Service de gestion des données de recherche à la Bibliothèque de l'EPFL: historique, services et perspectives

Eliane Blumer
researchdata@epfl.ch
https://orcid.org/0000-0002-0972-5396
Coordinatrice données de recherche, Bibliothécaire de liaison pour les Sciences de la Vie

Jan Krause
researchdata@epfl.ch
Data librarian

1. Introduction

Tout service réussi répond à un besoin. Dans ce cas, le besoin est double, d'une part celui des chercheurs et d'autre part celui de l'institution. En effet, au cours des dernières années, les principaux bailleurs de fonds ont mis en place des exigences plus poussées en matière de gestion de données. En parallèle, les éditeurs ont également élevé leurs critères en ce sens. De plus, la complexité et la quantité des données ne cessant d'augmenter, certains chercheurs ressentent le besoin d'être épaulés par leur organisation. Par ailleurs, depuis le début de l'année la nouvelle direction de l'EPFL a mis un point fort sur l'ouverture de la science et sa reproductibilité (*open science*), ce qui inclut une meilleure gestion ainsi qu'une ouverture plus large des données de la recherche.

Pour répondre à ces besoins croissants, le service de gestion des données de la recherche de la Bibliothèque de l'EPFL a mené deux projets en parallèle : la création de son offre de service et une participation très active au projet national Data Life-Cycle Management (DLCM). Chacun nourrissant l'autre et permettant le développement de compétences. Cet article décrit l'historique de la mise en place du service de gestion des données de recherche à la Bibliothèque de l'EPFL. Il s'arrête sur les services offerts à ce jour et se termine sur les perspectives d'évolution.

2. Historique de la mise en place du service

L'implication des bibliothèques universitaires dans la gestion des données de la recherche se situe dans la continuité de leurs missions. En effet, les données sont désormais considérées comme des éléments de publications scientifiques à part entière dans le cadre de la dissémination de la recherche. Le partage des données de recherche devient donc crucial. Il s'agit d'une extension des compétences en matière de métadonnées, de licences, de dépôts institutionnels, notamment. Pour cette raison, entre 2012 et 2013, le dossier « Données de la recherche » a été intégré dans les axes stratégiques de travail au sein de la Bibliothèque, qui a su anticiper les besoins institutionnels. Le travail a démarré d'un côté avec un état des lieux des services offerts dans d'autres institutions internationales, et de l'autre avec la mise en place de collaborations liés aux données de la recherche au sein de l'institution. En pratique, la gestion des données de recherche a été insérée dans le cahier des charges de deux personnes entre fin 2013 et début 2014, ce qui représentait une charge de 0.4 ETP au total.

Tout au long de 2014, le projet de mise en place des services de support pour les chercheurs en matière de gestion de données a été poursuivi, et une première action de sensibilisation a été réalisée avec la conférence « Open Research Data – The Future of Science ». Ce fut l'occasion de se réunir avec les chercheurs et les autres acteurs impliqués au sein de l'institution (le Research Office, le service informatique), et d'échanger avec d'autres acteurs concernés en Suisse (professionnels de l'information, etc.).

Cette même année, la Bibliothèque s'est focalisée sur la mise sur pied du projet national Data Life-Cycle Management (DLCM)[1]. Nous avons été particulièrement actifs dans la soumission commune de la proposition de projet auprès du fond swissuniversities P-2 Information Scientifique. Celle-ci a permis d'obtenir un fond de cinq millions au total destiné à répondre au niveau national aux besoins les plus imminents des chercheurs en matière de gestion de données, incluant le planification, la gestion active, la préservation ainsi que le training et consulting. Dans ce contexte, la Bibliothèque de l'EPFL s'est engagée activement en tant que

N° DOI : 10.55790/journals/ressi.2017.e1805

responsable d'un axe de travail, ce qui lui a permis d'augmenter ses forces de travail d'1.0 ETP via des fonds mixtes. De manière générale, la Bibliothèque de l'EPFL a bénéficié du projet DLCM qui a joué le rôle de moteur national pour les institutions participantes.

En 2015, le service se formalise. La Présidence de l'EPFL approuve en janvier un projet de service d'accompagnement des chercheurs pour la rédaction des Data Management Plans (DMP) en phase pilote pour six mois. L'objectif du service était de soutenir les chercheurs qui devaient répondre aux requêtes en matière de gestion de données de la recherche de la Commission Européenne dans le cadre de l'Open Data Pilot du programme de financement Horizon 2020. Le projet a été ensuite confirmé et stabilisé après le pilote. C'est également dans ce contexte et au cours de cette année que les effectifs ont pu être augmentés de 0.8 ETP.

