Les données de la recherche dans les thèses de doctorat - Livre blanc

Stéphane Chaudiron, Catherine Maignant, Joachim Schöpfel, Isabelle Westeel

Université de Lille 3 Sciences humaines et sociales

Villeneuve d'Ascq, juin 2015

Les données de la recherche dans les thèses de doctorat

Livre blanc

Valoriser les résultats de la recherche

Science numérique, science ouverte

Relever le défi

« Assurer une maitrise publique des données de la recherche », ainsi commencent les 10 propositions de la CPU pour les *Universités 3.0*. A l'heure où l'Etat français mène une politique exemplaire d'ouverture des données publiques et où les entreprises font face à la révolution *Big Data*, les universités relèvent le défi des données de la recherche. L'enjeu est double : mettre en place des infrastructures sécurisées pour la gestion et la préservation des données scientifiques, et assurer leur diffusion pour stimuler la recherche et l'innovation. L'irruption du numérique bouscule les modes d'apprentissage, l'organisation des universités, la définition du lien social mais avant tout aussi, la façon de faire de la recherche. Impossible aujourd'hui d'imaginer la science autrement que numérique. Elle est devenue ouverte, grâce à l'outil numérique, et elle est devenue participative. La diffusion libre des résultats de la recherche sur Internet renforce l'impact du chercheur et de l'institution. En même temps, elle correspond à un formidable retour sur investissement pour la société civile.

Rattraper du retard

Cependant, pour la circulation des données de la recherche, les universités ont pris du retard par rapport aux organismes, comme le fait remarquer Catherine Rivière, PDG de GENCI : « Les communautés (astrophysique, climat…) affichent des données. Les universités, pas vraiment. » Il est temps, selon les mots de la CPU, d'organiser et de systématiser la mise à disposition des résultats scientifiques et des données brutes de la recherche ; on s'oriente ainsi vers un statut de la donnée ouverte, pour la recherche et pour l'innovation.

Au cœur du dispositif, la thèse

Par où commencer ? Notre proposition est de valoriser le cœur même de l'enseignement et de la recherche universitaire et de construire une démarche « données de la recherche » autour des thèses de doctorat. Pour plusieurs raisons : les thèses appartiennent à l'enseignement supérieur, hors tout circuit commercial, et elles ont, du fait de leur nombre, leur richesse et qualité mais aussi leur représentativité, un grand intérêt pour la veille et l'innovation. Elles sont déjà, pour une grande part, dématérialisées, l'infrastructure numérique est en place ; par ailleurs, il y a une incitation forte à la diffusion en libre accès. Les conditions sont donc réunies pour monter un projet spécifique aux données de la recherche dans les thèses de doctorat. D'autant plus qu'investir dans la formation doctorale, c'est investir dans l'excellence scientifique et la recherche de demain.

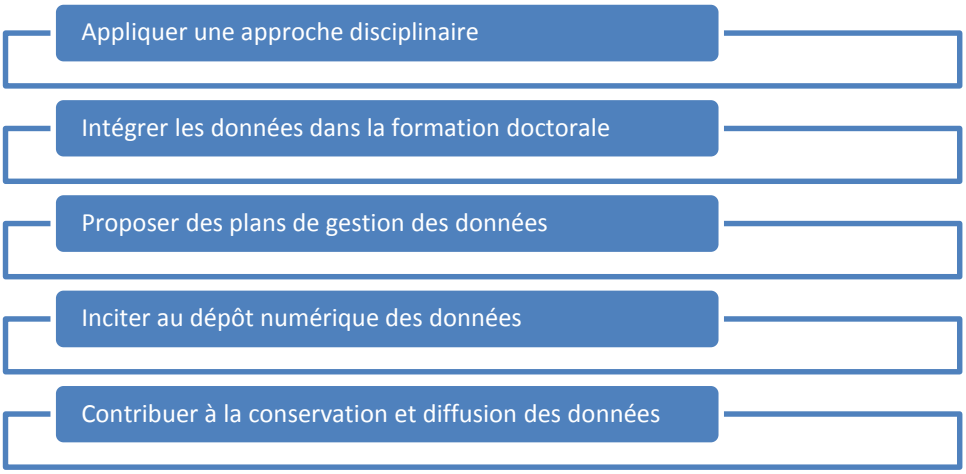
« Les données sont devenues une ressource, peu différente des matières premières. »
(Charles Huot, *GFII*)

