Projet DLCM: les données de la recherche en Suisse

Jan Krause

Eliane Blumer

Raphaël Rey

Cet article est publié dans [Hors-Texte](http://www.agbd.ch/hors-texte/) numéro 110 (année 2016).

Licence: Creative Commons, CC-By-SA.

# Introduction

Le Projet Data Lifecyle Management (DLCM) vise à mettre en place des services pour la communauté académique suisse en matière de gestion des données. DLCM est financé par [SwissUniversities](https://www.swissuniversities.ch/fr/) dans le cadre du programme SUC P2. Il y a un an environ, un [article](http://www.ressi.ch/num16/article_110) en détaillait les objectifs dans la revue Revue électronique suisse de science de l’information (Ressi). Ce texte, plus synthétique, se focalise principalement sur les avancées du projet au cours de cette dernière année. Nous commencerons cependant par une brève remise en situation.

Accepté en août 2015, DLCM vise à développer ses services en partant des besoins de la communauté. D’entrée de jeu, 49 chercheurs spécialisés dans des disciplines variées ont été interviewés à travers la Suisse. Concrètement, le projet rassemble les forces de l’EPFL, de la HEG / HES-SO, de l’UNIL, de l’UNIBAS, de l’UNIZH, de l’ETHZ, de l’UNIGE et de SWITCH. Pour répondre aux besoins identifiés dans ces interviews, cinq groupes de tâches, appelés *tracks*, ont été mis en place. Comme le suggère le nom du projet, ils couvrent l’ensemble du cycle de vie des données, tel qu’illustré dans la figure ??? ci-dessous. Nous allons les passer en revue au fil de cet article.

Mais tout d’abord, qu’entendons nous ici précisément par “données de recherche”? La réponse à cette question n’est pas évidente, car elle varie en fonction des disciplines et institutions, et ce projet vise justement l’ensemble des disciplines et institutions suisses. L’Organisation de coopération et de développement économiques (OCDE) propose cependant une [définition intéressante](http://www.oecd.org/fr/sti/sci-tech/38500823.pdf): il s’agit d’enregistrements factuels (chiffres, textes, images et sons), qui sont utilisés comme sources principales pour la recherche scientifique et qui sont souvent reconnus par la communauté scientifique comme nécessaires pour valider des résultats de recherche.



Un cycle de vie des données de recherche [crée par l’Université de Sydney](https://library.sydney.edu.au/research/data-management/) et commenté par l’[URFIST](https://drive.google.com/file/d/0BxKZLWq08xX-TW5VOEUtd2FSRE0/view?pref=2&pli=1).

# Réalisation et tâches en cours

À ce jour, où en est le projet? C’est ce que nous allons aborder, track par track, dans les sections suivantes.

## Track 1 : Guidelines, policies et portail web

Ce track a produit un modèle de politique institutionnelle sur la gestion des données de recherche. De telles politiques existent déjà à l’étranger; celle-ci à la particularité d’être adaptée au contexte suisse et réutilisable par toute institution. Dans la foulée, la mise en place d’une politique nationale est actuellement à l’étude.

Autre réalisation, une checklist destinée à assister les chercheurs dans la mise en place de leurs plans de gestion de données. En effet, les responsables de laboratoires se doivent de rédiger un Data Management Plan (DMP) pour chaque projet. Un DMP permet de se préparer aux divers défis survenant au fil du cycle de vie des données du projet de recherche. En d’autres termes, il s’agit d’une bonne occasion de trouver des solutions avant que des problèmes ne se posent. Citons par exemple les pertes de données, des coûts de stockage trop élevés, l’incompatibilité technique, des conflits de propriété intellectuelle ou de protection de la vie privée. Pour assurer une gestion de données de qualité, les DMP sont de plus en plus fréquemment exigés par les bailleurs de fonds.

De plus, le track 1 met en place une collection de ressources et outils pertinents à chaque étape du cycle de vie. Ce catalogue a déjà été conceptualisé et créé. Il est en ce moment dans une phase d’alimentation par les membre du projet. La collection couvre une grande variété d’outils: guidelines, policies, logiciels, répertoires, archives, etc. Ces ressources ont été choisies et sont indexées avec soin. Dès 2017, le catalogue sera ouvert aux contributions de la communauté et modéré par DLCM.

Finalement, une première version du portail DLCM a été mise en place rapidement au début du projet. Depuis, le nouveau portail [dlcm.ch](http://dlcm.ch) a été complètement re-conceptualisé et mis en linge peu avant la rentrée 2016. Les délivrables de l’ensemble du projet DLCM sont téléversés au fil de l’eau dans cette nouvelle mouture. Les rubriques principales du portail offrent un bon tour d’horizon de sa composition:

* News and events : événement organisés et autres actualités (voir ci-dessous, track 5).
* Data Lifecycle Management : les ressources sur le Data Management Planning (p.ex. la checklist DMP de DLCM mentionnée précédemment).
* Ressources : on y retrouve notamment policies et guidelines ainsi que la collection d’outils mentionnée ci-avant.
* Consulting : service de consulting (en cours de mise en place).
* Training and education : modules et ressources de formation (contenu en cours de création).