Une page web a été également insérée dans le site de la Bibliothèque, mettant en évidence les différents services impliqués dans la gestion des données à l'EPFL. De plus, un article présentant les enjeux de la bonne gestion des données ainsi que le service est publié dans le journal institutionnel[2].

Côté sensibilisation, la Bibliothèque a de nouveau organisé un événement, étalé sur quatre dates autour la thématique de l'Open Science, cette fois-ci avec une orientation plus pratique, avec entre autres des workshops sur [3] et la fouille de textes et de données (*data and text mining*).

Côté formation, une première offre a été mise en place, avec l'intention de couvrir les besoins institutionnels. Une formation généraliste d'une journée (« Optimizing Research Data Management »), accessible via le Service de Formation du Personnel et destinée à l'ensemble de l'EPFL avec un fort accent sur la recherche a été créée.

Dès 2016, un module de gestion des données de recherche dans la formation en compétences informationnelles à destination des doctorants a été initié ; il se focalise sur la publication de jeux de données. Des formations ciblées pour des groupes de recherches sont aussi offertes à la demande. Dans ces cas, une étude des besoins du laboratoire est effectuée au préalable, d'une part via un sondage auprès des membres du groupe et d'autre part en se basant sur une enquête plus approfondie auprès d'une ou deux personnes choisies. Finalement, des propositions d'améliorations sont faites et discutées avec l'ensemble du groupe pendant une demi-journée ou une journée entière. Le but étant de prendre des décisions débouchant sur des actions concrètes.

Sur le plan de la sensibilisation, la Bibliothèque de l'EPFL a notamment agi via la co-organisation de la conférence opendata.ch à Lausanne, en prenant en charge la partie dédiée aux données de recherche, les autres partenaires se focalisant sur les aspects gouvernementaux et commerciaux de l'open data.

Sur le plan pratique, pour renforcer les actions de support pour les chercheurs dans la rédaction des DMPs, la Bibliothèque a créé en 2016, en collaboration avec l'Ecole Polytechnique Fédérale de Zürich (ETHZ), une Data Management Checklist ainsi qu'un premier modèle de Data Management Plan (en partenariat aussi avec le Digital Curation Centre). Ce modèle répond au lancement du Open Data Pilot d'Horizon 2020[4], le programme européen de financement de la recherche et de l'innovation. Le pilote portait sur neuf domaines scientifiques[5] et comprenait la nouvelle exigence de fournir un DMP. Plus précisément, ce type de plan consiste à répondre à un ensemble de questions spécifiques, telles qu'illustrées dans le tableau 1 ci-dessous. En d'autre termes, il s'agit d'un document définissant pas à pas

N° DOI : 10.55790/journals/ressi.2017.e1805

la gestion des données, depuis leur création jusqu'à leur archivage. Étant donné que Horizon2020 est l'un des deux bailleurs de fonds les plus importants pour l'EPFL, il est évident que ceci a demandé une forte implication du service pour répondre au besoin des chercheurs. Il était également important de remplir ces exigences pour la direction et de maintenir le niveau de financement obtenu.

Par ailleurs, le projet DLCM ayant été accepté, une partie importante des forces de travail s'est orientée sur ce projet. Notamment, sur la réflexion autour du portail DLCM, ainsi qu'un travail préliminaire sur le projet de préservation à long terme. Ceci a permis à l'équipe de la Bibliothèque de renforcer son réseau et ses compétences

En 2017, Martin Vetterli a été nommé président de l'EPFL. Dès son entrée en fonction, celui-ci a fait remonter l'*Open Science* au premier plan des préoccupations stratégiques de l'institution. En particulier, la direction a apporté son soutien à la démarche de la Bibliothèque et de son service de gestion des données de recherche. Des actions pragmatiques ont été entreprises en étroite collaboration avec la Bibliothèque tel qu'un hackathon basé sur l'ouverture de données de l'EPFL[6] ou encore une école d'automne pour les doctorants, Open Science in Practice, a été mise en place. La bibliothèque s'y est fortement impliquée, notamment en proposant les modules sur le DMP, la publication et la préservation de données. De plus, la formation interne déjà en place sur le data management planning a vu sa demande augmenter et a été donnée quatre fois.

