Séminaire doctoral interdisciplinaire
2017-2018

Les données de la recherche dans les thèses de doctorat (DRTD)

Le séminaire est proposé comme séminaire « outils-méthodes ».

Compétences scientifiques :
À partir de leurs propres données (entretiens, statistiques, textes, images etc.), les doctorant(e)s apprendront à mettre en place un plan de gestion des données. Après un rappel des enjeux et du cycle de vie des données de la recherche, seront abordées successivement les étapes importantes en vue de l'archivage à long terme : la capture des données et leurs formats, la structuration des données, leur description et leur partage.

Compétences transférables :
Préparation et création d'un plan de gestion des données, à l'aide d'un outil en ligne (DMP OPIDoR).

Objectifs et évaluation :
- Réalisation d'un plan de gestion de données complet
- Pertinence par rapport à la thématique (pas de plan standard)
- Cohérence globale du plan
- Prise en compte des principes FAIR

D'une durée de 20 heures, le séminaire a lieu en sept séances entre janvier et juin 2018.
Le séminaire a lieu dans les locaux de la BU sciences humaines et sociales (Pont de Bois).

Responsables du séminaire :
- Bernard Jacquemin, maître de conférences, GERiiCO, Lille 3
- Joachim Schöpfel, maître de conférences, GERiiCO, Lille 3

Autres intervenants :
- Éric Kergosien, maître de conférences, GERiiCO, Lille 3
- Cécile Malleret, conservateur au SCD de l'Université de Lille SHS

Prérequis demandés/public ciblé :
- Tous les doctorants avec un projet de thèse produisant des données de la recherche.
- Connaissances d'anglais souhaitables.

(1) Introduction : Gérer les données de la recherche – pourquoi, comment ?

Mardi 30 janvier 2018 – 14-17 heures (C.Malleret, J.Schöpfel)

Cette séance aura pour but de définir et cerner les enjeux des données de la recherche en distinguant les données brutes, les données dérivées et les jeux de données (*datasets*) et en rappelant le contexte (national et européen) de l'Open Science. Un second temps sera consacré aux pratiques et besoins identifiés des doctorants et des chercheurs de Lille 3 dans la gestion de leurs données de recherche et enfin à l'évaluation des acquis des doctorants sur cette question.

(2) Créer un plan de gestion des données de la recherche

Mardi 20 février 2018 – 14-17 heures (C.Malleret, J.Schöpfel)

Le Plan de Gestion des Données des données (*Data Management Plan*, DMP) ne répond pas seulement à une obligation des financeurs mais se veut d'abord une aide à la conservation et, autant que possible, au partage des données de la recherche. L'objectif est donc d'abord de fournir un cadre qui permette, dans le processus de recherche, d'inclure la production, collecte et curation des données comme « bonne pratique scientifique ». Une partie de la séance sera consacrée à la formation à l'outil en ligne DMPOnline/OPIDoR et au lancement des travaux individuels.

(3) Le cycle de vie des données

Mardi 6 mars 2018 – 14-17 heures (B.Jacquemin, J.Schöpfel)

Les données de la recherche s'inscrivent dans le contexte plus large de la donnée numérique. Aussi est-il nécessaire d'étudier leur cycle de vie, depuis leur création jusqu'à leur archivage définitif, en prenant en compte deux propriétés essentielles que sont leur aspect digital d'une part, et leur lien à une activité de recherche de l'autre. Partant des besoins liés à l'archivage - et notamment l'archivage à long terme, qu'il s'agira d'identifier - nous étudierons donc l'identification et la description des données pour assurer leur (ré)utilisabilité à travers des jeux de métadonnées, les modèles existants pour la conservation et l'archivage des données numériques et les systèmes mis en place qui disposent des fonctionnalités nécessaires à un archivage efficace et pérenne. Un temps sera consacré à faire le point sur les plans de gestion des doctorants, par rapport à l'avancement de leurs propres projets de recherche.

(4) Décrire les données de la recherche

Mardi 3 avril 2018 – 14-17 heures (B.Jacquemin, E.Kergosien)

La description des données est une étape primordiale dans le plan de gestion. En effet, afin que les données de la recherche soient réutilisables, le contexte de leur production doit être documenté de manière précise et intelligible. Ainsi, il peut être décrit par :

- une documentation adéquate, sous la forme d'un fichier txt ou pdf qui rapporte des informations sur le projet (hypothèses, méthodologie, échantillonnage, instruments ...), sur les fichiers ou la base de données et sur les paramètres ;
- et des métadonnées (metadata) : ensemble structuré de données qui servent à définir ou décrire une ressource quel que soit son support. Les métadonnées répondent aux questions suivantes : qui, quoi, où, quand, comment, pourquoi ? Avec les métadonnées, le fournisseur de données apporte aux utilisateurs des informations sur le contexte de production et la qualité de ses données, tandis que l'utilisateur peut découvrir des ressources et évaluer leur pertinence par rapport à ses besoins.

Nous profiterons de cette séance pour traiter les règles de nommage des documents, la notion d'identifiant pérenne pour les données de la recherche et la façon de lier vos données aux publications scientifiques résultantes des travaux scientifiques.

(5) Structurer les données de la recherche

Mardi 17 avril 2018 – 14-17 heures (B.Jacquemin, E.Kergosien)

Afin de faciliter les échanges d'information, il est nécessaire d'utiliser un langage commun pour structurer les données. On parle alors de standards de métadonnées (metadata standard). Il existe différents types de standards de métadonnées : génériques, disciplinaires et technologiques. Nous étudierons le standard Dublin Core défini pour décrire de façon synthétique tout type de contenu et notamment les corpus de textes, les images et les enquêtes.

Nous présenterons le langage XML qui est un langage de balises permettant de décrire et structurer les données de la recherche. Après avoir détaillé quelques exemples de jeux de données structurés dans ce langage, des exercices permettront de mettre en pratique le langage XML et le standard Dublin Core sur des jeux de données de tests.

Nous aborderons les formats descriptifs notamment à travers des exemples de structuration de données, et nous montrerons comment baliser les données. Un TD sera mené sur trois types de données : corpus de textes, d'images, d'enquêtes.

(6) Conserver et partager des données

Mardi 29 mai 2018 – 14-17 heures (C.Malleret, J.Schöpfel)

Nous allons présenter un panorama des sites en ligne pour conserver et partager les données de la recherche, dans les domaines SHS. Nous allons aborder plusieurs aspects : comment trouver ces sites ? Comment déposer des données ? Comment les partager ? Une partie de la séance sera consacrée aux solutions pour les données des doctorants, y compris le dépôt sur Huma-Num via NAKALA (projet *D4Humanities*).

(7) Bilan et évaluation

Mardi 5 juin 2018 – 14-16 heures (B.Jacquemin, J.Schöpfel)

La dernière séance du séminaire sera consacrée à l'évaluation des plans de gestion des doctorants et à un échange avec les participants.