Traçabilité des données d'expérience pour les matériaux anciens et patrimoniaux

Laurent Romary, Inria
Projet Dopamine

Atelier Gérer les données dans des projets étudiant les matériaux anciens et patrimoniaux
EPHE, Paris, 7 février 2019, 9h15 – 17h00

Le numérique au cœur des recherches en patrimoine et matériaux anciens

- Place du numérique croissante dans tous les domaines scientifiques
  - Production d'objets numériques à toutes les étapes: de l'observation à la publication
  - La préservation et la réutilisation des données intervenant à toute étape du processus du recherche est un enjeu scientifique, économique et éthique
    - E.g. processus de réfutation (cf. exposé de Louise Le Meillour)
- Place particulièrement importante au sein des domaines couverts par le DIM MAP
  - Variétés des sources, multiplicité des acteurs
- Une première présentation du projet DOPAMINE
  - Concepts, état des lieux, pistes méthodologiques

Pourquoi suis-je ici?

- Recherche en informatique : ancrée sur une démarche pluri-disciplinaire
  - Traitement automatique des langues
  - Modélisation de données en sciences humaines (langue, documents)
- Du modèle au standard : participation à la normalisation internationale
  - *Text Encoding Initiative* : consortium de référence pour la représentation numérique de textes
  - Comité TC 37 de l'ISO (*Langue et terminologie*)
- Contribution aux infrastructures européennes
  - Initiateur (depuis 2006) et directeur (2014-2018) de l'infrastructure DARIAH ERIC
- Une implication (obstinée) dans la science ouverte
  - CNRS, Société Max Planck, Inria; membre du CoSO Tech
  - Déploiement de HAL, développement des centre de ressources numériques (bases des consortiums Huma-Num), obligation de dépôt dans HAL à Inria

Le projet DOPAMINE

- Données Patrimoniales : Méthodes, Infrastructures, Exploitations
- Partenariat
  - IPANEMA – Institut photonique d'analyse non destructive des matériaux anciens
  - DYPAC – Dynamiques patrimoniales et culturelles
  - EPHE – École pratique des hautes études
- Objectifs
  - Mettre en œuvre des méthodes pour améliorer la traçabilité des données d'expérience
  - Dynamique transversale au DIM « Matériaux anciens et patrimoniaux »

Un contexte national et européen « pressant »

- Le Plan national pour la science ouverte – juillet 2018
  - « Rendre obligatoire la diffusion ouverte des données de recherche issues de programmes financés par appels à projets sur fonds publics »
  - « La France recommandera l'adoption de licences ouvertes pour les publications et les données »
- De nombreuses initiatives européenne
  - Publications ouvertes : OpenAire, Plan S
  - Données
    - Obligation de la production d'un plan de gestion de données, pression pour disposer de données « FAIR »
    - RDA, GO FAIR, mise en place d'EOSC
  - Infrastructures européennes de la feuille de route ESFRI: DARIAH, E-RIHS, CLARIN, OPERAS, DiSSCo

Expérience acquise dans le cadre d'Iperion CH

- Iperion CH
  - Integrated Platform for the European Research Infrastructure ON Cultural Heritage
  - INFRAIA-1-2014-2015 - Integrating and opening existing national and regional research infrastructures of European interest
  - 19 équipements, dans 11 pays regroupés au sein de 3 plates-formes: ARCHLAB, FIXLAB and MOLAB
- Enquêtes effectuées au sein de la tâche 2.2 *Management plan of generated digital data*
  - Collaboration avec IPANEMA

Enquête Iperion CH – principaux résultats

- Double enquête sur les pratiques et les jeux de données
  - excellente couverture du consortium
  - 3 plates-formes et 29 instruments, 78 jeux de données mentionnées couvrant une très grande variété de matériaux et de méthodes d'analyse
- Difficultés exprimées par les répondants
  - documentation des données
  - formats standards et réutilisables
  - hébergement informatique pérenne
  - licences associées aux jeux de données

Archives départementales des Yvelines

Conditions de réutilisation des documents conservés aux Archives départementales des Yvelines

Vous êtes libre de réutiliser gratuitement et sans formalités les informations contenues dans les documents conservés aux Archives départementales ou les images de ces documents,

Vous pouvez les :

- reproduire, copier, publier et transmettre ;
- diffuser et redistribuer ;
- adapter, modifier, transformer, notamment pour créer des documents dérivés ;
- exploiter à titre commercial.

La réutilisation est gratuite mais la mise à disposition des informations donne lieu à la perception de frais techniques dans les cas où elle entraîne des opérations techniques (reproduction, extraction de données, compression et transfert de fichiers, …) à la charge du département.

Source: https://archives.yvelines.fr/article.php?larub=12&titre=reutilisation-des-archives

Archives Départementales des Yvelines (Suite)

Sauf s'il s'agit de documents dont vous avez obtenu communication par dérogation, de documents contenant des informations publiques comportant des données à caractère personnel, de documents sur lesquels s'exerce un droit de propriété intellectuelle, ou encore de documents entrés par don ou par dépôt.

…

CAS DES DOCUMENTS COMPRENANT DES DONNÉES À CARACTÈRE PERSONNEL

…

CAS DES DOCUMENTS SUR LESQUELS S'EXERCE UN DROIT DE PROPRIÉTÉ INTELLECTUELLE

…

Et à condition de respecter les conditions suivantes :

- ne pas dénaturer le sens des informations contenues dans les documents ;
- mentionner de manière visible la source des informations et leur lieu de conservation (de préférence sous la forme Archives départementales des Yvelines, précision de la cote) ;
- préciser la date de la production ou de la dernière mise à jour des informations ;
- mentionner le nom de (ou des) auteur(s), s'il y a lieu.

