Retour d’expérience sur la mise en place d’un inventaire et d’une politique de données

Monica Heintz & Hélène Gautier

C@fés Renatis, 7/04/2022

LESC
Laboratoire d’ethnologie et de sociologie comparative
UMR 7186 • Université Paris Nanterre • CNRS

Anthropen
Les frontières des données anthropologiques

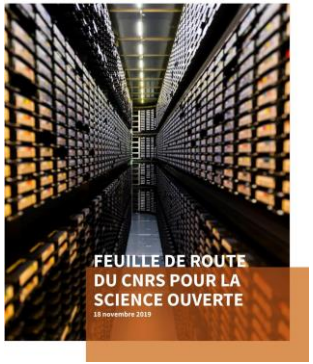
Présentation du Lesc

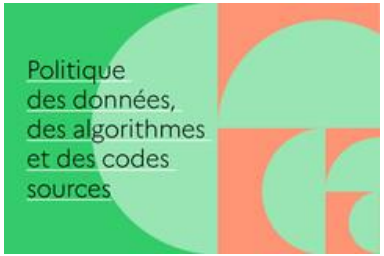
- Laboratoire d’anthropologie non spécialisé dans une aire géographique
- Environ 75 chercheurs et enseignants-chercheurs et 60 doctorants
- Pôle documentaire regroupant les ingénieurs en charge des ressources documentaires:
    - 2 ingénieurs de la Bibliothèque Eric de Dampierre (archives chercheurs + thesaurus GeoEthno)
    - Responsable du service audiovisuel (vidéos)
    - Chargée des ressources documentaires du CREM (Archives sonores du CNRS-Musée de l’Homme)
    - Chargée d’édition de corpus numériques

Des politiques en faveur de la science ouverte à plusieurs échelles

Variété des politiques et des niveaux:

- organisations internationales et nationales
- agences de financement
- universités et institutions de recherche
- revues



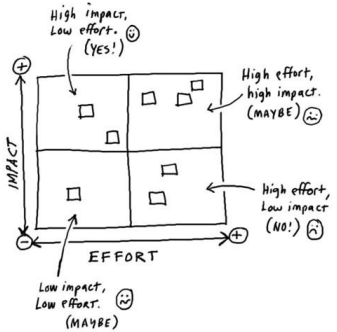


Une expérience dans le cadre d’ateliers de l’InSHS

- Cycle d’ateliers animé par Lionel Maurel et Emmanuelle Morlock
- Objectif: rédiger un document résumant les choix du laboratoire en matière de science ouverte, de gestion et de partage des données ainsi que les responsabilités des différents acteurs impliqués
- Plusieurs unités de recherche dont le Lesc (UMR 7186, Nanterre)
- Binôme chercheur-ingénieur
- Travail entre les ateliers et en collaboration avec d’autres collègues (notamment ingénieurs)

Une première étape: l’inventaire des données

- Recensement (non exhaustif) des jeux de données
  - Avoir une idée des types de données, des volumétries, de leur localisation, etc.
  - Inclure tout ce sur quoi le laboratoire investit des ressources (matérielles et/ou humaines) ou qui est pris en charge par un membre du laboratoire
- Outil pour la FAIRisation
  - Matrice impact/effort
- Aide à la rédaction de la politique
  - Identifier les priorités
  - Définir le périmètre des “données” pour le laboratoire



