSS) <<http://www.lockss.org/lockss/>>, basé à Stanford et le *National Digital Information Infrastructure and Preservation Program* (NDIIPP) <<http://www.digitalpreservation.gov/>>, coordonné par la Library of Congress. Une comparaison point par point avec ces initiatives dépasse cependant la portée de la présente expertise.

Bien qu'un des bénéfices attendus de l'infrastructure mise sur pied soit, à long terme, des économies d'échelle par la mise en commun des moyens et des ressources humaines compétentes [vidéo-mi-mai, d.10], nous n'avons pas évalué le projet pilote sous cet angle, ni essayé de prévoir l'ampleur de telles économies.

Les mesures recommandées ci-après nous semblent nécessaires au maintien des garanties de pérennité d'archivage et d'accès, mais nous n'avons pas la possibilité de chiffrer le coût de leur réalisation. Des analyses coûts/bénéfices plus poussées pourraient être nécessaires dans certains cas.

## Conclusions et recommandations

Compte tenu (1) de la conformité du projet aux plus importantes normes dans le domaine de la préservation numérique, (2) du haut niveau d'expertise et de compétence des équipes impliquées dans le projet, (3) de leur très bon fonctionnement collaboratif, (4) de l'existence d'énoncés et d'ententes clairs concernant le partage des responsabilités des différentes parties prenantes, (5) des caractéristiques propres de l'infrastructure d'archivage pérenne mise en place dans le projet et (6) des caractéristiques des données visées elles-mêmes (volumétrie, rythme de production, etc.), **nous sommes d'avis que les garanties de préservation à long terme offertes par l'infrastructure mise en place sont actuellement adéquates.**

Cependant, dans le but d'assurer le maintien de ces garanties dans le futur, nous recommandons que les mesures suivantes soient prises à court terme :

### Recommandation 1

Implanter le plus tôt possible la réplique mutuelle des données actuellement à l'étude entre le CINES et un autre site distant ayant vocation de préservation à long terme (pour l'instant, la BnF). Dans l'infrastructure actuelle du projet pilote, le nombre de copies des données est de quatre (CRDO, CINES + backup distant de moins d'un km, CC-IN2P3), dont une seule à vocation expresse de préservation à long terme (celle du CINES). À titre comparatif, l'approche LOCKSS recommande un minimum de six

copies pour assurer la préservation des données (en supposant cependant des dépôts vulnérables et des « adversaires » disposant de ressources considérables).<sup>1</sup> Malgré que, dans le cas présent, les dépôts puissent être considérés sécurisés, il nous semble que deux copies destinées à la préservation à long terme soit un strict minimum à viser à très court terme.

## **Recommandation 2**

Prévoir l'acquisition par le CINES et l'IN2P3, *avec des ressources internes* (par opposition aux CDD), de toute la connaissance et toute l'expertise nécessaires au maintien en fonctionnement, à la maintenance et aux évolutions futures éventuelles de l'infrastructure.

En prévision d'une montée en charge menant à la prise en charge de l'ensemble des données du CRDO, puis des données d'autres CRN (incluant Archéovision), nous recommandons que les mesures suivantes soient prises à moyen terme :

## **Recommandation 3**

*(Cette recommandation recoupe en partie la Recommandation 2.)* Compléter, consolider et systématiser toute la documentation de l'infrastructure mise en place, tant technique qu'administrative et organisationnelle, de façon à ce qu'elle soit (1) plus générique (on trouve actuellement plusieurs mentions du CRDO), (2) mieux délimitée et (3) plus facilement navigable. Actuellement, la documentation est répartie sur plusieurs documents du wiki (parfois des comptes-rendus) et est par endroits incomplète.

En particulier, certaines règles ou conventions d'écriture des métadonnées semblent se vérifier empiriquement – par exemple « NR » ou « Non applicable » en cas d'information non disponible pour une métadonnée obligatoire – mais ne sont simplement pas documentées (en tout cas pas sur le wiki). De façon générale, les règles d'écriture des métadonnées devraient être décrites plus précisément, avec plusieurs exemples. Cela contribuerait à minimiser la dispersion des formes inscrites (dont « NR » et « Non applicable » sont un exemple) et, partant, à améliorer la recherche en aval.

Il pourrait être opportun d'offrir des formations ciblées, en plus de la documentation, aux ingénieurs des CRN responsables des interactions avec l'infrastructure de préservation et d'accès.

## **Recommandation 4**

*(Cette recommandation développe un point particulier de la Recommandation 3.)* Inclure dans la documentation de l'infrastructure une justification des choix technologiques et organisationnels retenus. Ce point est important, d'une part pour guider l'évolution future de l'infrastructure (pour entre autres ne pas avoir à reprendre les mêmes décisions plusieurs fois), et d'autre part pour assurer à long terme la confiance en l'infrastructure d'éventuels partenaires et utilisateurs, confiance qui peut être une évidence

---

<sup>1</sup> Le nombre minimum de copies recommandé varie entre six et sept, selon les sources. Voir par exemple <[http://www.metaarchive.org/ddp/chapters/fenton-2\\_1.doc](http://www.metaarchive.org/ddp/chapters/fenton-2_1.doc)> et <[http://www.lockss.org/lockss/Selecting\\_and\\_Building\\_the\\_Collection](http://www.lockss.org/lockss/Selecting_and_Building_the_Collection)>.

au départ avec des équipes spécifiques, mais peut facilement s'effriter quelques « générations » plus tard.