Ce fut également le moment opportun pour lancer le nouveau site web http://researchdata.epfl.ch , déjà en préparation. Celui-ci comporte les sections suivantes : Planification et financement, Travailler avec les données, et Publication et préservation. Les informations-clés concernant le support et les formations y sont disponibles.

Toujours en 2017, le Fond national suisse de la recherche scientifique (FNS) exige des DMP dès la soumission des projets, et ceux-ci sont devenus une part importante du service. De façon similaire, Horizon2020 a étendu son Open Data Pilot à l'ensemble des disciplines, rendant le DMP systématiquement obligatoire. En réponse à ceci, [7] à été élaboré, toujours en collaboration avec l'ETHZ.

Ce n'est pas uniquement l'accroissement des forces de travail qui a permis de monter ce service, mais également la collaboration avec d'autres services et institutions ainsi que l'organisation d'un bon nombre d'événements clés internes et externes à l'EPFL sur plusieurs années.

N° DOI : 10.55790/journals/ressi.2017.e1805

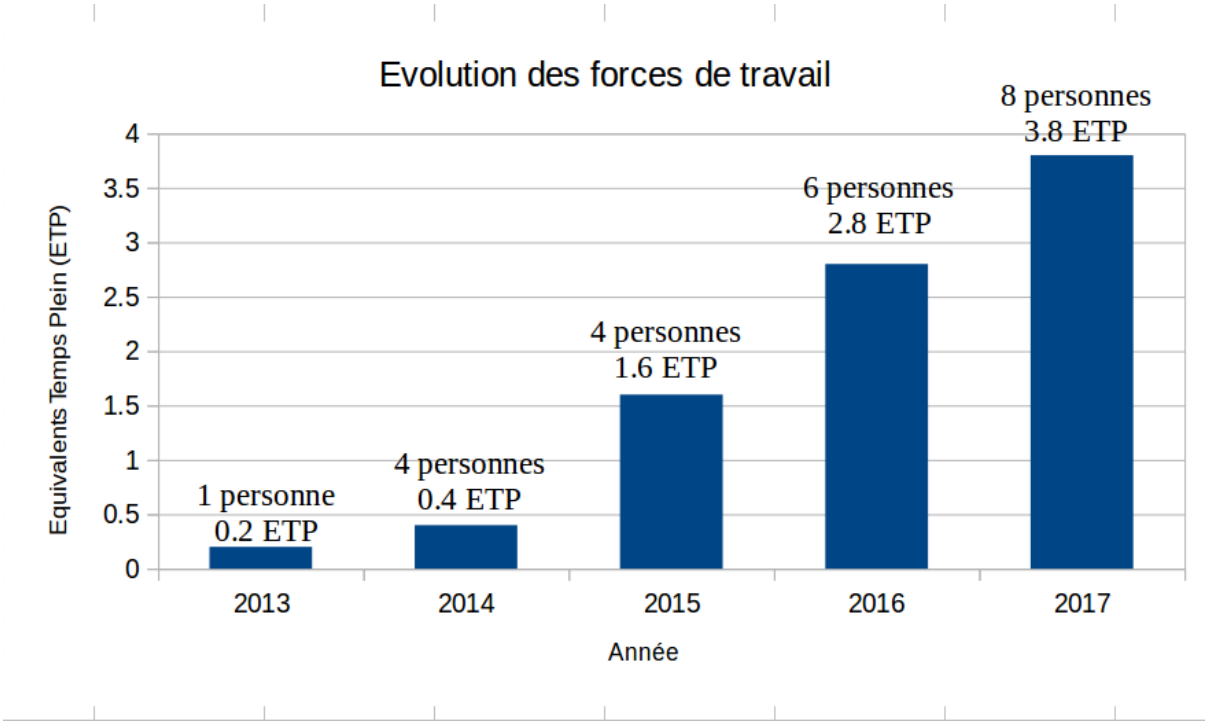

Figure 1 – Evolution des forces de travail de l'équipe, financement EPFL et DLCM confondus.

3. Les services à ce jour

Globalement, les services en matière de données de la recherche offerts par la Bibliothèque peuvent être divisés en deux catégories : le soutien et les formations.

3.1. Soutien

Le soutien se décline en plusieurs variantes.

Le plus fréquemment, il s'agit d'assister les unités de recherche à rédiger un DMP. Il s'agit souvent de répondre à l'exigence d'un bailleur de fonds, particulièrement le FNS ou Horizon2020. Moins fréquemment, les groupes désirent travailler sur un DMP dans le simple but d'améliorer leur pratique.

