Le Projet de Loi 12146 : Infrastructures et services numériques pour la recherche

Pierre-Yves Burgi :
Pierre-Yves.Burgi@unige.ch
https://orcid.org/0000-0002-4956-9279
Directeur adjoint des systèmes d'information, Université de Genève

Résumé

Le projet de loi (PL) 12146 du canton de Genève, en vigueur depuis janvier 2018, porte sur les infrastructures et services numériques pour la recherche. Une thématique d'actualité, sa rédaction en 2011 a nécessité une approche visionnaire, favorisée par le contexte du moment, aussi bien au niveau de l'Université de Genève (UNIGE) qu'au niveau suisse avec les prémisses en 2011 du prochain programme fédéral « CUS P2 2013-2016 : Information scientifique : accès, traitement et sauvegarde ». L'UNIGE a joué dès le début du programme fédéral un rôle déterminant avec le lancement du projet « Data Life-Cycle Management » (DLCM) en 2015, un des plus importants en nombre d'institutions partenaires (9 au total). Ce projet est entré dans sa deuxième phase en 2018. Le PL et le projet DLCM progressent par conséquent en synergie : la technologie de préservation long terme développée dans le projet DLCM a servi d'accélérateur pour la mise en production de Yareta, le système d'archivage long terme des données de recherche pour les institutions universitaires du canton de Genève. L'expérience acquise dans l'exploitation de Yareta va à son tour bénéficier à la mise en place de l'instance nationale, dénommée « OLOS », prévue début 2020.

Abstract

The bill 12146 of the Canton of Geneva, in force since January 2018, concerns digital infrastructure and services for research. A hot topic, its drafting in 2011 required a visionary approach, favored by the current context, both at the University of Geneva (UNIGE) and at the Swiss level, with the premises being laid in 2011 for the next federal programme "CUS P2 2013-2016: Scientific information: access, processing and preserving". Indeed, since the beginning of federal programme, UNIGE has played a decisive role with the launch of the "Data Life-Cycle Management" (DLCM) project in 2015, one of the largest in number of partner institutions (9 in total). This project entered its second phase in 2018. The bill 12146 and DLCM project are therefore progressing in synergy: the long-term preservation technology developed in the DLCM project served as an accelerator for the production of Yareta, the long-term archiving system for research data for academic institutions of the canton of Geneva. The experience acquired in the operation of Yareta will in turn benefit the est ablishment of the national version, called "OLOS", scheduled for early 2020.

Mots-clés

PL12146, HPC, Préservation long terme, OAIS, Données de Recherche, Programme CUS P2 et swissuniversities P5, Projet DLCM, Humanités numériques, Yareta

1. Introduction

La révolution numérique impacte et transforme la recherche scientifique dans tous les domaines, aussi bien dans les sciences dites dures que dans les sciences sociales et humaines. Les découvertes et avancées scientifiques majeures ne sont aujourd'hui plus possibles sans disposer de services et d'infrastructures informatiques à haute performance permettant la simulation numérique de phénomènes complexes ainsi que la recherche axée sur une exploitation intensive des données numériques (Hey et al., 2009). Afin de répondre à ces nouveaux défis et rester compétitive et attractive, il est impératif que l'Université de Genève (UNIGE) ainsi que les autres HE(1) du canton de Genève puissent mutualiser et développer les services et infrastructures numériques à disposition de tous leurs chercheurs, tout en gagnant en efficience. Cela nécessite des efforts coordonnés et des financements adéquats pour mettre en œuvre des solutions concrètes qui s'intègrent harmonieusement aux environnements des chercheurs et répondent aux exigences des bailleurs de fonds de la recherche scientifique.

Ces développements s'inscrivent dans la vision de l'UNIGE à l'horizon 2025 et contribuent pleinement au projet stratégique transversal de l'Université numérique (https://www.unige.ch/plan-strategique). Il est cependant utile de rappeler l'historique du Projet de Loi 12146, antérieur à cette vision numérique, comme expliqué dans la section suivante.

