Plan Overview

A Data Management Plan created using DMPonline

Title: CLSR ARCHIVES NUMERIQUES

Creator: Nam Pham

Principal Investigator: Nam Pham

Data Manager: Nam Pham

Project Administrator: Nam Pham

Contributor: Stéphane Pétermann

Affiliation: Université de Lausanne

Template: DMP UNIL en français

ORCID iD: 0009-0007-9635-7491

Project abstract:

Ce projet vise à assurer la **conservation pérenne** des archives **nativement numériques** du CLSR (fichiers textuels, iconographiques, sonores et audiovisuels). La structure de stockage repose sur le modèle OAIS, distinguant le **SIP** (données dans leur format d'origine) de l'**AIP** (données converties en formats de préservation). Pour garantir une sécurité maximale, chaque unité documentaire est encapsulée dans un **"BagIt"**, permettant un contrôle systématique de l'intégrité des fichiers.

ID: 193623

Start date: 05-01-2026

End date: 01-01-2100

Last modified: 06-01-2026

CLSR ARCHIVES NUMERIQUES

1. DESCRIPTION DES DONNEES [cette section se rapporte à la question 1.1 du DMP mySNF]

Des données seront collectées, étudiées, générées ou réutilisées dans le cadre de ce projet :

- Oui

Votre projet implique-t-il la réutilisation de données existantes (produites par vous ou des tiers) ?

- Oui

D’où proviennent les données existantes réutilisées, que contiennent-elles et quel est leur format ?

Les fonds sont constitués de versements effectués par les **producteur·rices** (auteur·rices, collectivités ou institutions) auprès du CLSR. Ces archives, nativement numériques, sont intégrées au répertoire **"01_ARCHIVAGE_NUMERIQUE"** après une phase de tri, de nettoyage et de renommage normalisé.

Bien que le corpus soit majoritairement composé de documents textuels et iconographiques, il est structuré pour accueillir des flux spécifiques, notamment les **courriels**, qui représentent une part croissante des archives contemporaines. Ces ressources sont destinées à être exploitées par la **communauté scientifique**, les professionnel·les de l’édition, les journalistes ainsi que par le grand public.

Indiquez le type des données produites dans le cadre du projet (nouvelles données) :

- Bases de données
- Autre (à préciser)
- Textuelles
- Vidéo
- Audio
- Tableurs
- Images

Le dossier pourrait accueillir des copies images d'ordinateurs, tablettes ou smartphone.

Décrivez brièvement le contenu des données produites ainsi que leur format :

Ces archives constituent une source précieuse sur l'activité littéraire et les engagements (socio-professionnels, politiques) des **producteur·rices**, ainsi que sur leur sphère privée. Pour les fonds collectifs, les données documentent l'histoire administrative et le déploiement opérationnel des entités concernées. Le corpus se compose majoritairement de documents textuels et iconographiques (**docx, doc, jpg, png, tiff, mp3, mp4**), bien que certains fonds puissent inclure des **formats propriétaires** spécifiques.

2. METHODOLOGIE ADOPTEE ET TRAITEMENT DES DONNEES [cette section se rapporte à la question 1.2 du DMP mySNF]

Décrivez comment les données seront collectées et/ou produites dans le cadre de votre projet :

Les archives intègrent les collections du CLSR sous forme de **dons** ou de **dépôts**. Les données font l'objet d'une **extraction** depuis divers supports physiques (clés USB, CD-ROM, disques durs externes, etc.) afin d'être traitées et conservées.

Décrivez la manière dont les données seront traitées, étudiées et analysées dans le cadre de votre projet :

Le traitement des données suit un protocole rigoureux : elles sont **analysées**, **décrites** automatiquement, **renommées** selon les cotes attribuées et, si nécessaire, **réorganisées** avant d'être **converties** dans des formats pérennes. Les doublons sont identifiés par voie logicielle ; leur élimination est soumise à la validation conjointe de la direction du CLSR et de l'ayant droit.