Dans d'autres cas, notre apport consiste à amener une expertise spécifique quant à une problématique précise. Il peut par exemple s'agir de choisir un dépôt pour la publication de données, ou une licence appropriée, ou encore un renseignement quant à la politique d'un éditeur en matière de partage de données.

Afin d'optimiser le support, nous avons créé différents guides. Parmi les plus utilisés, il y a le [8], élaboré en collaboration avec la bibliothèque de l'ETHZ. Ce document propose des recommandations ainsi que des exemples de réponses pour chaque question du bailleur de fonds (voir tableau 1).

N° DOI : 10.55790/journals/ressi.2017.e1805

Tableau 1 - Questions à développer pour le Data Management Plan du Fonds national suisse de la recherche scientifique

1 Data collection and documentation
1.1 What data will you collect, observe, generate or reuse?
1.2 How will the data be collected, observed or generated?
1.3 What documentation and metadata will you provide with the data?
2 Ethics, legal and security issues
2.1 How will ethical issues be addressed and handled?
2.2 How will data access and security be managed?
2.3 How will you handle copyright and Intellectual Property Rights issues?
3 Data storage and preservation
3.1 How will your data be stored and backed-up during the research?
3.2 What is your data preservation plan?
4 Data sharing and reuse
4.1 How and where will the data be shared?
4.2 Are there any necessary limitations to protect sensitive data?
4.3 I will choose digital repositories that are conform to the FAIR Data Principles.
4.4 I will choose digital repositories maintained by a non-profit organization.

Par exemple, concernant la question sur le copyright et la propriété intellectuelle, la recommandation suivante est proposée :

« Attaching a clear license to a publicly accessible data set allows other to know what can legally be done with its content. When copyright is applicable, Creative Commons licenses are recommended. However, Creative Commons licenses are not recommended for software. Amongst all Creative Commons licenses, CC0 "no copyright reserved" is recommended for scientific data, as it allows other researchers to build new knowledge on top of a data set without restriction. It specifically allows aggregation of several data sets for secondary analysis. Several data repositories impose the CC0 license to facilitate reuse of their content. In order to enable a data set to get cited, and therefore get recognition for its release, it is recommended to attach a CC-BY "Attribution" license to the record, usually a description of the dataset (metadata). To get recognition, data sets can be cited directly. However, to increase their visibility and reusability, it is recommended to describe them in a separated document licensed under CC BY "Attribution", such as a data paper or on the institutional repository.

When the data has the potential to be used as such for commercial purposes, and that you intend to do so, the license CC BY-NC allows you to keep the exclusive commercial use.

Reuse of third-party data may be restricted. If authorised, the data must be shared according to the third party's original requirement or license.

If you need guidance in the publication and license choice, you can check the suggested "Data publication decision tree"

Pour ce même point, trois exemples sont aussi donnés :

Example 1:

The research is not expected to lead to patents. IPR issues will be dealt with in line with the institutional policy. As the data is not subjected to a contract and will not be patented, it will be released as open data under Creative Commons CC0 license.

Example 2:

This project is being carried out in collaboration with an industrial partner. The intellectual property rights are set out in the collaboration agreement. The intellectual property generated from this project will be fully exploited with help from the institutional Technology Transfer Office. The aim is to patent the final procedure and then publish the work in a research journal and to publish the supporting data under an open Creative Commons Attribution (CC BY) license.

Example 3:

Data is suitable for sharing. They are observational data (hence unique) and could be used for other analyses or for comparison for climate change effects among many things. Reuse opportunities are vast. For this reason, we aim to allow the widest reuse of our data and will release them under Creative Commons CC0.

De plus, d'autres documents ont été élaborés et mis à disposition sur le site web, dont :

- Modèle pour Horizon 2020
- Formats de fichier recommandés pour la préservation à long terme
- Arbre de décision pour choisir une stratégie de publication des données

Un cas type de demande d'assistance pour la rédaction d'un DMP

Souvent, les demandes de soutien débutent par un mail reçu sur la boîte email researchdata@epfl.ch Prenons l'exemple fictif d'un message avec pièce jointe envoyé par une chercheuse en Sciences de la Vie constitué du modèle DMP FNS pré-rempli ainsi que de plusieurs questions. Imaginons que, dans le DMP, la nature des données produites ainsi que les droits de partage (embargo, données sensibles etc.) ne sont pas détaillés. Comme il y a beaucoup de points à couvrir dans ce cas, nous contactons la personne en lui proposant un entretien afin de discuter de vive voix de ses questions. Dans des cas plus simples, nous répondons directement par email ou par téléphone.