2. Historique

Les prémisses du projet de loi remontent à l'été 2011. Avec mon collègue Jean-François Rossignol, alors responsable des infrastructures informatiques de l'UNIGE, nous avions rédigé une première version du projet de loi (PL) sur les thèmes du calcul haute performance (HPC) et du Research Data Management (RDM). Ces deux thèmes avaient déjà été jugés importants, tant l'analyse et la conservation des données scientifiques nous semblaient complémentaires. Sur la base de cette version préliminaire, nous avions organisé le 13 octobre 2011 une présentation à la Commission informatique (COINF) de l'UNIGE des concepts et idées contenus dans ce nouveau projet.

La COINF est la commission consultative du Rectorat en matière de politique institutionnelle des systèmes d'information (SI) et des services numériques qui lui sont associés. Elle oriente l'évolution des SI au sein de l'institution en tenant compte des besoins de l'enseignement et de la recherche ainsi que de l'administration universitaire, tout en veillant à l'optimisation des ressources. Elle favorise une large participation facultaire de par sa composition et son articulation avec les Commissions Informatiques Facultaires (CIF). Aussi, suite à cette présentation, il était naturel de présenter le projet plus largement dans les CIF des différentes facultés. Cette consultation plus large des chercheurs au sein de leur faculté a duré plusieurs mois et a permis de dresser une image qualitative plus précise des besoins en HPC et RDM. Il est cependant intéressant de relever qu'à cette période la question de la gestion des données de recherche n'était pas nécessairement une priorité et que l'information récoltée sur ce sujet n'a pas amené des éléments majeurs dans la rédaction du PL. Tout au plus, cette information n'a pas infirmé ce que nous avions déjà rédigé.

Quant au HPC, les informations récoltées sont venues renforcer une enquête quantitative réalisée précédemment au printemps 2011 qui était axée sur les besoins en calculs scientifiques (simulations, analyse de données, etc.). En effet, durant les 15 années précédentes, le nombre de chercheurs utilisant l'informatique pour la modélisation et l'analyse de données n'a cessé de croître. Plusieurs dizaines de serveurs de calculs ont été acquis et se sont retrouvés dispersés dans plusieurs salles machines de l'Université, voire dans des bureaux et laboratoires. Cette dispersion entraînait de nombreux problèmes et désagréments, notamment de fortes difficultés à assurer une administration système efficace. En vue de regrouper les machines servant au calcul, une enquête des besoins en termes de consommation électrique et d'espace nécessaire pour la période 2012 à 2015 a donc été réalisée. La tâche s'est avérée complexe, car les différentes structures de l'Université sont habilitées à acquérir des machines de calcul de manière décentralisée (achats souvent effectués au niveau des groupes de recherche ou des départements). Par ailleurs, les besoins futurs sont extrêmement difficiles à estimer à priori, vu le caractère hétérogène des applications et des équipements. L'enquête a néanmoins révélé un fort besoin des utilisateurs pour un serveur de calcul centralisé. Les réponses ont indiqué un besoin de plus de 30 millions d'heures de calcul annuel, qui correspond à une ferme de calcul d'environ 4'000 cœurs(2).

L'évolution technologique a été intégrée dans les prévisions des besoins en appliquant la loi de Moore qui prévoit que le nombre de transistors pouvant être placés sur une puce à un prix acceptable est multiplié par 2 tous les 2 ans (Moore, 1965). Le corollaire de cette augmentation de la densité des composants électroniques est que la consommation électrique et la place occupée par processeur sont divisées en principe par 2 tous les 2 ans (mais voir Waldrop, 2006 pour l'évolution de cette loi de Moore qui se modifie dans le contexte technologique actuel). Au vu de ces estimations, on a pu estimer les caractéristiques nécessaires d'une salle machines regroupant tous les serveurs de calcul de l'Université, avec une projection de 35 racks pour une puissance totale de 520 KW en 2015. À noter qu'à ce jour (2019), et grâce aux progrès technologiques, la ferme (serveurs) de calcul de l'UNIGE (dénommée « Baobab ») possède de l'ordre de 4'400 cœurs, occupe 6 racks et consomme environ 60 kWh. L'étape suivante (d'ici à 2020) vise à acquérir 3'000 cœurs de nouvelle génération, plus puissants et donc nécessitant moins de cœurs comparé à la génération précédente, dans 3 racks pour une consommation d'environ 37 kWh.

