SWISS CLOUD

LES CINQ PILIERS DU PARTAGE

DANS LE CADRE DU PROGRAMME P5 DE SWISSUNIVERSITIES, UNE CINQUANTAINE DE CHERCHEURS PLANCHENT DEPUIS 2015 SUR UNE **STRATÉGIE NATIONALE** EN MATIÈRE DE GESTION DES DONNÉES DE RECHERCHE. PRÉSENTATION.

Pierre-Yves Burgi

Directeur adjoint de la Division du système de l'information

Les universités suisses n'ont pas attendu les directives du Fonds national (FNS) pour se pencher sur la question de la préservation et de l'accessibilité des données scientifiques. Lancé en septembre 2015 sous l'égide de Swissuniversities (association qui regroupe les responsables des hautes écoles universitaires, spécialisées et pédagogiques de Suisse depuis 2012), le projet *Data Life-Cycle Management* (DLCM) a en effet pour mission in fine d'élaborer une stratégie nationale en matière d'Open Access. Celle-ci devant être compatible avec les principes FAIR, autrement dit permettre que les données soient trouvables, accessibles, interopérables et réutilisables.

Pour y parvenir, il s'agit de coordonner les efforts que déploient actuellement les hautes écoles de manière dispersée pour mettre à disposition et traiter des informations scientifiques. Il convient également de créer un certain nombre d'outils permettant aux chercheurs d'assurer la sauvegarde et l'accès sur le long terme à leurs données. Placé sous la direction de Pierre-Yves Burgi, directeur adjoint de la Division du système de l'information de l'UNIGE, et regroupant sept autres institutions*, le projet DLCM repose sur cinq axes principaux. Tour d'horizon.

« AUJOURD'HUI, 80 % DU VOLUME DES DONNÉES SONT PRODUITES PAR 20 % DES CHERCHEURS. »