Réunir les forces

Transversalité et synergie

Un projet modulaire

L'enjeu des données de la recherche est important, la question est posée, et la réponse ne peut attendre. L'objectif de notre livre blanc est de relever le défi et de proposer un cadre pour la valorisation des données de la recherche sur le campus de l'Université de Lille 3, en mettant l'accent, dans un premier temps, sur les données de la recherche produites par les doctorants. La proposition est portée par la conviction qu'il faut réunir les forces de notre université et faire jouer la synergie entre les laboratoires de



recherche, l'Ecole doctorale, le Service commun de la documentation et l'Atelier National de Reproduction des Thèses. La synergie transversale entre ces services sera indispensable pour la réussite. Mais elle ne suffira pas, et il faudra développer d'autres partenariats pour s'appuyer sur des compétences, outils et services externes au campus.

L'Université peut mettre en place cette démarche dès 2015, sans attendre la création du futur Learning Centre ou la construction de l'Université de Lille. Nous prônons un projet modulaire en cinq axes, qui apporte une solution immédiate aux besoins des doctorants ; il servira de modèle pour l'ensemble de la communauté scientifique de l'Université de Lille 3, il positionnera aussi notre université dans le domaine des données de la recherche tant au niveau local (UDL, COMUE) que national (CPU), tout en restant flexible et ouvert à d'autres développements.

Les auteurs

Stéphane Chaudiron (GERiiCO)

Catherine Maignant (Ecole doctorale SHS)

Joachim Schöpfel (ANRT)

Isabelle Westeel (SCD)

Avec la contribution d'Eric Kergosien, Bernard Jacquemin, Hélène Prost, Florence Thiault (GERiiCO) et Cécile Malleret (SCD), et avec le soutien financier de la MESHS (projet DRTD-SHS).

(1) Faire face à la diversité

Pour une approche disciplinaire, proche du terrain

Pas de solution unique

Les enquêtes et retour d'expériences concordent sur un point essentiel : dans le domaine des données de la recherche, il n'y a pas de solution unique, et une stratégie *top-down* est vouée à l'échec. Pour répondre aux besoins des chercheurs, il faut prendre en compte leurs pratiques. Certaines communautés sont plus avancées que d'autres dans l'archivage et le partage des résultats scientifiques. Aussi, notre étude des thèses de Lille 3 révèle différents profils disciplinaires selon les sources utilisées et selon les types de données produites. Un historien ne produit pas le même type de données qu'un linguiste ou psychologue. De même, le volume des données et la façon de les publier varient d'une discipline à l'autre. Dès le départ, il faut abandonner l'idée d'une solution unique en faveur d'une approche modulaire, différenciée, par option.

A partir du terrain

Pour développer une offre de service utile aux doctorants, nous devons partir du terrain, évaluer les attentes et usages avec les étudiants, les directeurs de recherche et les laboratoires. Dans certains cas, il faudra intégrer d'autres paramètres, l'existence de réseaux scientifiques, de projets structurants, d'infrastructures disciplinaires. L'idée est une proposition transversale, avec des prestations cœur, destinées à l'ensemble des doctorants, et d'autres plus spécifiques par domaine ou discipline. Le fil directeur de cette offre de services se nourrit obligatoirement des perceptions et initiatives des doctorants et leurs directeurs pour être répercutées, déclinées et prises en compte par les structures opérationnelles de pilotage.