Ce travail d'étude des besoins, mené sous l'égide de la COINF et en étroite collaboration avec les facultés, a conduit à préaviser positivement en mai 2012 la nécessité d'une part de réaliser ces investissements majeurs en infrastructure pour répondre aux nouveaux besoins de la recherche et d'autre part de consolider à terme des postes d'ingénieur possédant des compétences pointues en HPC, indispensables pour les chercheurs pour mener à bien leurs calculs.

Dans sa séance du 18 juin 2012, le Rectorat a quittancé ce travail et décidé d'aller de l'avant avec le dépôt auprès de l'Etat de Genève du PL pour l'ouverture d'une subvention d'investissement de l'ordre de 15 millions sur 5 ans, couvrant l'achat d'équipements informatiques de calcul et de stockage, ainsi que l'engagement du personnel requis pour le développement et la mise en œuvre des nouveaux services. La procédure de dépôt a été ajustée avec la Présidence du Département de l'Instruction Publique (DIP) en réunion DIP-UNIGE-HESGE(3). La version définitive du « Projet de Loi HPC & DM » a été transmise le 29 novembre 2012 à la division des finances de l'UNIGE pour dépôt auprès de la direction des finances du DIP le 21 mai 2013.

N°DOI : 10.55790/journals/ressi.2019.e2006

Ce dossier a néanmoins été bloqué jusqu'en juillet 2016, date à laquelle le vice-recteur Denis Hochstrasser, en charge du système d'information de l'UNIGE, a transmis un argumentaire au DIP afin d'appuyer le PL dans le cadre du plan d'investissement de l'Etat. Le dépôt du PL par le Conseil d'Etat est devenu effectif au 21 juin 2017 avec une re-planification du projet sur la période 2018-2022. Le 22 septembre 2017 le Grand Conseil a envoyé la proposition de loi à sa commission des travaux pour être mis à l'ordre du jour du Grand Conseil du 23-24 novembre(4). Elle a finalement été acceptée sans débat le 24 novembre 2017, pour un début officiel qui a été fixé au 1er janvier 2018.

3. Périmètre

Le périmètre du PL a été conçu afin de répondre au 4ème paradigme de la recherche que représente l'utilisation intensive des données pour progresser dans les découvertes scientifiques et qui vient compléter les méthodes classiques que sont l'expérimentation, la théorie et la simulation (Hey et al., 2009). Cette nouvelle manière de faire ne concerne pas uniquement les sciences dites « dures » (physique, astronomie, génomique, informatique, neurosciences, sciences de l'environnement, etc.), mais également les sciences sociales et humaines. En effet, par le biais de ce que l'on nomme les « digital humanities » (humanités numériques), la puissance de calcul des ordinateurs rend possible l'examen de larges corpus (textes ou autres types de média), y compris de millions de livres numérisés. Bien qu'émergente, cette approche transdisciplinaire, qui bénéficie de l'initiative du libre accès, et plus généralement de « l'Open Science », conduit déjà à de nouvelles connaissances en linguistique, en histoire, en archéologie, dans l'interprétation d'anciens textes, etc. (see e.g., Borgman, 2010).

Ce 4ème paradigme représente donc un élément déterminant pour l'exploration et le progrès de la connaissance en général, ainsi que pour promouvoir une recherche interdisciplinaire à fort potentiel pour résoudre des problèmes scientifiques et sociétaux complexes. Pour faciliter le passage à ce nouveau paradigme, l'évolution des infrastructures et services associés de calcul à haute performance et de stockage long terme afin d'optimiser et faciliter l'utilisation des données issues de la recherche dans les hautes écoles universitaires genevoises constitue l'essence du PL 12146. Aussi, ce PL se décline en quatre objectifs prioritaires :

1. Mettre en place une infrastructure pouvant répondre de manière optimisée aux besoins en matière de calcul scientifique et de gestion du cycle de vie des données de la recherche (qui comprend également la gestion des données dites « actives ») ;
2. Mettre en place une architecture de stockage sécurisée construite sur les standards internationaux permettant la conservation à moyen et long terme des données scientifiques ;
3. Développer des interfaces logicielles qui répondent aux besoins des chercheurs et facilitent l'utilisation de ces infrastructures aussi bien pour le calcul que pour le dépôt, la gestion, et l'accès aux données ;
4. Concevoir des environnements informatiques favorisant la collaboration entre chercheurs, facilitant l'exploration, la visualisation et le partage des données, ainsi que leur utilisation dans l'enseignement.