Le rendez-vous fixé, deux collègues de l'équipe données de recherche ainsi que le/la bibliothécaire de liaison de la discipline concernée se réunissent pour préparer ensemble le retour. L'équipe se rend compte que d'autres parties du DMP ont été traitées de façon superficielle ou manquent. C'est souvent concernant la préservation à long-terme, les métadonnées, les formats de préservation et les licences.

Lors de la réunion, les questions de la chercheuse sont traitées, puis, l'ensemble du document est passé en revue et des suggestions d'amélioration sont faites. En sciences de la vie, lorsque le projet est réalisé avec des partenaires industriels, les droits doivent être discutés avec tous les acteurs impliqués. On insiste également sur le fait que les données personnelles ou sensibles, sont soumises à des exigences légales, et en fonction de la situation, nous renvoyons la chercheuse vers le *Human Research Ethics Committee's (HREC)[9]*. Pour la gestion des données actives, nous renseignons sur les différents outils appropriés, et, le cas échéant nous renvoyons vers le service informatique concerné (stockage, cahier de laboratoire électronique, gestion des versions, etc.). En ce qui concerne les métadonnées, nous proposons des standards appropriés, en soulignant l'importance de décrire les jeux de données de façon à ce qu'ils soient FAIR[10], comme l'exigent la plupart des bailleurs de fonds. Pour la partie préservation à long-terme, nous rappelons que l'EPFL ne propose pas (encore) de solution interne. Dans le cas présent, une fois les données anonymisées, elles pourront être publiées gratuitement sur Zenodo.org [11]. Nous profitons de l'occasion pour discuter un peu plus en détails de l'initiative du FNS, ainsi que du budget disponible pour la gestion des données du projet.

A la suite de la réunion, nous renvoyons un message contenant tous les documents et informations utiles en pièce jointe.

Le soutien tel que décrit ci-dessus s'est avéré pertinent et apprécié. Cependant, au sein d'une institution forte de milliers de chercheurs organisés en environ 353 groupes de recherche, il n'est pas possible de revoir en détail le data management plan de chaque projet. C'est une des raisons pour lesquelles il est important de former les collaborateurs concernés.

3.2. Formation

- Présentation : court exposé non-interactif d'une thématique visant un public spécifique.
- Atelier court : d'une durée de deux heures, souvent combiné avec une pause sandwich à midi, visant généralement les chercheurs.
- Atelier d'une demi-journée : souvent utilisé pour la formation sur mesure d'un groupe de recherche, ou un module pour l'école doctorale.
- Atelier d'une journée et plus : visant généralement un public hétérogène : personnel EPFL en général, une école doctorale, bibliothécaires spécialistes.

La Bibliothèque propose une gamme de formations, tant à l'interne qu'à l'externe. La priorité est évidemment de servir l'institution, néanmoins, l'équipe « données de recherche », s'efforce de répondre aux sollicitations externes dans la mesure de ses forces. En effet, les interventions externes constituent une occasion précieuse de s'enrichir et d'échanger.

Nos formations prennent différentes formes, en fonction du contexte. Voici la panoplie que nous utilisons :

N° DOI : 10.55790/journals/ressi.2017.e1805

Un cas type : la formation destinée au personnel EPFL

Description

Savoir gérer et organiser ses données de recherche devient l'une des conditions s*ine qua non* pour garantir la qualité, la pérennité ainsi que la reproductibilité de ses données dans la durée. De plus, de nombreux bailleurs de fonds exigent aujourd'hui de savoir préparer un plan de gestion des données pour obtenir un financement. Ce cours d'introduction vous permettra d'acquérir les connaissances et ressources indispensables pour améliorer et optimiser l'organisation ainsi que le partage de vos données de recherche pour le long terme.

Objectifs

- Acquérir les connaissances de bases et découvrir les ressources à votre disposition pour mieux gérer vos données.
- Se sensibiliser aux bonnes pratiques en matière de gestion des données de la recherche tout au long du cycle de vie.
- Saisir les bénéfices d'une telle pratique et savoir répondre aux nouvelles exigences des bailleurs de fond dans ce domaine.
- S'approprier les outils à votre disposition pour optimiser la gestion de vos données au quotidien et à long terme.
- Connaître les étapes et outils pour optimiser la préparation d'un plan de gestion des données.
- Partager votre expérience avec vos collègues.