L'apport des humanités numériques

Quand on parle du *big data* de la recherche, on pense d'abord et surtout aux grands équipements et projets scientifiques STM, comme le CERN, la *Protein Data Bank* ou le *Human Genome Project*. La nature de *big data* STM (tel que les « 3 V » volume, vitesse et variété) a largement influencé notre manière d'appréhender l'enjeu des données. Même si certaines solutions et procédures sont transposables, il convient d'insister sur les particularités des sciences humaines et sociales. Dans les thèses, pas de *big data* « 3V » mais des milliers de pages d'annexes riches d'une grande variété de *small data*. L'environnement émergent des humanités numériques offre un cadre intéressant pour des services et infrastructures adaptés. C'est un aspect important aussi par rapport à l'intégration à venir des structures et services de l'Université de Lille.

« Le numérique a facilité la mise à disposition des données de la recherche, corpus hétérogène qui offre des possibilités de développement quasiment illimitées. »

(CPU, *Universités 3.0*)

(2) Apprendre les enjeux

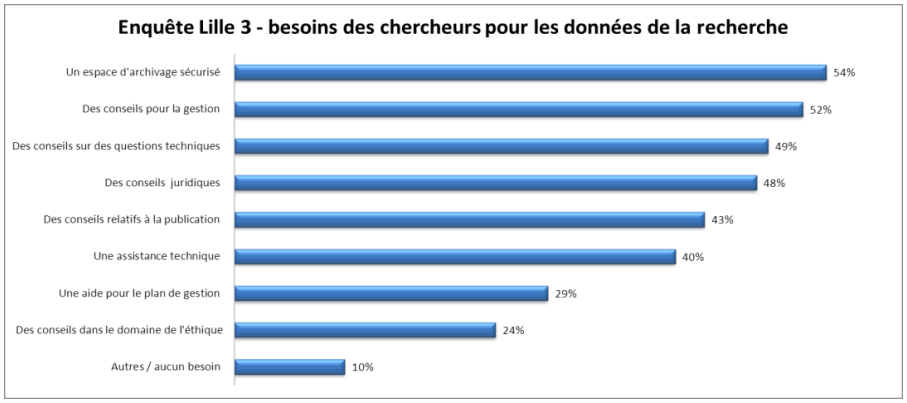
Pour une formation des doctorants à la gestion des données

De l'apprentissage « sur le tas »…

Souvent, les étudiants, doctorants et jeunes chercheurs apprennent « comment faire avec les données » en pratiquant, sur le terrain, au sein d'une équipe, avec les collègues. Pour le quotidien et les besoins personnels, cela peut suffire. Pour aller plus loin, établir un plan de gestion, archiver et partager, il faut d'autres compétences en méthodologie scientifique. Les chercheurs sont conscients de ce besoin, en particulier pour les aspects techniques, juridiques et documentaires.

… vers une offre de formation cohérente …

Rares sont à ce jour les démarches systématiques pour former les jeunes chercheurs à la gestion et valorisation des données de la recherche. Par manque d'expérience, de recul,



de légitimité ? Il est certain qu'il faut réunir les compétences du SCD, des laboratoires et des réseaux scientifiques et documentaires, pour développer une offre de formation cohérente et exhaustive, couvrant l'ensemble des procédures : la recherche, création, description et indexation, préservation, diffusion et réutilisation. En SHS, le cadre des humanités numériques lié à la science ouverte (*open access*) peut favoriser une telle offre.

… à plusieurs niveaux

Intégrer la gestion des données dans les études doctorales est l'objectif premier; il se réalise par la coordination transversale des séminaires disciplinaires et interdisciplinaires et des compléments de formation. D'autres actions : des cours dans les Masters de recherche, des journées d'études ciblées, des ressources en ligne (*webinars*, guides…). Ce programme fera appel à d'autres acteurs (CNRS, réseaux…). De même, il s'articulera avec la démarche de conseil et d'assistance personnalisée.