N°DOI : 10.55790/journals/ressi.2019.e2006

Le principe général est de mettre à disposition des membres de la communauté scientifique, au travers d'un « cloud académique », des infrastructures et services mutualisés et sécurisés. En effet, plutôt que de continuer de financer des infrastructures dédiées exclusivement à des domaines scientifiques particuliers, qui s'avèrent souvent insuffisantes, la voie prise par le « cloud académique » est de mutualiser autant que possible les différentes couches de services, à savoir IaaS (Infrastructure as a Service), PaaS (Platform as a Service), et SaaS (Software as a Service), afin d'accroître et d'optimiser l'offre de services.

D'autre part, afin de permettre un usage adéquat de ces nouvelles infrastructures et des logiciels associés, des services d'accompagnement et d'expertise doivent être mis en place pour aider les chercheurs : conseils en HPC, interfaces utilisateurs, environnements informatiques favorisant la collaboration, l'échange de données et leur utilisation en conformité avec les nouvelles exigences des bailleurs de fonds et des éditeurs, etc.

Selon la volonté du DIP, ces infrastructures et services mutualisés, gérés par l'UNIGE, sont ouverts aux partenaires académiques genevois (HES-SO Genève et IHEID). Ils sont également conçus pour s'intégrer dans l'écosystème national, mais aussi au-delà, du fait qu'ils sont basés sur des standards internationaux. À noter que pour des raisons de pérennité des solutions informatiques, les logiciels sous-jacents à ces infrastructures sont pour la majorité des logiciels libres(5) soutenus par des communautés internationales. Il serait en effet illusoire de compter sur des solutions propriétaires, sujettes aux aléas des marchés économiques.

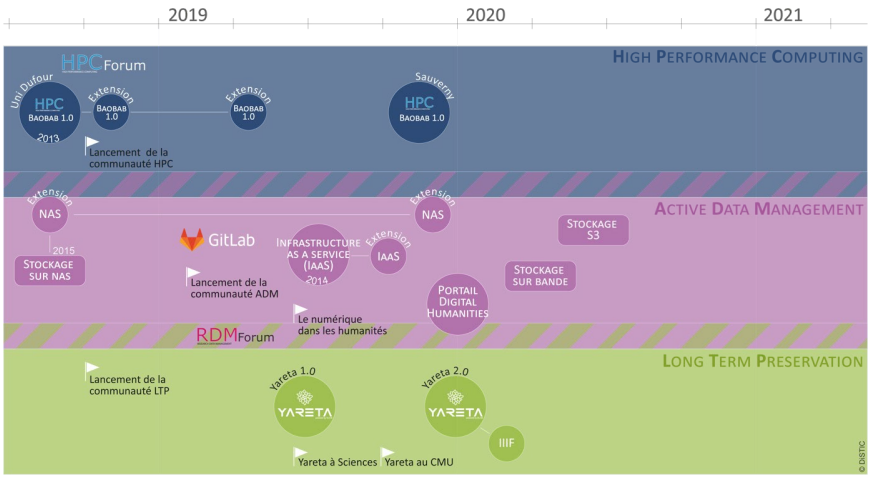
Plus concrètement, ces quatre objectifs prioritaires se déclinent par :

