Data Life-Cycle Management: The Swiss Way

Burgi, Pierre-Yves

Data Life-Cycle Management: The Swiss Way

Pierre-Yves Burgi, directeur SI adjoint, Université de Genève

L'information contenue dans les données de recherche constitue un bien précieux pour le chercheur. En revanche, le chercheur ignore trop souvent l'importance du cycle de vie de ces données, un concept qui englobe une multitude de facettes, dont certaines dépendent de la discipline considérée.

D'une manière générale, ce cycle débute par l'acquisition de données dites «brutes»[1]. Par la suite ces données sont analysées pour les confronter à des hypothèses et modèles. Les résultats de ces études conduisent généralement à des publications, un jalon important dans la vie des données, tout comme la fin d'un projet (ou de son financement). En effet, ces données sont alors souvent délaissées, typiquement sur l'espace de stockage du chercheur, et/ou d'un serveur institutionnel, pour être rapidement oubliées, en omettant de les valoriser dans d'autres contextes.

Collecter, sélectionner et sécuriser

Le phénomène de «numérisation» de la recherche s'accélérant, il devient urgent d'aborder le cycle de vie complet des données. A titre d'exemple, le radiotélescope SKA (Square Kilometer Array[2]) devrait, dès 2024, collecter chaque jour 14 exabytes[3] de données dont 1 petabyte sera sauvegardé quotidiennement. Cet exemple illustre les besoins gigantesques en stockage que certains consortia internationaux devront couvrir dans un futur proche. La plupart des laboratoires de recherche en Suisse ont certes des besoins plus modestes, couramment décrits comme «the long tail»[4]. Reste que la gestion de cette information, aussi minime soit-elle, est sujette au second principe de la thermodynamique selon lequel le désordre (ou entropie) a tendance à irrémédiablement augmenter, avec pour conséquence que toute donnée devient inévitablement corrompue avec le temps. Cela implique que pour conserver l'information il faut contrecarrer ce désordre naturel. En l'occurrence, l'ordre peut être rétabli par la capacité de la matière à effectuer des calculs pour rétablir une information corrompue[5]. Cependant ce mécanisme a des limites, ne serait-ce que par l'énergie énergie requise pour son fonctionnement. Ainsi une sélection de l'information devient une nécessité, tout comme la durée de l'archivage dans le temps.

Origine du projet et ses intervenants

Le projet Data Life-Cycle Management (DLCM) s'est concrétisé dans le contexte du programme national suisse CUS P-2 (2013–2016) «Information scientifique: accès, traitement et sauvegarde». En novembre 2013, l'auteur de cet article a initié des contacts avec des experts du domaine DLCM dans les universités suisses dans le but de former un partenariat avec l'objectif de déposer une proposition de projet CUS P-2 en 2014. Au terme de cette démarche, un partenari-

1 Une dénomination ambiguë par le fait que les instruments scientifiques prétraitent souvent les données à différents niveaux. La NASA a par exemple défini 6 niveaux de prétraitement (science.nasa.gov/earth-science/earth-science-data/data-processing-levels-for-eosdis-data-products).

2 www.skatelescope.org

3 1 Exabyte (EB) correspond à 1000 Petabytes (PB), soit 1 000 000 Terabytes (TB). Aujourd'hui, un disque de 1 TB coûte environ CHF 100.

4 «The long tail» fait référence à la loi de Pareto selon laquelle 80% des effets sont le produit de 20% des causes (e.g., fr.wikipedia.org/wiki/Principe_de_Pareto). Pour les données de la recherche, cela se traduit par le postulat que 20% des producteurs de données comptent pour 80% du volume produit; «the long tail» rappelle qu'il ne faut pas oublier les 80% des chercheurs qui produisent plus modestement des données, mais constituent une population importante ayant des besoins avérés en DLCM.

5 Sachant que cette même matière est aussi sujette à l'augmentation de l'entropie et nécessite aussi un traitement pour être maintenue opérationnelle.