« L'information est un bien commun dont la diffusion doit être ouverte au plus grand nombre pour l'intérêt général. »

(J.-C. Cointot & Y. Eychenne, *La révolution Big Data*)

(3) Gérer les données

Aider à mettre en place des plans de gestion des données

Des pratiques personnelles

Chaque chercheur, chaque doctorant a ses propres pratiques par rapport à « ses » données, plus ou moins efficaces, plus ou moins sécurisées. Sauvegarde sur une ou plusieurs disques durs, parfois une copie sur un serveur externe, rarement de documentation (curation), encore plus rarement du partage… Ce que les enquêtes montrent, relève davantage des techniques du *personal knowledge management* que d'une gestion réfléchie et scientifique des données de la recherche. A titre individuel, cela peut paraître satisfaisant, sur mesure, *just enough*. A l'échelle d'une structure ou d'une communauté, cela ne suffit plus.

Des contraintes imposées

Aujourd'hui, les appels à projets (H2020 et autres) commencent à imposer des *data management plan* ou DMP. Certains éditeurs emboitent le pas quand ils demandent le dépôt des données (ou du moins un lien vers le dépôt) avec le manuscrit. La capacité de produire et préserver des données dans une norme et format accessible, documentée et réutilisable est devenue en quelques années un critère d'acceptation pour des projets, articles et communications. Tout cela nécessite une démarche méthodologique, l'anticipation des enjeux, une recherche de solutions adéquate, des procédures standards etc., autrement dit, un plan de gestion. Or, un tel plan ne s'improvise pas.

D'autres contraintes sont liées au fonctionnement normal d'un laboratoire : la fin d'un contrat de recherche, la fin d'un projet scientifique, le départ d'un collègue, la fusion (ou suppression) d'une structure, la soutenance d'une thèse : chaque fois, la question est posée : que deviennent les données ? Parfois tout simplement, où sont-elles ? Parfois plus compliqué, comment comprendre les données brutes ? Comment lire les fichiers ?

Des solutions à proposer

Pour ces DPM, il existe des cahiers des charges et des modèles, on trouve des guides et des outils de création. Pour aider les doctorants à créer un plan de gestion des données pour leur projet de thèse (ou tout autre projet), nous préconisons un triple service : la mise à disposition d'un guide DPM adapté aux SHS, un outil de création interactif (sur le modèle du *Research Data Planning Tool* du DCC (JISC) et une assistance lors de la rédaction du plan (conseil, évaluation).

« Ouvrir les données publiques c'est mettre à disposition des ressources stimulant l'innovation. »

(Laure Lucchesi, *Etalab*)

(4) Déposer avec la thèse

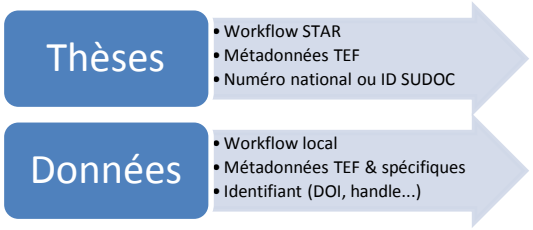
Pour un traitement différencié

Etape cruciale

Que deviennent les données de la recherche après la soutenance ? Pour certaines, elles sont déposées avec la thèse, dans les annexes ou le corps du texte, souvent en PDF, mal structurées, peu documentées. Souvent les résultats vont rester simplement « propriété » du doctorant, sur un disque dur ou une clé USB, virtuellement déjà perdus. Or, le dépôt de la thèse est un moment crucial pour le cycle de vie des données de la recherche doctorale. C'est l'opportunité unique pour l'institution de découvrir, identifier et sélectionner ces données. A condition toutefois d'insérer cette étape dans le processus du doctorat et de créer un workflow spécifique aux données, différent de celui des thèses.

Deux workflows

Il ne faut pas mélanger documents et données. Au contraire, à l'instar d'initiatives modèles comme à l'Université d'Emory, Virginia Tech ou Carnegie Mellon (projet *ETDplus*), il convient de séparer thèses et données pour les injecter dans deux *workflows* différents, avec d'autres traitements documentaires et informatiques. Plusieurs raisons à cela : d'autres caractéristiques (typologie, format, taille, diversité), d'autres finalités, d'autres conditions juridiques aussi. Le *workflow* pour les thèses est opérationnel (STAR). Il s'agit donc de créer un *workflow* local pour les données liées aux thèses, à partir de l'infrastructure existant (SCD, ANRT, Ecole doctorale) et avec l'aide de la DSI. Il ne s'agira pas de créer une archive de données locale mais de proposer une interface pour déposer et décrire les résultats, avec un premier contrôle de qualité et d'uniformité, avant un stockage définitif sur un site dédié.