En particulier, il importe de situer l'architecture retenue par rapport à celle des réseaux LOCKSS privés <[http://www.lockss.org/lockss/Private\\_LOCKSS\\_Networks](http://www.lockss.org/lockss/Private_LOCKSS_Networks)>, utilisée notamment par LOCKSS, CLOCKSS, Data-PASS et [Beaudry-2009], et le profil de métadonnées pour l'archivage retenu par rapport à la norme PREMIS (Preservation Metadata: Implementation Strategies) <<http://www.loc.gov/standards/premis/v2/premis-2-0.pdf>>, mise de l'avant par Library of Congress. Tant l'approche LOCKSS que PREMIS ont été développés de façon entièrement cohérente avec le modèle de référence OAIS.

Il nous semble également important d'expliquer le choix d'un logiciel commercial (Arcsys) pour la gestion de l'archive au CINES et, surtout, de présenter les solutions de repli prévues en cas de cessation ou de modification importante (temporaire, définitive, volontaire ou involontaire) des activités du fournisseur.

En ce qui concerne l'accès aux données archivées, nous recommandons que les mesures suivantes soient prises à moyen terme :

#### **Recommandation 5**

De façon à ce que le plein accès aux données archivées soit possible selon le modèle de fonctionnement prévu à l'origine du projet pilote [[CH-doc-synthèse-12-mars](#)], réaliser les tâches identifiées comme « restant à réaliser » dans le rapport d'avancement du 2 avril 2009 [[CH-avancement-2-avril](#)]. Un ordre de priorité entre ces tâches est suggéré plus loin dans le présent rapport.

#### **Recommandation 6**

De façon à pérenniser l'accès aux données dans toute la mesure du possible, étudier la possibilité que les volets essentiels du paramétrage et de la programmation de Fedora et des portails métiers soient eux-mêmes mis en paquets et versés au CINES pour préservation à long terme (éventuellement avec un statut spécial). Compte tenu de l'obsolescence rapide des plateformes logicielles, cette pérennisation est toute relative; cependant, cela nous semble être une mesure peu coûteuse propre à assurer des garanties d'accessibilité intéressantes.

Finalement, pour l'ensemble des développements relatifs à l'infrastructure, y compris les recommandations qui précèdent, nous recommandons de :

#### **Recommandation 7**

Utiliser les critères et la liste de vérification présentés dans le document *Trustworthy Repositories Audit & Certification: Criteria and Checklist* (TRAC) <<http://www.crl.edu/PDF/trac.pdf>>, élaborés par OCLC et CRL (The Center for Research Libraries), pour orienter les choix et évaluer périodiquement l'infrastructure. Les outils proposés sont entièrement cohérents avec le modèle de référence OAIS.

Un projet de norme ISO pour l'audit et la certification des dépôts numériques basé sur TRAC est actuellement en cours, sous l'impulsion du *Consultative Committee for Space Data Systems* (CCSDS). Les documents relatifs à ces développements peuvent être consultés au [<http://wiki.digitalrepositoryauditandcertification.org/>](http://wiki.digitalrepositoryauditandcertification.org/).

---

## Sources d'information

Les principales sources d'information pour réaliser l'expertise ont été les suivantes:

1. [Bärring-2008] BÄRRING, Olof. *Hosting of IT services and data for Human and Social Sciences in France*, 31 janvier 2008.
2. [Beaudry-2009] BEAUDRY, Guylaine (Directrice Centre d'édition numérique - Université de Montréal). Communication personnelle.
3. [CH-doc-synthèse-12-mars] HUC, Claude; et al. *Mutualisation de la pérennisation et de l'accès au données en SHS : Projet pilote sur les données orales, Version 1.3*, 12 mars 2009. En ligne <http://www.tge-adonis.fr/wiki/uploads/9/9d/ArchivageMutualise-document-synthese-v1.3.pdf>.
4. [CH-avancement-2-avril] HUC, Claude; et al. *Mutualisation de la pérennisation et de l'accès au données en SHS : Rapport d'avancement au 02/04/2009 du projet pilote sur les données orales*, 2 avril 2009. En ligne <http://www.tge-adonis.fr/wiki/uploads/9/96/EvaluationProjetPilote-rapport-avancement-v9.pdf>.
5. [vidéo-mi-mai] *Collectif*. Vidéo de présentation du projet pilote pour fins d'évaluation, mi-mai 2009. En ligne [http://nicolas.risc.cnrs.fr/bricolage/flashvideo/presentation\\_adonis.htm](http://nicolas.risc.cnrs.fr/bricolage/flashvideo/presentation_adonis.htm); diapositives au <http://www.tge-adonis.fr/wiki/uploads/7/77/Presentation-evaluation-v7.ppt>.

Deux conférences téléphoniques ont été tenues les 19 et 27 mai 2009, avec les participants suivants:

- Bernard Bel, LPL, pour CRDO Aix
- Pascal Calvat, ingénieur, CC-IN2P3 (27 mai seulement)
- Pascal Dugénie, CINES
- Benoît Habert, directeur adjoint sortant, TGE-Adonis
- Claude Huc, consultant, TGE-Adonis
- Michel Jacobson, CRDO Paris (27 mai seulement)
- Thomas Kachelhoffer, ingénieur, CC-IN2P3
- Nicolas Larrousse, ingénieur RICS, pour CRDO Paris (27 mai seulement)
- Olivier Rouchon, CINES, responsable PAC

Nous avons également pu contacter directement les différents participants au projet et poser des questions.