- Le renouvellement, l'extension et l'amélioration de la ferme de calcul Baobab, à savoir augmenter la puissance de calcul et la capacité de stockage pour les données actives, diversifier le matériel avec l'ajout de GPU(6), et bénéficier des nouvelles technologies de serveurs moins énergivores ;
- Du « cloud computing » permettant aux chercheurs d'utiliser les ressources d'une manière flexible, selon leurs besoins, à la demande ;
- Du conseil et de l'expertise en HPC afin de garantir aux chercheurs une meilleure adéquation entre leurs algorithmes et l'architecture des machines disponibles ;
- L'archivage des données selon des standards internationaux en vigueur dans le domaine archivistique (tels que l'OAIS Reference Model – ISO 14721:2012), permettant la conservation à moyen terme (moins de 10 ans) et long terme (au-delà de 10 ans) des données scientifiques, répliquées sur plusieurs sites dans différentes technologies ;
- L'accompagnement des chercheurs dans leurs projets d'exploitation de leurs données selon de nouveaux algorithmes, particulièrement dans les sciences sociales et humaines (digital humanities), domaine qui est amené à se développer et dont le potentiel, largement sous-exploité à ce jour, intéresse de plus en plus les chercheurs de l'UNIGE (par exemple, en facultés des lettres, de théologie, de traduction et d'interprétation, et des sciences économiques et sociales) ;
- Le développement d'interfaces permettant aux étudiants et plus largement aux citoyens d'accéder et de visualiser des données issues de la recherche et rendues spécifiquement compréhensibles pour cette population d'utilisateurs.

La feuille de route (roadmap) de la réalisation de ces objectifs est présentée dans la Figure 1 pour la période 2018 à mi-2020 (le PL doit se terminer en décembre 2022).

N°DOI : 10.55790/journals/ressi.2019.e2006

Figure 1 : Feuille de route intermédiaire



4. Réalisations

Le PL se concentre sur deux axes principaux : pour commencer, un axe « Infrastructures », dans lequel l'acquisition de matériel est effectuée sur la base d'appels d'offres publics. Cela concerne l'évolution des serveurs de calcul de Baobab avec un objectif de 4'000 cœurs, l'extension du stockage « Network Attached Storage » (NAS) à 3.5 PB(7) (répliqué sur un deuxième site), et destiné au traitement des données actives (Active Data Management), de l'acquisition d'un robot pour le stockage sur bande de grands volumes de données (plusieurs dizaines de PB), ainsi que du stockage de type « S3 »(8) destiné à la préservation long terme.

Le deuxième axe du PL concerne le développement des environnements et interfaces logiciels permettant aux chercheurs de bénéficier des nouvelles infrastructures.

Après deux ans d'activité du PL, nous pouvons citer deux réalisations principales : Yareta, le système d'archivage long terme des données de recherche et la plateforme « DH » (Digital Humanities).

4.1. Yareta

Yareta (https://yareta.unige.ch) est le nouveau dépôt de données FAIR(9) (Findable, Accessible, Interoperate, Reusable – Wilkinson et al., 2016) disponible depuis juin 2019 pour la communauté de chercheurs genevoise, permettant de promouvoir le partage des données et la reproductibilité scientifique. « Powered by the DLCM technology » (https://www.dlcm.ch), Yareta se compose d'une architecture OAIS (norme ISO 14721:2012) ouverte et modulaire pour la conservation à long terme (voir Figure 2), qui est centrée sur l'utilisateur afin de faciliter le dépôt des données de recherche. De plus, il a été conçu pour s'intégrer facilement aux systèmes de gestion de l'information des laboratoires et/ou s'interfacer avec des environnements de gestion active des données (« ADM »), puisque basé sur la technologie « Web services » (ou API(10)) (voir Figure 3).

L'architecture est composée de 3 parties : la soumission, l'archivage, et la dissémination, qui correspondent respectivement aux composants « Submission Information Package » (SIP), « Archival Information Package » (AIP), et « Dissemination Information Package » (DIP) de la norme OAIS. Chaque sous-module (Pre-Ingest, Ingest, Archival Storage, Data Management,

Access) est indépendant et peut s'exécuter dans le cloud sur des serveurs distincts. Le DOI (Digital Object Identifier) d'un dataset peut être réservé lors du « pre-ingest », ou assigné lors de l'« ingest ». Les métadonnées au format Datacite (https://schema.datacite.org/), complétées par les données administratives PREMIS (https://www.loc.gov/standards/premis/), sont indexées et moissonnables au travers du protocole OAI-PMH. Lorsque les données sont soumises, le processus inclut entre autres le calcul d'un checksum assurant l'intégrité des informations, le contrôle par un antivirus, et l'identification et la qualification du format. Le module « data management » définit les modalités de stockage physique (nombre de réplications, technologie, durée, etc.).