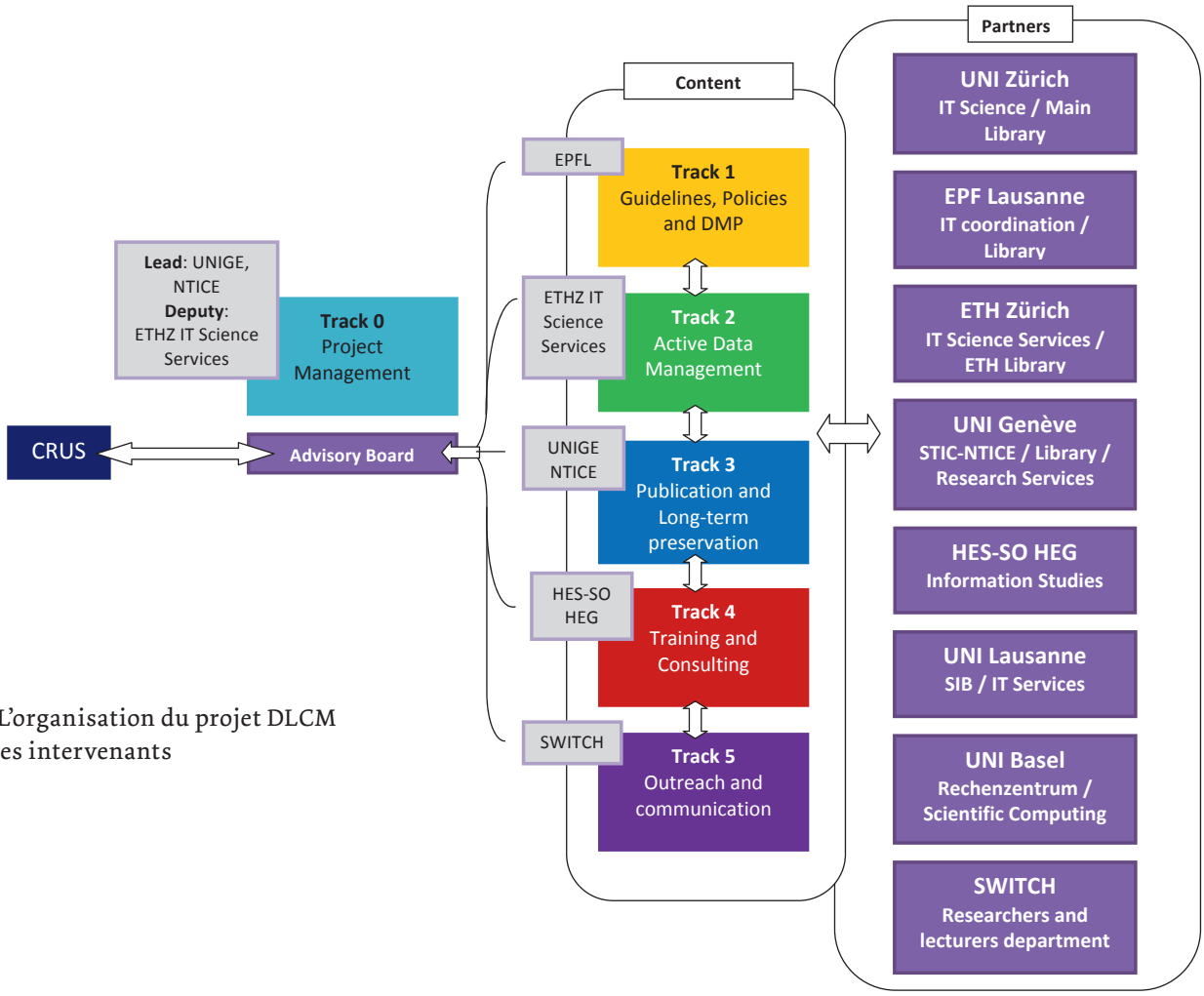

Fig. 1 L'organisation du projet DLCM avec ses intervenants

at entre les deux écoles polytechniques fédérales (ETH-Z et EPF-L), les universités de Zurich, Bâle, Lausanne, et Genève, la HEG de la HES-SO, et SWITCH s'est finalement établi (Fig. 1), avec à la clé une proposition aboutie et soumise en février 2015. Suite à l'acceptation de cette proposition en juillet 2015, le démarrage officiel du projet a eu lieu le 1er septembre 2015.

Objectifs du projet

Concrètement, le projet DLCM aborde tous les aspects du cycle de vie des données, à savoir l'acquisition (e.g., LIMS[6], ELN[7]), la normalisation et description des données en vue d'un référencement pérenne, leur traitement, et conservation sur le long terme (Fig. 2). Ces différentes étapes du cycle de vie sont couvertes dans cinq parties distinctes (Fig. 1 et 2):

Le «track 1» traite de la dimension politique du DLCM («guidelines, policies»), de plus en plus présente dans les instances qui financent la recherche (e.g., Horizon 2020, Fonds national suisse, National Institutes of Health, etc.), et qui nécessite l'écriture de «data management plans» (DMP);

Le «track 2» porte sur le traitement «actif» des données;

Le «track 3» considère la publication et conservation (long terme) des données;

Le «track 4» concerne la formation des utilisateurs sur l'usage des outils du DLCM et l'élaboration de DMP;

Le «track 5» traite de la dissémination au niveau national des services nouvellement établis.