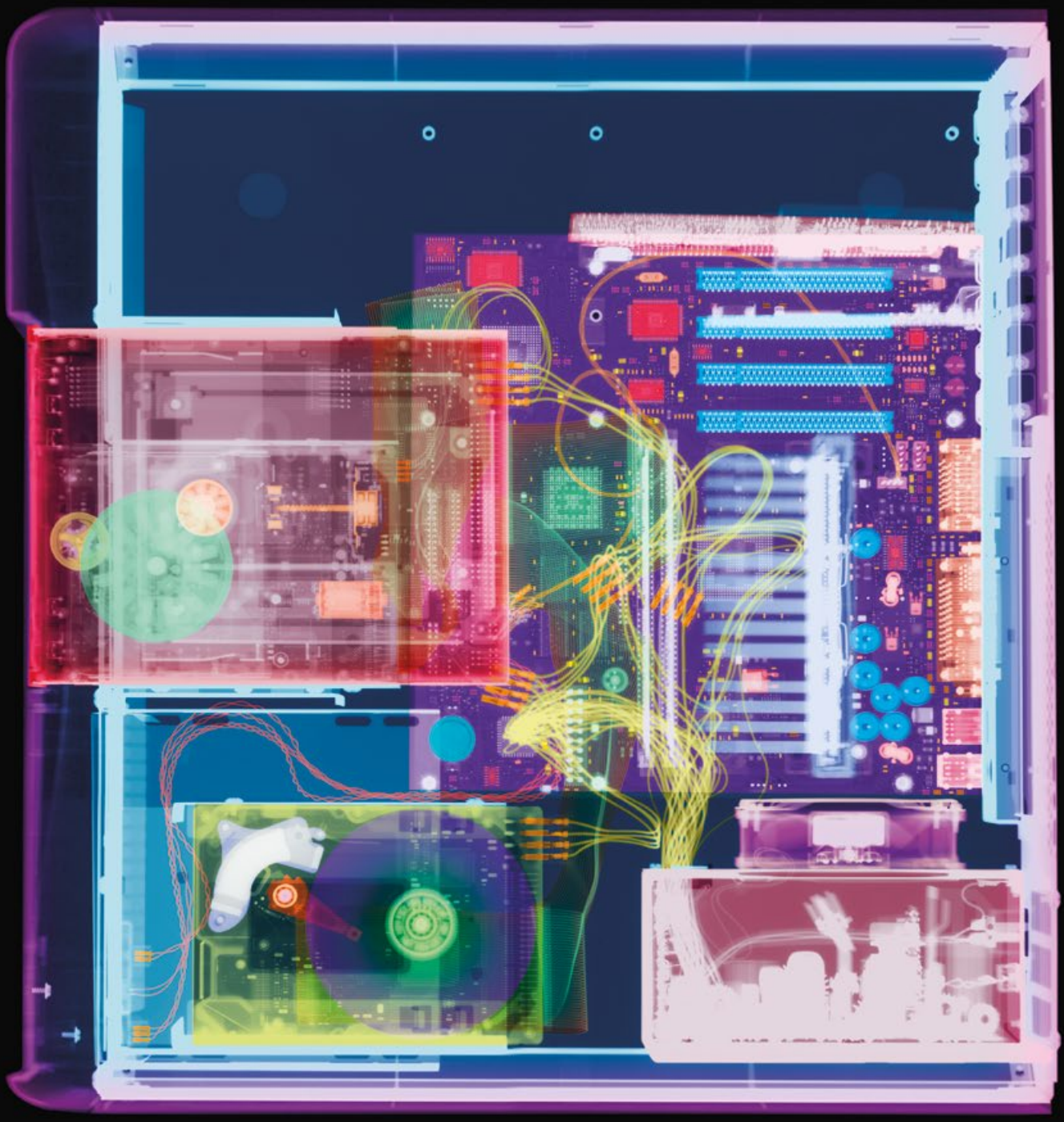
1. Le plan de gestion des données

Depuis l'automne 2017, chaque projet soumis au Fonds national doit être accompagné d'un document décrivant la façon dont seront traitées les données utilisées pour une publication, le DMP – pour *Data Management Plan* (lire page 22). Or, la chose ne va pas de soi pour tout le monde. *« Aujourd'hui, 80 % du volume des données sont produits par 20 % des chercheurs,* explique Pierre-Yves Burgi. *Certaines disciplines comme l'astronomie, la physique des particules, la géographie ou les sciences de l'environnement sont déjà très bien organisées. Elles collectent des données depuis plusieurs dizaines d'années et peuvent se débrouiller sans nous. À l'inverse, les 80 % de chercheurs qui gèrent des volumes de données relativement faibles sont assez démunis. Ils ne disposent pas des outils ni des procédures nécessaires et ils ne peuvent pas non plus s'appuyer sur leur communauté. »*

Afin de leur faciliter la tâche, les équipes du programme P5 ont donc commencé par élaborer deux modèles génériques, l'un concernant le DMP et l'autre la politique de gestion des données, pouvant être repris au niveau de chaque institution partenaire et leur permettant de définir leur propre politique en la matière. Pour compléter ces documents, un formulaire type répondant aux exigences du FNS a par ailleurs été mis à la disposition des chercheurs. Un portail national sur lequel on peut trouver divers documents de référence a en outre été ouvert (www.dlcm.ch).

2. La gestion active des données

Conserver des données en vue de leur partage n'a de sens qu'à deux conditions. La première est de savoir ce qu'elles contiennent et la seconde est de pouvoir les retrouver facilement. *« Dans les laboratoires et les centres de recherche, les données brutes subissent de nombreux traitements avant d'être analysées,* explique Pierre-Yves Burgi. *Or, il est essentiel que ce*



KEYSTONE / SCIENCE PHOTO LIBRARY

processus de transformation soit documenté selon une procédure bien établie non seulement pour permettre leur identification et leur réutilisation mais également pour assurer la reproductibilité des résultats.» C'est dans cette perspective que les équipes du Programme P5 ont élaboré une série de directives pratiques portant sur deux outils de plus en plus répandus dans le monde scientifique : les cahiers de laboratoire électroniques (ELN) et les systèmes de gestion des informations de laboratoire (LIMS).

Raccordé directement sur les instruments de mesure scientifiques (spectromètre, IRM, scanner ou microscope électronique), le LIMS capte les données à la source via une interface et assure leur gestion ainsi que leur traçabilité. Complémentaire, l'ELN permet ensuite de stocker, de relier et d'annoter toutes les données numériques générées au cours du processus de recherche.