Curation et identification

Du point de vue gestion documentaire, c'est l'étape cruciale : la création des métadonnées décrivant les données, partiellement identiques à celles de la thèse correspondante ; et l'attribution d'un identifiant unique et pérenne pour faciliter la publication et la citation des données mais aussi surtout, pour garder le lien avec la thèse. Le modèle ici pourrait être l'initiative DataCite avec l'attribution d'un DOI mais d'autres solutions sont envisageables (ARK…).

« All scientists put a lot of effort into collecting and processing data. And quite rightly so. But what happens next, after publication? »

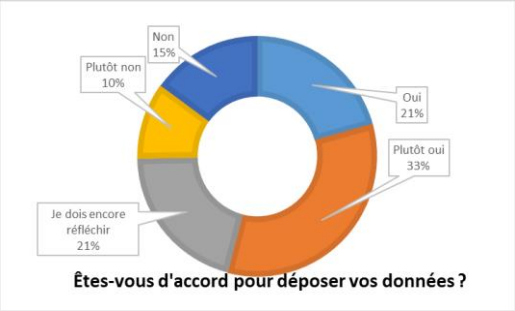
(Neelie Kroes, *CE*)

(5) Conserver et diffuser

Pour des solutions d'archivage et de réutilisation

Stocker ? Sauvegarder ? Préserver ?

Les enquêtes de terrain montrent des pratiques souvent peu satisfaisantes et incompatibles avec les exigences d'un DMP. En tant qu'employeur, l'Université se doit de proposer et promouvoir des solutions pour le stockage de données hétérogènes des chercheurs, pour la durée d'un projet, pour une conservation à plus long terme (5-10 ans) mais aussi en vue d'une diffusion. Fiabilité des supports et sécurisation de l'environnement sont les maîtres-mots, avec des sauvegardes régulières. Mais attention, il ne s'agit pas d'un archivage pérenne au sens du CINES mais d'outils, sur site ou à distance, pour encourager des pratiques plus efficaces. Il en existe déjà ; il faut s'assurer qu'ils conviennent aux besoins et qu'ils soient connus et acceptés.



Partager ? Diffuser ?

Les chercheurs sur le campus protègent leurs données mais sont majoritairement prêts à partager, avec leur équipe, avec d'autres collègues et/ou à la demande. Ils n'ont pas de préférence pour tel ou tel site et les avis sont partagé entre une archive internationale ou nationale et une solution locale (laboratoire, université). Le problème : en SHS, il n'y a pas beaucoup d'alternatives à ce jour. Certains types données pourraient, soit être déposés dans HAL, soit être stockés sur d'autres sites comme, par exemple, figshare.com en attendant une plateforme locale dédiée aux publications et données. Une veille systématique des archives de données et autres sites de stockage et partage des résultats, par discipline ou pour l'ensemble des SHS, par exemple en partenariat avec l'initiative re3data.org et le réseau Huma-Num, pourrait s'avérer aussi nécessaire qu'utile pour les doctorants et jeunes chercheurs.

Réutiliser ?

Partager les données, c'est contribuer à la « science ouverte », à rendre la recherche plus transparente, reproductible. Reste la question de la réutilisation des résultats de la recherche, par d'autres chercheurs, par la veille, les outils du *data mining* ou pour l'innovation. C'est une question juridique (quelle licence de diffusion ?) aussi bien que technique (quel format ? quelle documentation ?) et éthique (science et société). D'autant plus que tous les résultats ne se prêtent pas à une diffusion libre (données personnelles etc.). Ici l'idée serait d'aider les doctorants à trouver la bonne formule, dans son contexte, pour son projet, pour ses données.