## Track 2: Gestion des données actives

Que sont les données actives? Il s’agit des données avec lesquelles les chercheurs travaillent au quotidien. Il est clair que leur type et les workflows qui leur sont associés sont très variés en fonction des disciplines. Par exemple, en sciences de la vie, les systèmes de gestion de l’information de laboratoires sont très populaires (on parle en général de Laboratory Information Management System, les LIMS). En sciences sociales, ce sont plutôt les environnements de recherche virtuels (ou Virtual Research Environnements, les VRE) qui ont le vent en poupe.

Dans le cadre de DLCM, les LIMS tiennent une place importante. Plus précisément, un LIMS facilite la gestion d’un laboratoire et de ses données. “Il gère par exemple la saisie sur le terrain et la traçabilité des échantillons, les utilisateurs, les instruments, les stocks et approvisionnements, le suivi des produits et des équipements, les dossiers des patients.” ([Wikipedia 2016](https://fr.wikipedia.org/wiki/Syst%C3%A8me_de_gestion_de_l%27information_du_laboratoire)). De plus, cet outil assure le stockage et l’organisation des données, notamment dans le cadre de l’exécution des protocoles de laboratoire et de la gestion de leur versionnement. Des modules de cahier de laboratoire électronique remplacent le traditionnel document en papier. De ce fait, les LIMS deviennent quasiment incontournables pour le travail collaboratif, car ils permettent de systématiser la documentation des équipes scientifiques fréquemment caractérisées par un fort turn-over. Cela permet aux personnes clés, les rares membres permanents des laboratoires, de s’y retrouver.

En pratique, DLCM travaille par exemple sur:

* Une analyse de marché des solutions LIMS, avec publication des résultats et de la méthodologie (en cours, bien avancé).
* Une généralisation du buisness model des LIMS en sciences de la vie à d’autres disciplines (en cours, bien avancé).
* Le développement du logiciel libre OpenBIS, le LIMS de l’ETHZ (en cours, déjà en production).
* Du consulting entre institutions (l’EPFL a par exemple reçu l’UNIBAS pour lui présenter sa solution SLIMS et l’UNIGE a reçu l’UNIBAS pour en savoir plus sur leur LIMS de LabKey Software).
* L’amélioration de l’interopérabilité entre SLIMS et les archives de données (un module de publication en quelques clics d’expériences entières de SLIMS vers [Zenodo](http://zenodo.org) va être développé l’année prochaine).
* La mise en place de tutoriels vidéos en ligne (en cours d’élaboration).

## Track 3: Préservation et publication

Losrqu’qu’un chercheur a terminé d’analyser et d’exploiter ses données et qu’il considère avoir obtenu un jeu de données démontrant ou infirmant son hypothèse, il se préoccupe de sa publication, de façon à valider son étude, mais aussi d’être crédité par sa communauté de recherche, c’est à dire d’être cité. La publication des données peut se faire de différentes manières. La plus répandue est de publier les données directement avec l’article scientifique dans une revue spécifique, comme par exemple [PLOS ONE](http://journals.plos.org/plosone/). Une autre approche consiste à lier les données à l’article publié au moyen d’un Digital Object Identifier ([DOI](https://www.doi.org/)). Il’s d’agit d’identifiants pérennes offrant des liens stables, y compris lorsque l’hébergeur de données effectue des modifications importantes sur son système. De nombreux dépôts numériques, tels que par exemple [Figshare](https://figshare.com/), [Dryad](http://datadryad.org/) ou [Zenodo](https://zenodo.org/) attribuent systématiquement des DOIs à leur ressources. Aujourd’hui, des milliers de tels dépôts sont disponibles. Pour se faire une idée de leur diversitié, on peut consulter le répertoire [re3data.org](http://www.re3data.org/). Ainsi, les chercheurs ont l’opportunité de recevoir un double crédit: l’un pour la publication traditionnelle, l’autre pour les données. Ce qui est tout à fait justifié, comme le conditionnement de données de qualité, prend beaucoup de temps mais est essentiel à la reproductibilité des études. Une autre voie de publication de données, plus récente, sont les *Data Papers*. Ces derniers trouvent leur place dans des journaux hybrides ou dédiés, ou encore au sein d’espaces numériques spécialisés dans la publication de données. Donnons l’exemple du journal [Scientific Data](http://www.nature.com/sdata/) de Nature.

Dans le cadre des interviews DLCM, une tendance c’est dégagée: une fois les données publiées, le chercheur estime son travail achevé. Néanmoins, ces données devraient demeurer recherchables, accessibles, intègres, interoperables et réutilisables dans le futur (au moins proche et si possible sur le long terme). Ce qui n’est généralement pas le cas.