Parcours de formation :

- Définitions et cycle de vie des données de recherche.
- Avantages d'une meilleure gestion et positions des bailleurs de fonds.
- Présentation théorique des bonnes pratiques pour organiser et partager ses données de recherche.
- Mise en pratique de solutions en fonction des disciplines des participants à cette formation.
- Conseils personnalisés et partages d'expériences.

Public cible

Tout collaborateur souhaitant acquérir de nouvelles compétences pour une meilleure gestion de ses données de recherche. Cette formation est un prérequis pour suivre le cours pratique qui porte sur l'utilisation du logiciel de cahier de laboratoire électronique.

Autres informations

Cette formation offrira des bases théoriques et pratiques. Tout au long de la journée, des exemples concrets seront discutés et mis en pratique selon les besoins précis des participants. Chaque participant aura l'opportunité de pratiquer sur un ordinateur mis à sa disposition et pourra explorer les ressources et solutions disponibles pertinentes dans le cadre son domaine de recherche.

N° DOI : 10.55790/journals/ressi.2017.e1805

4. Conclusion

En résumé, la Bibliothèque a su anticiper les évolutions de la recherche scientifique et les nouveaux besoins des chercheurs afin de mettre en place un service adapté. Il a été nécessaire de rassembler plusieurs facteurs pour y parvenir. Tout d'abord, comme l'ensemble des compétences requises pour un tel service n'était pas centralisé à la bibliothèque, il a été essentiel d'établir des collaborations internes et externes. Cela a été l'occasion de devenir une plateforme centralisant les demandes. Ensuite, le contexte politique a été favorable. Premièrement, au niveau institutionnel, avec une volonté claire de soutenir l'Open Science. Deuxièmement, le contexte politique suisse et international a permis de s'engager dans le projet DLCM et de répondre au besoin des chercheurs, lié principalement aux exigences des bailleurs de fond FNS et Horizon2020.

Dans ce domaine en évolution perpétuelle, il est nécessaire de continuer d'innover. Pour commencer, les activités liées à la gestion des données de recherche de la Bibliothèque s'inscrivent dans une stratégie plus large, et qui est actuellement en train d'être approfondie. De plus, si l'on se restreint plus strictement au sujet de cet article, divers développements sont prévus. Pour commencer, les services et outils mis en place continueront d'être améliorés, notamment une nouvelle version du modèle de DMP pour le FNS est prévue. Par ailleurs, des nouveautés sont prévues, citons par exemple la mise en place d'un dépôt institutionnel pour les données, d'un cours crédité pour les doctorants, ou encore d'un service de data stewards travaillant en étroite collaboration avec les laboratoires.

5. Remerciements

Nous tenons à remercier chaleureusement tous nos collègues de l'équipe des données de recherche et au-delà, qui ont pris le temps de relire et améliorer cet article. Merci en particulier à Lorenza Salvatori d'avoir complété cet historique.

NOTES

[1] 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 » Krause & Blumer, Hors-Texte 110, Novembre 2016, p. 27sq. Les cinq tracks sont : Lignes directrices et politiques ; Données actives de recherche ; Préservation à long terme ; Conseil, formation et enseignement ; Diffusion. Plus d'information : https://dlcm.ch

[2] « Les données de recherche: une mine d'or à domestiquer et valoriser » Flash No. 04, mai 2015. https://mediacom.epfl.ch/files/content/sites/mediacom/files/Flash/Flash%2004-2015.pdf

[3] Voir : http://jupyter.org

[4] Voir : https://ec.europa.eu/programmes/horizon2020/

[5] Voir : http://www2.unavarra.es/gesadj/servicioBiblioteca/piloto_de_datos/4.%20p...

[6] Voir : http://datajamdays.org/

[7] Voir : https://researchdata.epfl.ch/files/content/sites/researchdata/files/doc/EPFL_SNSF_DMP_Template.docx

[8] Voir : https://researchdata.epfl.ch/files/content/sites/researchdata/files/doc/EPFL_SNSF_DMP_Template.docx

[9] Voir : https://research-office.epfl.ch/research-ethics/research-ethics-assessment/epfl-human-research-ethics-committee/hrec

[10] Les données FAIR, signifient: Findable, Accessible, Interoperable and Reusable. Voir: https://www.force11.org/group/fairgroup/fairprinciples

[11] Dépôt de données de recherche. Voir: https://zenodo.org

N° DOI : 10.55790/journals/ressi.2017.e1805