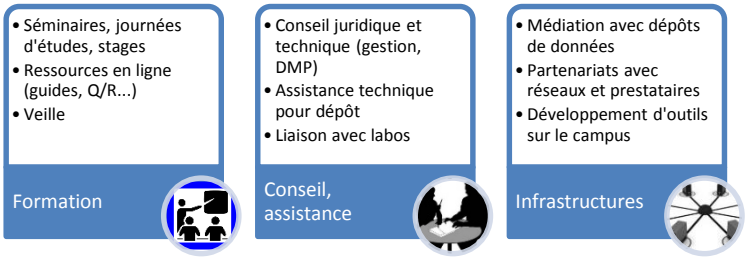
> « Dans un monde de la recherche de plus en plus numérique, la question de la conservation et de la préservation des données est devenue un enjeu majeur. » (*Huma-Num*)

Mettre en œuvre

Pour un projet transversal et fédératif

Architecture de l'offre

Contribuer à la gestion des données de la recherche dans les thèses de doctorat en SHS de l'Université de Lille 3 est un projet certes limité mais d'une grande valeur symbolique, d'autant plus qu'il témoignera de l'ambition – et de la capacité – de notre Université à relever le défi des données. L'offre de service à mettre en œuvre s'articule autour de trois axes – la formation des doctorants et jeunes chercheurs ; le conseil et l'assistance personnalisée (tout au long du cycle de vie de la donnée, de la préparation d'un *DMP* au partage en ligne) ; la mise à disposition des infrastructures nécessaires pour la sauvegarde, le stockage et le partage des données. C'est une offre modulaire, avec le potentiel de généraliser les services à l'ensemble du personnel scientifique du campus et ouverte aux évolutions et changements à venir.



Support

Une telle offre est possible par la concertation étroite des services directement concernés (laboratoires, école doctorale, SCD et ANRT), pilotée par le Conseil Académique et coordonnée par le SCD. Au moins pour la mise en chantier, ce projet pourra fonctionner à moyens constants. Assez vite, d'autres services seront associés (DR, DSI, DUNE, Communication). De même, le projet mobilisera des partenariats avec des organismes et réseaux opérationnels proposant des ressources et outils pertinents (Huma-Num, DARIAH, CNRS…).

Planning

- Dès maintenant : poursuivre les études sur les données.
- Rentrée 2015 : mettre en place une offre de formation sur les données.
- 2015-2016 : adapter un DMP pour les SHS, avec guide, outil et conseil.
- 2016-2017 : ajuster le workflow des thèses pour prendre en charge les données.
- 2016-2018 : développer des partenariats et infrastructures.

« Affordable and easy access to the results of research we fund is important for the scientific community and for innovative business. »

(Robert-Jan Smits, *CE*)

Université de Lille 3

Sciences humaines et sociales

Une recherche d'excellence

14 laboratoires labellisés, dont 4 unités mixtes Université/CNRS.

Plus de 500 enseignants-chercheurs publiants.

579 doctorants inscrits dans 3 écoles doctorales et 55 thèses soutenues en 2013.

Membre de 3 projets d'investissement d'avenir :

- SCIENCES ET CULTURES DU VISUEL – iCAVS : un cluster de recherches bâti sur un concept unique en France et IrDIVE : une plateforme technologique de niveau international labellisée Equipement d'excellence (EquipEX).
- LABEX DISTALZ : Développement de stratégies innovantes pour une approche transdisciplinaire de la maladie d'Alzheimer.
- Le SIRIC : ONCOLille - Site de Recherche Intégrée sur le Cancer.

Un engagement fort pour la science ouverte…

Mise en place d'une archive institutionnelle en 2013.

Presque 2000 publications en libre accès, dont plus de 150 thèses de doctorat.

Une démarche pour la gestion des données de la recherche.

… et pour les humanités numériques

De nouvelles formations en Licence et Master.

Une recherche d'excellence priorisée sur 3 thématiques :

- santé – éthique – vulnérabilité
- sciences et culture du visuel
- processus de création et d'innovation

La création d'un Learning Centre à horizon 2020.