En résumé, Yareta est une solution d'archivage long terme des données de recherche :

- non-commerciale, sur sol suisse et qui répond aux exigences FAIR du Fonds National Suisse (FNS) ;
- conforme aux normes internationales pour l'interopérabilité des données (OAIS, DOI, OAI-PMH, PREMISE, Datacite, etc.) ;
- compatible avec tous les formats en vigueur dans les différentes disciplines scientifiques et qui constitue donc un dépôt générique des données de recherche ;
- conçue « *the swiss way* », à savoir décentralisée tout en assurant une indexation à l'échelle nationale ;
- flexible quant au nombre de copies et leur stockage physique dans différentes technologies pour assurer la pérennité des données ;
- basée sur une technologie moderne qui s'interconnecte aux environnements des chercheurs (Web services).

Dans les objectifs futurs de développement de Yareta, on peut citer :

- l'amélioration de l'ergonomie des interfaces utilisateur ;
- l'extension des fonctionnalités offertes afin de faciliter l'accès aux données de recherche, par exemple le développement de "plug-in" permettant de visualiser les données scientifiques via le module « *International Image Interoperability Framework* » (IIIF) pour les humanités numériques, un module 3D pour la visualisation de molécules, etc. ;
- le développement d'un module pour implémenter la politique de préservation permettant de compléter le cycle de vie des données (migration, fin de vie, etc.) et une plus grande souplesse dans le choix du nombre de réplications et des supports physiques ;
- l'ajout de mécanismes de "data privacy" permettant de gérer les données sensibles.

Figure 2 : Architecture de Yareta

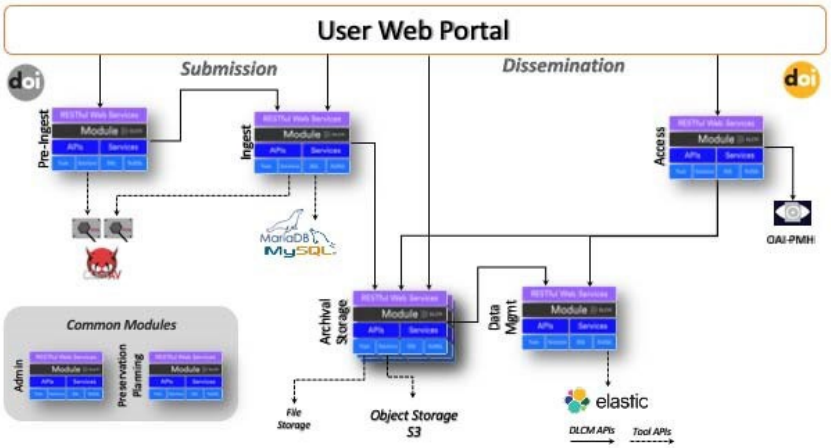
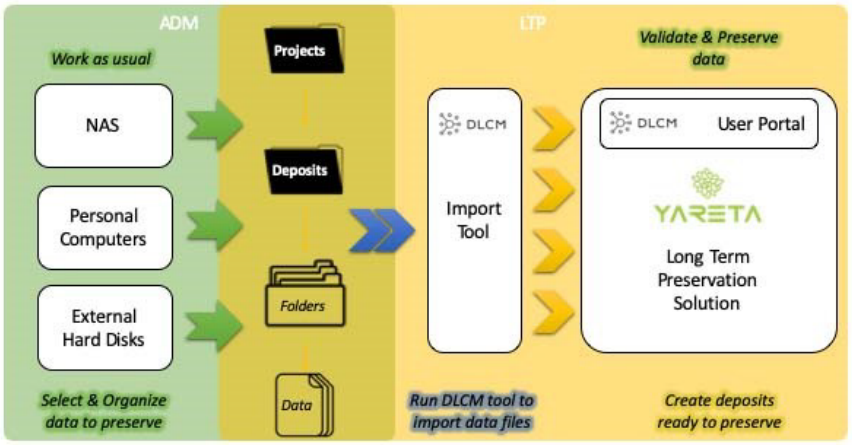


Figure 3 : De l'active data management (ADM) à l'archivage long terme (LTP)



4.2. La plateforme DH

La plateforme DH (Figure 4) a été conçue pour que les chercheurs en humanités numériques puissent gérer leurs données en format RDF(11) (Resource Description Framework). Ce format apporte beaucoup plus de possibilités que la forme plus traditionnelle, qui consiste à organiser les données dans des bases de données relationnelles. Développé par le W3C(12), RDF est le langage de base du Web sémantique représenté sous la forme de graphes composés d'associations « sujet, prédicat, objet », dénommées « triplets ». L'avantage d'une telle représentation est de pouvoir faire des requêtes dans un langage de plus haut niveau sémantique que celles des bases de données relationnelles (Structured Query Langage – SQL).