«Bien qu'elles soient principalement basées sur nos expériences avec les laboratoires des sciences de la vie, les directives que nous avons publiées peuvent facilement être étendues à d'autres domaines de recherche», note Pierre-Yves Burgi.

Afin de déterminer quel outil est le plus approprié pour tel ou tel laboratoire, les chercheurs du programme P5 ont par ailleurs dressé une liste des différents produits actuellement disponibles sur le marché et peuvent offrir leur expertise à toute personne intéressée.

3. La sauvegarde des données

Principalement collectées pour répondre aux besoins des chercheurs, et non en vue de leur partage ou de leur préservation, les données scientifiques se présentent également sous des formats très divers (fichiers vectoriels, vidéos, audios, images, textes, graphiques, streams, etc.). Cette double complexité est loin de faciliter leur archivage.

«Les bibliothèques disposent d'un savoir-faire très précieux dans ce domaine, explique Pierre-Yves Burgi. *Notre idée est de reprendre en grande partie le travail de standardisation qui a été effectué pour les publications et de l'appliquer aux données.»*

En utilisant notamment les LIMS (voir ci-contre) afin de disposer de points de contrôle là où on s'attend à ce que le chercheur donne des informations sur les données, le

LA COLLECTION BODMER S'OUVRE AU MONDE

Fondé en 2015 au sein de l'Université de Genève, le Bodmer Lab s'est donné pour objectif de numériser une partie importante des ouvrages rares qui constituent le cœur de la collection construite entre 1920 et 1971 par le bibliophile suisse Martin Bodmer. Deux ans et demi après le début des travaux, 220 000 pages, soit environ 1500 ouvrages, sont d'ores et déjà accessibles à la communauté des chercheurs de l'UNIGE.

«Notre idée n'est pas de produire des données au kilomètre pour ensuite les voir se perdre dans l'immensité de la Toile, mais de donner une seconde vie à ces objets en profitant des opportunités offertes par les technologies numériques», explique Radu Suciu, collaborateur scientifique à la Faculté des lettres et codirecteur du Bodmer Lab. Pour donner corps à cette vision, les responsables du Bodmer Lab ont accepté dès le lancement du projet de servir d'étude de cas aux concepteurs du projet *Data Life Cycle Management* (ou DLCM, lire ci-dessus).

Cette alliance a tout d'abord permis d'adopter un système – le même que celui utilisé pour les archives ouvertes de l'UNIGE – garantissant la sécurité des données numérisées sur le long terme et respectant les standards internationaux actuellement en vigueur. Elle a par ailleurs été capitale pour assurer leur interconnectivité. Après le scannage des 70 000 fiches papier du catalogue originel de la collection, les informaticiens du Bodmer Lab, en collaboration avec les équipes du DLCM ont créé un outil permettant de lier les données émanant de la Fondation Bodmer au répertoire de l'European Library, qui rassemble toutes les métadonnées des bibliothèques européennes dans une immense base de données. Facilitant grandement le processus d'indexation des documents et l'établissement d'un nouveau catalogue numérique, le procédé permet également de faire cohabiter sur une même plateforme des ensembles de données provenant des différentes institutions. Autrement dit : de comparer une édition rare de Shakespeare appartenant à la Fondation Bodmer avec un ouvrage similaire conservé à la Bibliothèque nationale de France ou à la British Library.

La plateforme numérique du Bodmer Lab, qui sera lancée à l'été 2018, proposera au visiteur deux types de navigation. La première – qui repose sur un moteur de recherche type « Google » – s'adresse aux chercheurs expérimentés qui savent ce qu'ils cherchent. Elle permet d'arriver rapidement aux documents désirés et de les visualiser, si besoin en les comparant avec un document provenant d'une autre bibliothèque.

La seconde, destinée à un public plus large et aux intentions moins définies, est basée sur des dossiers thématiques parfois agrémentés de vidéos explicatives.