6 Laboratory Information Management System (LIMS).

7 Electronic Laboratory Notebook (ELN).

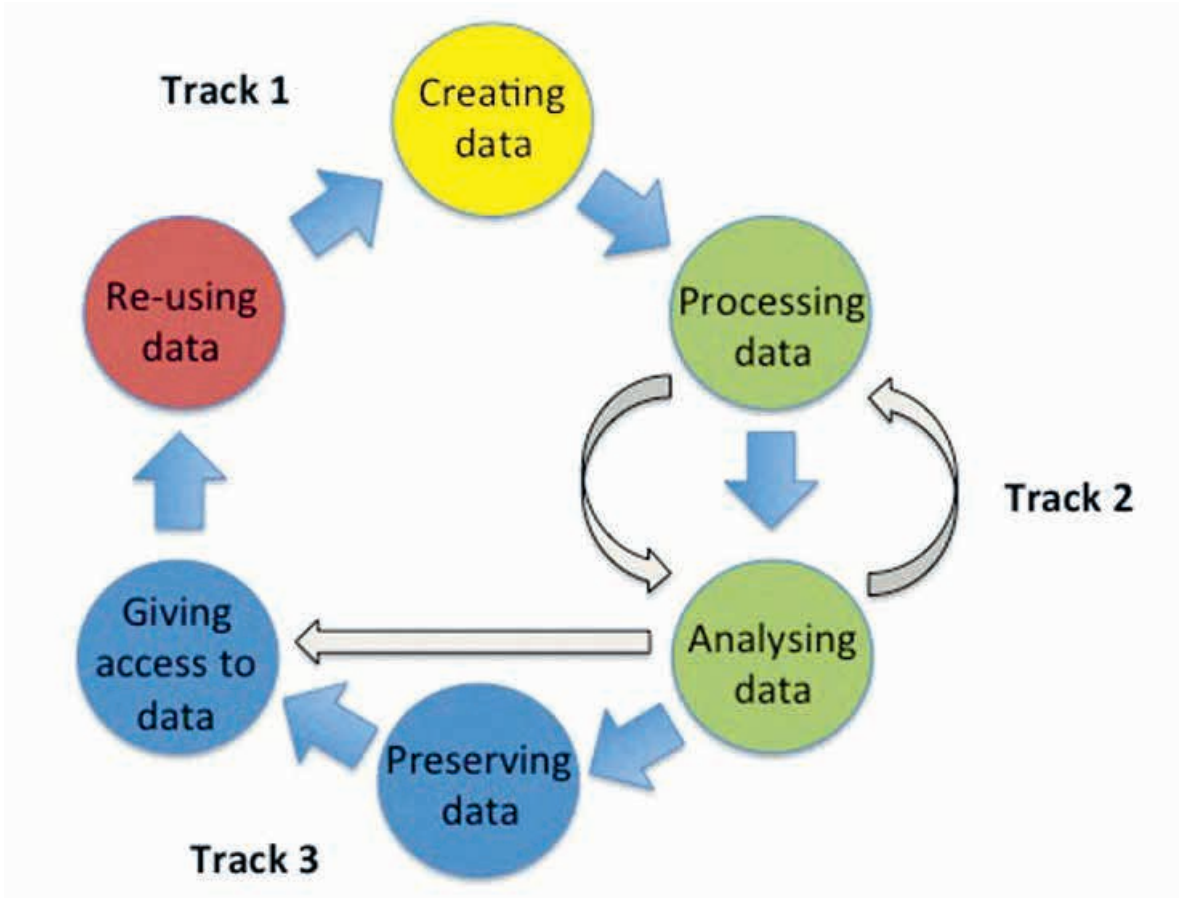

Fig. 2 Les éléments clés du cycle de vie des données du projet DLCM

Les axes forts du projet

Un point vital du programme CUS P-2 est qu'au terme des différents projets de nouveaux services nationaux puissent être proposés à la communauté universitaire. Le projet DLCM, défini sur 3 ans, a pour ambition de traiter chaque étape du DLCM, avec à la clé la mise en place d'une pluralité de services construits sur la base d'une collaboration entre des informaticiens, des bibliothécaires, et des services à la recherche. Pour éviter cependant de trop se disperser dans les objectifs, deux disciplines «pilotes» sont considérées: les sciences de la vie et les humanités numériques. Ces disciplines, de prime abord antinomiques, vont au contraire forcer une conceptualisation de plus haut niveau de l'information, pour conduire à des solutions plus génériques. Par exemple, des réseaux sémantiques basés sur le modèle RDF sont agnostiques au domaine considéré et s'appliquent aussi bien en génomique qu'en sciences humaines. Reste que pour développer des services nationaux «génériques» couvrant le domaine du DLCM à large échelle, des modèles économiques robustes sont encore à concevoir, ce représente un défi supplémentaire.

L'auteur

Pierre-Yves Burgi

Dr Pierre-Yves Burgi est directeur SI adjoint, responsable depuis juin 2003 du service des Nouvelles Technologies de l'Information, Communication, et Enseignement (NTICE) dans la division du Système de l'Information de l'Université de Genève. Il a reçu son diplôme d'ingénieur en informatique de l'Ecole Polytechnique Fédérale de Lausanne en 1986, et le titre de docteur ès sciences de l'Université de Genève en 1992. Ses études doctorales ont été suivies par un post-doctorat d'une durée de 5 ans en Neurosciences de la vision, d'une part au Smith-Kettlewell Eye Research Institute, San Francisco, CA, et d'autre part au CNRS à l'Université Paul Sabatier, Toulouse, France.