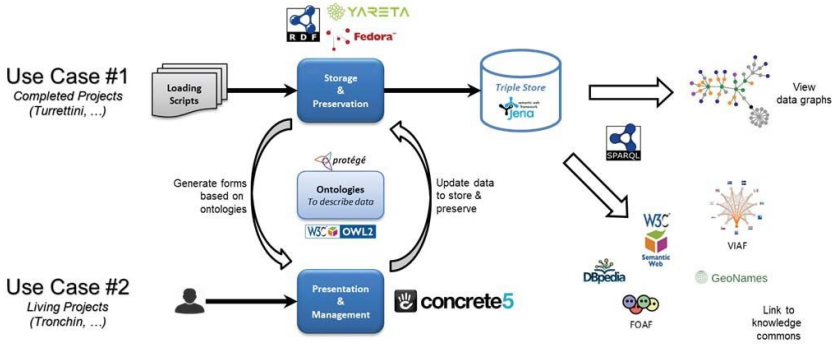
Dans la plateforme DH, deux approches sont proposées pour passer des bases de données relationnelles à la représentation RDF :

(1) une conversion automatique qui génère des relations sémantiques entre les objets, ces relations correspondant aux liens relationnels existants entre les différents champs de la base de données ; cette manière de faire ne bénéficie pas de tout le potentiel de la représentation sémantique, mais permet de minimiser le travail du chercheur lorsque les bases de données sont préexistantes ;

(2) la conception d'un modèle de données par les chercheurs avec l'aide d'experts du domaine RDF. Ce modèle est ensuite traduit en RDF.

Le format RDF permet de faire du « Linked data » (« Web des données » en français) qui consiste à publier sur le Web les données structurées non pas sous la forme de silos de données isolés les uns des autres, mais en reliant les données entre elles pour constituer un réseau global d'informations. L'usage d'ontologies prédéfinies disponibles sur le Web permet d'avoir des interprétations communes de ces informations. Cela permet par exemple de décrire des personnes (VIAF) et les relations qu'elles entretiennent entre elles (FOAF), ou des lieux géographiques (GeoNames), ainsi que des réseaux de connaissances (e.g., DBpedia).

Figure 4 : Architecture de la plateforme DH



Les données RDF sont gérées par le logiciel « Fedora Commons », le système avec lequel est construite l'Archive Ouverte de l'UNIGE (https://archive-ouverte.unige.ch/). La mise à jour des données RDF se fait via le logiciel Concrete 5, qui est le Content Management System (CMS) de l'UNIGE et avec lequel les pages Web de l'UNIGE sont gérées. L'avantage d'utiliser des outils institutionnels préexistants est double : parcimonie dans l'exploitation des outils et prise en main plus aisée par les utilisateurs qui les utilisent déjà dans d'autres contextes. Finalement, la préservation sur le long terme de ces données RDF est possible au travers d'un connecteur vers Yareta.

À ce jour (novembre 2019), deux projets pilotes bénéficient de cette plateforme DH : « Turrettini », la correspondance de Jean-Alphonse Turrettini, et « Tronchin », la publication des archives du collectionneur genevois François Tronchin.

5. Gouvernance

La gouvernance du PL se mène de manière transversale et agile dans laquelle la priorisation des tâches, contenues dans un « backlog », est essentielle (Al-Baik & Miller, 2015). Cette gouvernance, représentée dans la Figure 5, est par ailleurs construite sur la base de trois communautés de chercheurs (HPC, ADM, LTP) qui ont pour vocation d'impulser une dynamique ascendante et interdisciplinaire. Elle nécessite conjointement l'implication de nombreux partenaires et équipes d'expertises diverses (services « recherche », « information scientifique », « système d'information », etc.) travaillant en étroite collaboration afin d'assurer une adéquation avec les besoins des chercheurs des différentes facultés, ainsi que des institutions bénéficiant du PL. Pour ces raisons, une gouvernance impliquant des représentants de ces différentes entités a été mise en place sous la forme d'un comité de pilotage. Ce comité, qui se réunit environ 4 fois par année, est composé d'un représentant de chaque faculté et institution.