C’est dans ce contexte que le track 3 du projet DLCM entre en matière. Il se charge de la publication et de l’archivage des données, en prenant en compte les différentes technologies implémentées dans les différentes institutions partenaires. Ce track travaille notamment sur:

* Un *gap analysis* selon la méthodologie [POWRR](http://digitalpowrr.niu.edu) pour déterminer les lacunes de chaque dépôt institutionnel dans le but de les rendre compatibles *Open Archival Information System* (OAIS). L’[OAIS](http://www.iso.org/iso/catalogue_detail.htm?csnumber=57284) est le modèle de référence pour les systèmes et l’organisation de l’archivage dans une organisation donnée. Cette tâche est terminée et les résultats mettant en évidence différences et similarités entre les systèmes sont finalisés. Voire figure ???.
* En se basant sur l’analyse POWRR précédemment mentionnée, les spécifications de la structure des paquets d’information à archiver (paquets de soumission, de diffusion et d’archivage) sont en cours de développement.
* De plus, des spécifications de l’interopérabilité entre les différents dépôts et d’autres systèmes souhaités vont faire partie des développements (cette tâche vient de commencer).
* Dans un futur proche, un rapport de concept pour une infrastructure nationale de stockage sera rédigé et publié (tâche à venir).
* Finalement, le développement d’un prototype et des *Application Programming Interfaces* (APIs) autour de ce prototype aura lieu. L’idée ce cachant derrière cela est d’utiliser ces APIs comme façades normalisées, permettant aux diverses archives suisses d’être interopérables (à venir également).



 Résultat du Gap Analysis du track 3, basé sur la méthodologie POWRR.

## Track 4: Consulting et formations

La partie 4 du projet DLCM s’occupe principalement de la création de formations. En particulier de formations, qui n’ont pas encore été développées dans les autres branches du projet. Dans un deuxième temps, le track 4 vise à mettre en place des services de conseil (consulting) pour la gestion des données de recherche au bénéfice des institutions qui en ont besoin.

Très concrètement, les réalisations du track 4 sont:

* Des formations, comme par exemple la collaboration entre l’EPFL, l’UNIGE et le projet Train2Dakar (des ateliers sont proposés sur deux ans, en 2015 et en 2016).
* Un catalogue exhaustif de formateurs et formatrices, ainsi que de formations existantes (ceci est en cours).
* Une analyse des besoins concernant la gestion de données au sein des HES (en cours également).
* Un concept pour un service de conseil générique, pouvant être mis en place dans des institutions de types divers (juste commencé).

## Track 5: Communication et dissémination

Tous les services décrits ci-dessus concernent le cycle de vie des données. Le track 5, quant à lui, se charge de la communication et la dissémination du projet DLCM dans son ensemble.

Pour ce faire, il organise un événement de communauté “DLCM” le 29 novembre 2016 à l’EPFL. L’objectif de cet événement est d’une part la présentation des acquis du projet, mais également des discussions, workshop et présentation d’autres projets internationaux dans le même domaine, le tout en vue de lancer une nouvelle communauté autour des “données de recherche”.

# Conclusion

Vous avez pu vous faire une idée des délivrables de ce projet, via les exemples mentionnés au fil de ce texte. Ces livrables sont variés, nombreux et exigeants. Au vu de la complexité de la gestion des données, aucune institution suisse n’aurait été à même de prendre en charge à elle seule un tel dossier dans son entier. Il est donc tout à fait pertinent de répartir le travail de façon à mettre en place les solutions et services nécessaires à une bonne gestion des données tout au long de leur cycle de vie.

Le gain le plus important du projet DLCM réside peut-être dans les collaborations mises en place en son sein. Et cela sur plusieurs plans. D’une part, à notre conaissance, c’est la première fois qu’autant d’institutions académiques s’unissent dans le cadre d’un projet en Suisse. D’autre part, de nombreuses disciplines et métiers collaborent à cette fin. En voici deux illustrations: un bio-informaticien discutant avec une chercheuse en humanités numériques, un bibliothécaire scientifique entrant dans des laboratoires de sciences de la vie pour en analyser le mode de fonctionnement.

Cependant, la complexité du projet et le nombre d’acteurs impliqués posent des défis. Tout d’abord, une grande taille implique une grande inertie, et la mise en route de certaines tâches est parfois lente, ne serais-ce que parce qu’il faut s’assurer que la majorité des marins soit bien à bord du bateau. Un autre défi est, comme c’est souvent le cas pour les projets à large échelle, la pérennité des solutions mises en place. C’est pourquoi, au sein de DLCM, nous avons pris soin de développer le Business Model le plus rapidement possible. En d’autres termes, il s’agit des solutions de financement et de maintenance après la fin du projet en 2018, de façon à assurer la durabilité des services. Concrètement, nous suivons une méthodologie spécifique, le Business Model Canvas et Value Proposition Design de [Strategyzer](http://www.businessmodelgeneration.com/canvas/bmc). Cette approche agile permettra d’adapter nos service en fonction d’évaluations régulières et du feedback des utilisateurs finaux.

D’ici 2018, l’ensemble des services DLCM deviendra réalité. D’ici là, les délivrables sont mis en ligne au fil de l’eau. Des résultats ont déjà été obtenus à ce jour et nous vous invitons à venir les découvrir sur le portail [DLCM.ch](http://dlcm.ch) dès maintenant.

# References