3. ORGANISATION ET REGLES DE NOMMAGE [cette section se rapporte à la question 1.2 du DMP mySNF]

Décrivez les règles de gestion et l'organisation des dossiers et des fichiers de données adoptée (arborescence classificatoire) :

Structure
Les archives sont organisées par fonds et réparties selon deux types de paquets d'information, conformément au modèle **OAIS** :

- **L'AIP (Archival Information Package) :** Regroupe les fichiers convertis dans des formats de conservation pérennes (tels que PDF/A, TIFF ou WAV) afin d'en garantir la lecture à long terme.
- **Le SIP (Submission Information Package) :** Conserve les fichiers dans leur format d'origine, tel qu'ils ont été reçus lors du versement.

Nomenclature et Interopérabilité

Au sein de ces répertoires, les sous-dossiers sont nommés selon la **cote archivistique**. Cette nomenclature assure une correspondance directe avec l'outil de description **Phœbus**, qui centralise l'inventaire des fonds du CLSR.

Intégrité et Sécurité des Données (BagIt)

Chaque dossier identifié par une cote complète (ex: P022-A-5-VIE) est structuré selon la spécification **BagIt**. Cette organisation comprend :

- Un répertoire data/ contenant les objets numériques.
- Des fichiers manifestes (.txt) générant des empreintes numériques via l'algorithme **SHA-256**.

Ce dispositif permet d'effectuer des contrôles d'intégrité réguliers, garantissant que les fichiers n'ont subi aucune altération ou corruption au fil du temps.

Décrivez les règles de nommage des fichiers adoptées (nomenclature) :

Chaque dossier principal contient des sous-dossiers nommés d'après la **cote de l'archive**. Cette nomenclature assure le lien avec l'outil **Phœbus**, où sont répertoriés et décrits les fonds conservés au CLSR.

Par exemple :

- BOULANGER_Mousse_P022
  - AIP_BOULANGER_Mousse_P022
    - P022-A-5-VIE
    - P022-C-2-b-BIB
  - SIP_BOULANGER_Mousse_P022
    - P022-A-5-VIE
    - P022-C-2-b-BIB

4. DOCUMENTATION ET METADONNEES [cette section se rapporte à la question 1.3 du DMP mySNF]

Allez-vous rédiger de la documentation qui accompagne les données ?

- Oui

Sous quelle forme envisagez-vous de rédiger la documentation qui accompagne les données ?

- Un fichier texte (.docx, .odt, .pdf)

Quelles informations cette documentation fournira-t-elle ?

- La table de codage
- Les logiciels utilisés (en mentionnant quelle version, ainsi que les paramètres mis en place)
- Les instruments utilisés et les manipulations effectuées
- Les procédures d’analyse des données
- Les formats et les types de données
- La licence de réutilisation des données
- Les standards et les variables utilisés (noms, questions, descriptions, algorithmes, syntaxes, etc.)

La structure des données, le renommage, les procédure de conversion, les scripts python

Indiquez quelles métadonnées ou standards de métadonnées accompagneront les données :

- Je ne sais pas encore

Dans la perspective du déploiement d’un **SAE (Système d’Archivage Électronique)**, nous envisageons d'adopter des standards de métadonnées reconnus tels que le **Dublin Core** ou l'**EAD**.

5. STOCKAGE [cette section se rapporte à la question 3.1 du DMP mySNF]

Veuillez fournir une estimation du volume nécessaire pour stocker les données :

- De 501Go à 1023Go

Quelles infrastructures de stockage utiliserez-vous ?

- Infrastructure de stockage de la Division calcul et soutien à la recherche (DCSR)

6. SAUVEGARDE [cette section se rapporte à la question 3.1 du DMP mySNF]

Par quels moyens les données seront-elles sauvegardées ?

- NAS DCSR (double copie)

7. QUESTIONS ETHIQUES ET LEGALES [cette section se rapporte à la question 2.1 du DMP mySNF]

Veuillez cocher les questions d’ordre éthique soulevées par votre projet :

- Mon projet ne soulève aucune question d’ordre éthique

Votre projet implique-t-il la récolte ou le traitement de données personnelles et/ou sensibles ?

- Oui

8. SECURITE DES DONNEES [cette section se rapporte à la question 2.2 du DMP mySNF]

Décrivez les mesures mises en place pour assurer la sécurité informatique des données tout au long du projet :

La gestion des données est soumise à des droits d'accès différenciés : seul l'archiviste du CLSR est habilité à les **modifier** ou les **supprimer**. Les autres collaborateurs disposent exclusivement de droits de **consultation et de téléchargement**. Par ailleurs, la sécurité informatique du système est **déléguée** à la DCSR.