Identification du jeu					Caractérisation des données produites					Informations complémentaires				Modes de diffusion / archivage				Modalités de partage	
Programme de recherche	Désignation du jeu	Année de collecte ou de production	Nom du chercheur principal ayant collecté ou produit les données	Personne ou service en charge de la gestion des données	Caractérisation des données produites	État de finalisation (en cours/ publié)	Format(s)	Taille (nombre de fichiers, espace disque)	Données brutes ou primaires (oui /non)	Type	Observations particulières (origine des données, réutilisations potentielles, etc.)	Plan de gestion de données (oui/non/en cours)	Zones géographiques couvertes	Serveur et modalités de stockage / sauvegarde	Diffusion via le projet de publication numérique	Partage / archivage comme jeu de données réutilisables	Référencement/indexation par des méta-moteurs, des portails, etc.	Périmètre du partage potentiel	Licence attribuée
Projet Inscriptions	Inscriptions - transcriptions TEI	2020		IR pôle ingénierie	Transcription narrative (inscription) et métadonnées descriptives des objets support (commentaires de l'épigraphiste en prose)	en cours	XML	1000 fichiers, moins de 2 Mo	non	données textuelles	Les échanges TEI des données textuelles et des métadonnées apportent un potentiel de réutilisation et de valorisation dans la communauté des épigraphistes et en dehors potentiellement très large.	ct PGD (projet en cours)		Huma-Num Sharedocs	Huma-num (hébergement web)	Zenodo / Huma-num Nakala (prévu)		Tout le monde	CC BY 4.0
	Inscriptions images	2019		IR pôle ingénierie	collection d'images	en cours	TIFF, JPEG	1000 fichiers, plusieurs Go de fichiers image				ct PGD (projet en cours)		Huma-num (hébergement web)	Huma-num (hébergement web)	Nakala si autorisation confirmée		Toutes communautés scientifiques	cf. Louvre-RMN pour les images du Louvre / CC BY pour les images dont les chercheurs sont auteurs.
	Inscription système de publication	2020		IR pôle ingénierie	application web personnalisée	en cours	XQuery, CSS, XSLT, XML	moins de 200 Mo	non	codes informatiques, programmes		ct PGD (projet en cours)		Huma-num (gitlab)	Huma-num (hébergement web, gitlab)	Huma-num (gitlab), Zenodo		Tout le monde	© Huma-num pour les codes modifiant TEI publisher.
Plateforme Halmeda	Plateforme Halmeda	2017		IE développement d'applications, IR resp. documentaire	Photographies de terrain et reproduction de documents d'archives pour des projets de recherche sur les villes de Tunis, Alger et Oran et métadonnées descriptives associées.	publié	JPEG	1224 fichiers		Images fixes	les fichiers et métadonnées publiés sur Halmeda sont pour une part importés de Lightroom.			Huma-num (hébergement web	Huma-num (hébergement web)	Huma-num (hébergement web)	si Libre	Tout le monde	CC-BYNC et sous droit pour les images. CC-BY pour les métadonnées.

https://bit.ly/3KdR05l

L’inventaire de données au Lesc en pratique

- Adaptation du modèle proposé avec ajout de champs
- Sollicitation de l’ensemble des membres du laboratoire
  - Remplir l’inventaire ne constitue pas un engagement à ouvrir ses données
  - Informations de nature documentaire et technique
  - Aucune colonne obligatoire, donc aussi possible de compléter si toutes les informations ne sont pas disponibles
  - Possibilité d’avoir plusieurs lignes pour un même projet/jeu de données (selon le type de données, le périmètre de partage, l’état de finalisation, le lieu de stockage, etc.)
- Question de la granularité pertinente

L’inventaire de données, bilan

- Données surtout pensées en termes de projet (ANR) ou de terrain, non numériques et “passives”
- Certaines colonnes peu remplies: périmètre de partage, licence
- Côté ingénieurs, prise de connaissance de jeux de données
- Côté chercheurs, prise de conscience de l’anticipation nécessaire des questions de gestion et de partage des données

Politique de données vs DMP

Politique de données

- Niveau: laboratoire
- Porte sur ce qui relève de la science ouverte (les données mais pas que!)
- Rubriques axées sur la science ouverte
- Communique une vision stratégique
- Révisable mais vocation à durer dans le temps

Data Management Plan (DMP)

- Niveau: projet
- Porte sur les données
- Rubriques axées sur le cycle de vie des données
- Outil de planification/gestion
- Document évolutif (a minima 3 versions)

Un modèle de politique à adapter

- Proposition du modèle OpenAIRE (2018) pour les infrastructures de recherche traduit en français et adapté par les animateurs de l'atelier: https://bit.ly/3x3TXlf
- Sections
  - Préambule
  - Périmètre d'application de la politique
  - Droits, obligations et responsabilités (laboratoire et membres)
  - Accès ouvert aux publications
  - Accès ouvert aux données de la recherche
  - Science ouverte
  - Infrastructure et services d'appui
  - Evaluation des travaux de recherche et des chercheurs
  - Formation
  - Application de la politique
  - Annexes

La politique de données en pratique

- Des questions
  - sur les tutelles (ont-elles des politiques? quels sont leurs engagements? sous quelle forme?...)
  - sur l'environnement (y a-t-il des services dédiés dans l'environnement du laboratoire (MSH, université, etc.)?)
  - sur l'unité (les questions de science ouverte font-elles partie du plan quinquennal? comment sont gérées les données en interne?)
  - sur les pratiques (départ à la retraite, changement d'unité: que deviennent les données? Qui prend les décisions relatives aux données? Quand?)
- Peu de modifications mais ajout d'informations concrètes (personne ou service ressource sur telle question)
- Choix de ne pas inclure de mesures contraignantes

Bilan

- Outil le plus utile actuellement (et le plus simple à mettre en oeuvre): l’inventaire de données
  - Réflexions sur sa poursuite et sa mise à jour régulière

- Un point d’entrée particulièrement pertinent pour les anthropologues: la mission