Séminaire doctoral DRTD

« Les données de la recherche dans les thèses de doctorat »

École doctorale SHS - Année 2015 – 2016

Faisant suite au séminaire tenu en 2014-2015 sur « les données de la recherche dans les thèses de doctorat en sciences humaines et sociales » financé par la MESHS (projet Partenariat) financé par la MESHS (projet Partenariat) et au Livre blanc[1] qui en a résulté, le laboratoire GERiiCO (Stéphane Chaudiron) et le SCD de Lille 3 (Isabelle Westeel) organisent un séminaire doctoral méthodologique sur la mise en place d'un plan de gestion des données de la recherche.

Contexte

Dans le mouvement des Humanités numériques, le libre accès aux publications scientifiques et aux données de la recherche est à l'ordre du jour. Les agences de financement (ANR, programme H2020 notamment) exigent de la part des porteurs de projets soumettant une demande de financement que soit explicitement prévu un plan de gestion des données qui seront recueillies et produites dans le cadre de la recherche. Pour répondre à cette exigence, des répertoires de données se mettent en place, en France (Nakala dans le cadre d'Huma-Num) et à l'étranger (par exemple le *DANS* aux Pays-Bas, *HISTAT* en Allemagne dans le domaine des statistiques historiques ou l'*Archeology Data Service* en Grande-Bretagne). Le registre des répertoires de données (re3data.org[2]) identifie plus de 1300 répertoires à ce jour, dans de multiples domaines.

Les enjeux liés à l'archivage et à l'accessibilité des données de la recherche sont nombreux : l'enjeu est d'abord de nature patrimoniale car décrire et archiver de manière pérenne les données permettra une réutilisation future par d'autres chercheurs. L'enjeu est ensuite de nature heuristique car il s'agit de permettre l'exploration des données (corpus textuels ou oraux, données brutes, images…) avec des techniques numériques (*text mining*, classification, cartographie, visualisation…) afin de construire un sens nouveau. Il répond enfin à une exigence en termes de politique scientifique car, imposé dans les programmes de financement, il est devenu un élément clé dans les réponses aux appels à projets.

Objectifs

À partir de trois exemples de données (entretiens, textes et images), 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.
D'une durée de 20 heures, le séminaire est prévu sur 7 séances entre janvier et juin 2016. Le programme prévisionnel est le suivant.

Programme

Séance 1 : Pourquoi gérer les données de la recherche ? (18 janvier 2016, de 14h à 17h, salle B2.468)

Intervenants : Cécile Malleret, Joachim Schöpfel

[1] Séminaire organisé par le laboratoire GERiiCO avec le soutien de l'École doctorale SHS

[2] http://www.re3data.org/

Cette séance aura d'abord 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 (ou *dataset*) 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.

Séance 2 : Créer un plan de gestion des données de la recherche (1 février 2016, de 14h à 17h, salle A1.419)

Intervenants : Cécile Malleret, Joachim Schöpfel

Le Plan de Gestion des Données des données (ou *Data Management Plan)* ne répond pas seulement à une obligation des financeurs mais se veut d'abord une aide à l'archivage 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 et la collecte des données. Une grande partie de la séance sera consacrée à la méthodologie mise en œuvre au travers d'exemples.

Séance 3 : Le cycle de vie des données (14 mars 2016, de 14h à 17h, salle B2.460)

Intervenants : Bernard Jacquemin

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.

Séance 4 : Décrire les données de la recherche (21 avril 2016, de 14h à 17h, salle B2.472)

Intervenants : Bernard Jacquemin, Eric 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, que, 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.

Séance 5 : Structurer les données de la recherche, (23 mai 2016, de 14h à 17h, salle B2.472)

Intervenants : Bernard Jacquemin, Eric 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.

Quels formats descriptifs ?;

Montrer des exemples de structuration de données ;

Comment baliser les données ;

TD sur 3 types de données : corpus de textes, images, enquêtes.