Décrivez les mesures de sécurité supplémentaires mises en place pour assurer les impératifs de protection des données personnelles et/ou sensibles :

Les descriptions (métadonnées) sont exemptes de toute **donnée à caractère personnel**. Quant aux données sensibles, elles sont placées sous **"accès restreint"** : leur consultation ou exploitation est soumise à la double autorisation de la direction du CLSR et des ayants droit.

9. PROPRIETE INTELLECTUELLE [cette section se rapporte à la question 2.3 du DMP mySNF]

Sous quelles conditions la réutilisation des données créées par des tiers est possible ?

Les données étant généralement régies par le **droit d’auteur**, leur exploitation requiert l'autorisation préalable des **ayants droit**. À cet effet, le CLSR assure l'interface entre les **demandeur·euses** et les titulaires de droits.

Les données créées dans le cadre de votre projet seront-elles soumises à des restrictions en lien avec les brevets ou les inventions, ou encore à un contrat ?

- Non

10. CONSERVATION A LONG TERME [cette section se rapporte à la question 3.2 du DMP mySNF]

Selon votre estimation, quelle proportion des données de recherche issues de votre projet est destinée à être conservée à long terme (plus de 10 ans) :

- Toutes les données

Sur quels critères de sélection se base cette estimation ?

Ces archives numériques revêtent une dimension **patrimoniale**. À ce titre, leur conservation est garantie de manière **pérenne**, pour toute la durée de vie de l'institution.

Quels formats d’archivage seront utilisés ?

- Texte : XML, PDF/A, HTML, ASCII, UTF-8
- Sons : WAVE, AIFF, MP3, MXF
- Archive web : WARC
- Containers : TAR, GZIP, ZIP
- Données tabulaires : CSV
- Films : MOV, MPEG, AVI, MXF
- Images: TIFF, JPEG 2000, PDF, PNG, GIF, BMP

Il s'agit en majorité de Texte et d'Images.

11. PARTAGE ET DEPOT DES DONNEES [cette section se rapporte aux questions 4.1 et 4.2 du DMP mySNF]

Le partage des données de votre projet de recherche est-il soumis à des restrictions d’accès (embargo, partage sur demande, données qui ne peuvent pas être partagées en raison de clauses légales, éthiques, etc.) ?

- Oui

Précisez ces restrictions et justifiez-les :

La majeure partie de ces archives est régie par le **droit d'auteur**, rendant leur exploitation conditionnée à l'accord préalable des **ayants droit**. Dans ce cadre, le CLSR agira en tant qu'**intermédiaire** pour faciliter les démarches entre les demandeurs et les titulaires des droits.

Sous quelle licence les données seront-elles mises à disposition ?

- Je ne sais pas encore

Le CLSR ne détient pas les droits de la majorité des archives qui sera mise en ligne, c'est pourquoi l'utilisation de licence Creative Commons n'est pas conseillée, hormis pour les documents tombés dans le domaine public (PDM).
A la place, nous utiliserons les RightsStatements qui proposent des "déclarations des droits distinctes dont les institutions de gestion du patrimoine culturel peuvent se servir pour informer le public sur le statut des objets numériques en ce qui concerne le droit d'auteur et leur réutilisation." (https://rightsstatements.org/page/1.0/?language=fr)
Voici les déclarations RightsStatements et CC qu'on pourrait utiliser :

InC : In Copyright [protégé par le droit d'auteur]
InC-RUU : Unknown Rightsholder [protégé par le droit d'auteur - titulaire(s) des droits impossible(s) à localiser ou à identifier]
PDM : Public Domain Mark [marque du domaine public]
CNE : Copyright Not Evaluated [droit d'auteur non évalué]
CC-BY 4.0 : Creative Commons - Attribution

Pour information, cette information reste masquée et sert uniquement à la gestion interne des droits.

Dans quel dépôt (*repository*) les données seront-elles déposées ?

- Je ne vais pas déposer mes données dans un dépôt (expliquez pourquoi)

Contrairement aux projets de recherche clôturés, les archives numériques ne constituent pas des jeux de données figés. Elles doivent être perçues comme des **"données de recherche activables"**, dont le contenu a vocation à être **enrichi et valorisé** de manière continue au fil des années.