«En apportant un certain nombre d'informations contextuelles, nous voulons guider le public pour éviter que les gens ne se perdent en chemin, explique Radu Suciu. *Mais il est aussi essentiel de lui laisser la liberté de tomber sur des choses inattendues et de favoriser ainsi une certaine forme de sérendipité. Nous serions d'ailleurs ravis si les documents que nous rendons accessibles servaient à des travaux graphiques ou à des expériences artistiques comme nous en avons fait l'expérience l'an dernier dans le cadre d'une collaboration avec la Haute école d'art et de design de Genève (HEAD).»*

http://bodmerlab.unige.ch/

AFIN D'ASSURER LA PRÉSERVATION DES DONNÉES DANS LA DURÉE, LES CHERCHEURS DU PROGRAMME P5 VISENT LA CRÉATION D'UN «CLOUD» NATIONAL QUI DEVRAIT ÊTRE OPÉRATIONNEL À PARTIR DE L'ÉTÉ 2018.

processus pourrait être largement automatisé. Il demandera cependant un certain nombre d'interventions manuelles. C'est ainsi au chercheur qu'il reviendra de décrire la structure des données en question, leur utilité et les modalités liées à leur utilisation.

Autre piste envisagée : l'assignation d'un identifiant comparable à celui dont est dotée chaque publication scientifique, de manière à ce qu'il soit possible d'associer à un article les données utilisées au cours du travail de recherche.

Afin d'assurer la préservation de ces innombrables fichiers numériques dans la durée, les chercheurs du programme P5 visent la création d'un « cloud » national qui devrait être opérationnel à partir de l'été 2018. Répondant à la norme OAIS *(Open Archival Information System),* ce dernier est basé sur un algorithme capable de fonctionner malgré un nombre assez élevé de bugs dans le système comparable à celui qui a été mis en place l'an dernier pour assurer la pérennité des thèses des archives ouvertes de l'UNIGE.

« Les thèses produites au sein de l'Université n'ont, pour la plupart, pas été publiées ailleurs, précise Pierre-Yves Burgi. *Elles ont donc une valeur patrimoniale qui doit être préservée. Nous avons opté pour un système à sept copies qui assure un taux de sécurité maximum. Le problème, c'est qu'avec des fichiers plus gros contenant des données de recherche, le coût des copies, qui est à la charge du chercheur et/ou de l'institution, peut devenir prohibitif. »*

Plus réaliste, la solution vers laquelle le projet se dirige devrait se contenter de trois copies, ce qui permettra également de rester compétitif sur un marché en pleine ébullition. En effet, outre les grands éditeurs scientifiques (Elsevier, Springer…) intéressés par cette nouvelle manne, certaines sociétés comme Amazon se sont également lancées dans la bataille avec des offres très agressives. *« Amazon propose un service avec des tarifs très attractifs pour le stockage à long terme mais qui s'envolent lorsqu'il s'agit de récupérer des données, prenant en quelque sorte les chercheurs au piège,* commente Pierre-Yves Burgi. *C'est un scénario que l'on veut éviter à tout prix. »*

4. La formation et l'expertise

Afin d'accompagner au mieux les chercheurs dans cette mutation appelée à transformer leurs habitudes de travail, les équipes du programme P5 ont mis sur pied des formations visant notamment à expliquer les tenants et les aboutissants du fameux DMP. Dispensées aussi bien à Genève qu'à l'EPFL ou à l'Université de Zurich, elles s'adressent en priorité aux doctorants et autres membres de la relève académique. Par ailleurs, un bachelor sur la conservation des données est à l'étude au sein de la Haute école de gestion (HEG). Un helpdesk sera également mis sur pied. Il permettra de trouver de l'assistance, par exemple pour remplir le questionnaire du FNS, et, si besoin, de se voir diriger vers un expert.

5. La sensibilisation et la communication

En plus de faire connaître les prestations offertes par le DLCM à l'ensemble de la communauté des chercheurs, les responsables du projet souhaitent élargir leur réseau à toutes les hautes écoles de Suisse. Ils entendent également développer de nouvelles collaborations, notamment avec SWITCH, la fondation qui a mis sur pied les autoroutes de l'information au niveau des hautes écoles, ou le nouveau Science Data Center commun aux deux écoles polytechniques fédérales. Une structure au sein de laquelle vont prochainement œuvrer une cinquantaine de « data scientists » chargés de développer les pratiques d'analyse de données.

* Écoles polytechniques de Zurich et de Lausanne, Université de Zurich, de Lausanne et de Bâle, Haute école de gestion/Haute école spécialisée de Suisse occidentale, Fondation SWITCH.