Séance 6 : Partager et réutiliser des données (6 juin 2016, de 14h à 17h, salle B2.472)

Intervenants : Cécile Malleret, Joachim Schöpfel, Sofia Papastamkou

Nous allons présenter un panorama des sites en ligne pour conserver et partager les données de la recherche. Nous allons aborder plusieurs aspects : comment trouver ces sites ? Comment déposer des données ? Comment les partager ? Nous allons parler de différents types de sites (entrepôts), en montrant plusieurs exemples. Nous allons également évoquer quelques aspects éthiques et juridiques du partage, et nous allons finir par quelques alternatives, dont notamment les "*data papers*".

Séance 7 : Bilan du séminaire (date à préciser)

La dernière séance du séminaire sera consacrée à un échange avec les participants au séminaire.

Les intervenants

Stéphane Chaudiron, professeur, GERiiCO, Lille 3
Bernard Jacquemin, maître de conférences, GERiiCO, Lille 3
Éric Kergosien, maître de conférences, GERiiCO, Lille 3
Cécile Malleret, conservateur au Service commun de documentation de Lille 3
Sofia Papastamkou, MESHS Lille Nord de France
Joachim Schöpfel, maître de conférences, GERiiCO, Lille 3
Isabelle Westeel, directrice du Service commun de documentation de Lille 3

Bibliographie indicative

André, F., 2015. Déluge des données de la recherche ? In: Calderan, L., Laurent, P., Lowinger, H., Millet, J. (Eds.), Big data : nouvelles partitions de l'information. Actes du Séminaire IST Inria, octobre 2014. De Boeck; ADBS, Louvain-la-Neuve, pp. 77-95.

COMETS, 2015. Les enjeux éthiques du partage des données scientifiques. Comité éthique du CNRS, Paris.
URL http://www.cnrs.fr/comets/spip.php?article123

European Commission. Guidelines on Data Management in Horizon 2020(dec. 2013)
http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

European Commission. Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020 (dec. 2013)
http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf

Gaillard, R., 2014. De l'Open data à l'Open research data : quelle(s) politique(s) pour les données de recherche ? Master's thesis, ENSSIB, Villeurbanne.
URL http://eprints.rclis.org/22746/

Kindling, M., 2013. Doctoral theses' research data and metadata documentation. In: ETD 2013 Hong Kong 16th International Symposium on Electronic Theses and Dissertations 25 September 2013.
URL http://lib.hku.hk/etd2013/presentation/Maxi-ETD-20130925.pdf

Kuipers, T., van der Hoeven, J., 2009. Insight into digital preservation of research output in europe. survey report. PARSE insight, n/a.
URL http://www.parse-insight.eu/downloads/PARSE-Insight_D3-4_SurveyReport_final_hq.pdf

Les données de la recherche dans les appels à projets Horizon 2020 : Produire un Data Management Plan, Université Paris Diderot, Université Paris Descartes, Université Paris Sorbonne
http://www.isore.cnrs.fr/IMG/pdf/2014_ANF_5-2.pdf

Naegelen, P., 2015. Données de la recherche : quel positionnement et quels rôles pour les bibliothèques ? In: Données en partage : enjeux et acteurs des données de la recherche. URFIST Toulouse, 15 juin 2015.
URL http://fr.slideshare.net/pierrenaegelen/donnes-de-la-recherche-quel-positionnement-et-quels-rles-pour-les-bibliothques

Principes et lignes directrices de l'OCDE pour l'accès aux données de la recherche financée sur fonds publics(2007)
http://www.oecd.org/fr/science/sci-tech/38500823.pdf

Prost, H., Schöpfel, J., 2015. Les données de la recherche en SHS. une enquête à l'Université de lille 3. rapport final. Université de Lille 3, Villeneuve d'Ascq.
URL http://hal.univ-lille3.fr/hal-01198379v1

Special issue on "Data Sharing, Data Publication and Data Citation." *Journal of Librarianship and Scholarly Communications*, Volume 3 - Issue 2, 22 sep 2015
Site : http://jlsc-pub.org/10/volume/3/issue/2/

Tenopir, C., Dalton, E. D., Allard, S., Frame, M., Pjesivac, I., Birch, B., Pollock, D., Dorsett, K., Aug. 2015. Changes in data sharing and data reuse practices and perceptions among scientists worldwide. PLoS ONE 10 (8), e0134826+.
URL http://dx.doi.org/10.1371/journal.pone.0134826