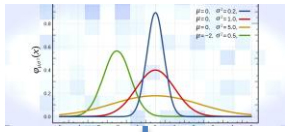
Et le chercheur/la chercheuse dans tout ça ?

- Recentrer le débat sur l'individu et le projet de recherche
  - Les principes FAIR sont trop centrés sur les jeux de données, mais pas sur les acteurs du processus de recherche
  - L'administration des données ne doit pas prendre le pas sur la recherche elle-même
- Identifier les questions qui peuvent se poser sur le terrain
  - Documentation, attribution, hébergement, réutilisation
- Apporter des réponses simples et concrètes
  - Intégrer la gestion des données à la pratique de recherche

Accompagner le processus de recherche – vers un *certificat d'identité des données*

- Vision: agréger à chaque étape tous les éléments pertinents pour la traçabilité d'un jeu de données
  - Acteurs et responsabilités
  - Sources
  - Données techniques, formats
  - Traitements effectués
  - Conditions de diffusion et de réutilisation, licences
- Démarche pragmatique
  - Identifier comment rendre cette gestion naturelle (mais pas transparente) pour la chercheuse ou le chercheur

Tracer la création des jeux de données



Projet de recherche

Jeux de données primaires

Jeux de données secondaires

Publication

Echantillon
Origine
Propriétés

Appareil
Format
Paramètres utilisés

Logiciel
Format
Paramètres utilisés

Littérature secondaire

Littérature primaire

Documentation

Données d'enquête

Et quand on croit être arrivé au bout…

- Hébergement
  - Données, méta-données
  - Identification, archivage à long terme
  - Autorité?
- Référencement (publications)
  - Citer les jeux de données (et sources) primaires et secondaires
  - Citer les différents acteurs du processus
  - Réutiliser les contenus?

Objectifs de la journée

- Initier une interrogation individuelle concernant le circuit de gestion de données
- Recueillir des éléments permettant de compiler des recommandations aux chercheurs
- Avancer ensemble dans la définition d'une implémentation de la charte de ré-utilisation des données

Aborder le système en amont

- De la demande initiale à la réalisation du projet de recherche
- Possible utilisation de SciencesCall
  - Une plate-forme de gestion d'appel à projet
    - Développement à l'UMS CCSD (cf. HAL, Episciences, SciencesConf)
    - Construit suivant un modèle classique de dépôt et évaluation

Pistes pour une utilisation de SciencesCall pour le traçage des données

- Fiche de gestion de données au moment du dépôt
- Intégration d'un point de validation par les acteurs du processus
  - Etablissement patrimonial, équipement, futur hébergeur des données
- Agrégation de métadonnées de la part de ces acteurs

Intégrer tous les acteurs du dispositif

- Le chercheur, la chercheuse
  - Qualifie la recherche et définit le cycle de vie de ses données
- L'établissement patrimonial
  - Détermine les contraintes liées à l'utilisation de ses fonds et les attentes en retour
- L'équipement
  - Détermine une politique d'utilisation de la part des projets de recherche
- L'hébergeur de données
  - Exprime des contraintes sur la taille, les formats, la documentation, les conditions d'accès
- Et bien sûr les politiques européennes, nationales et institutionnelles

Un outil méthodologique

- La charte de réutilisation des données
  - un contrat entre les différents acteurs du processus de recherche
  - fluidifier les échanges et faciliter la réutilisation des données patrimoniales
- Implication d'institutions européennes
  - DARIAH, Europeana, CLARIN, APE, E-RIHS

Les principes fondateurs

Tentative d'implémentation

- Partir des rôles et des exigences variés entre acteurs pour faire des recommandations centrées sur les besoins de chacun, mais profitables à tous (approche *bottom-up*)
- Par exemple, pour la « *citability* » :
  - Pour le chercheur ou la chercheuse → citer l'institution d'où viennent les données et l'équipement / l'infrastructure qui est intervenue au cours du projet
  - Pour l'institution patrimoniale → fournir un modèle de citation et communiquer autour des recherches faites à partir de leurs collections
- Tester ces principes à travers **SciencesCall**, plateforme de gestion des appels à projet

Et pratiquement?

- Priorité au travail avec les chercheurs
  - Doctorants et post-doctorants financés dans le cadre du DIM MAP
    - Séminaire de travail commun le 7 février
    - Echanges personnalisés en mai-juin
- Travailler à une déclaration préalable à intégrer aux demandes de projet sur SciencesCall
  - Intégration des grandes catégories de la charte
- Définir un concept pour intégrer des contraintes multi-acteurs dans ScienceCall
- Vers un guide de gestion des données de recherche à destination des jeunes chercheurs
  - Centré sur les processus de recherche et les types de donnée

Perspectives

- Gestion raisonnée des données
  - Un point de départ pour la transparence scientifique
    - Dans l'espace et le temps
  - Nécessite des compétences et des infrastructures
  - Du temps, de l'argent, des personnels
- Aborder la gestion des données de façon humaine et pragmatique
  - Accompagner le changement en impliquant les chercheurs
  - Permettre une amélioration progressive des conditions de gestion des données de la recherche
  - Mettre en œuvre des solutions pérennes