5.1. Modèle de coûts

Les services couverts par le PL12146 et pour lesquels un financement s'avérera nécessaire afin d'assurer leur exploitation et leur renouvellement à son terme sont les suivants :

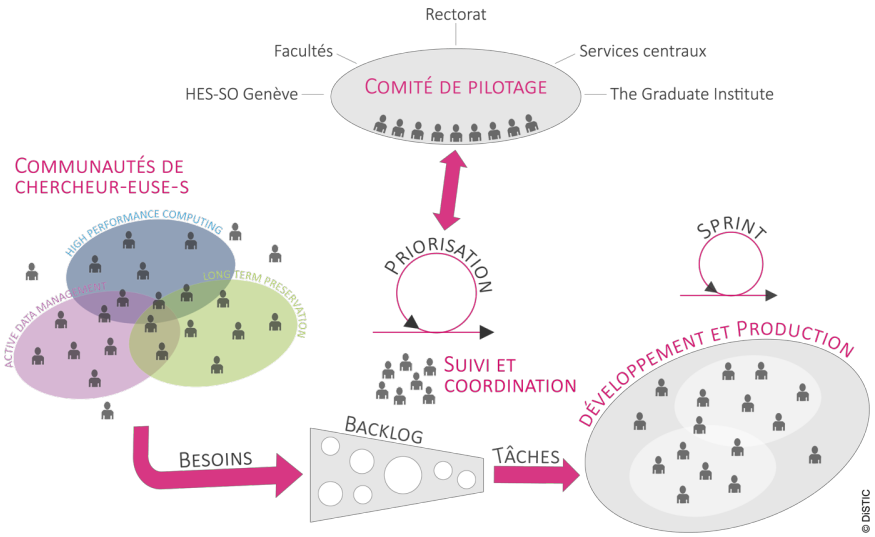
- Le calcul haute performance (HPC) ;
- Les solutions de type IAAS qui permettent aux chercheurs de bénéficier de « machines virtuelles » sans nécessiter d'acquérir et gérer des serveurs dédiés, souvent sous-exploités ;
- Le stockage puis l'archivage des données de la recherche ;
- La maintenance évolutive du portefeuille de solutions logicielles qui seront créées pour la gestion des données actives (e.g. la plateforme « Digital Humanities ») ;
- Le conseil aux chercheurs sur les infrastructures et services numériques ainsi que l'animation des communautés de pratiques.

Après consultations auprès du Rectorat et du service des finances de l'UNIGE, les tarifs des services numériques issus du PL qui utilisent des ressources de manière exclusive ont récemment été fixés et communiqués à la communauté universitaire du canton. Il s'agit en particulier des services de stockage et d'archivage des données de la recherche, dont les tarifs annuels sont les suivants :

- stockage disque (backupé) : 75 CHF / TB
- stockage bande (backupé) : 25 CHF / TB
- préservation long terme : 100 CHF / TB (copies sur deux technologies différentes)

Afin de limiter le travail administratif et de se conformer aux pratiques du marché, 50 GB sont mis à disposition gratuitement. Ces tarifs seront révisés annuellement afin de rester compétitifs avec les tarifs du marché.

Figure 5 : Gouvernance du PL



N°DOI : 10.55790/journals/ressi.2019.e2006

Appel à projets

Afin de faire participer les communautés de chercheurs à la mise en place de nouvelles solutions numériques pour la gestion des données, un appel à projets a été lancé en novembre 2019. Celui-ci s'adresse à toutes les disciplines scientifiques.

À cette fin, les chercheurs sont appelés à soumettre des propositions de projets informatiques en vue de faciliter la collecte, le traitement, la gestion, le partage ou la diffusion des données de recherche, activités communément désignées sous le terme de gestion des données actives (« ADM », voir ci-dessus).

Les projets proposés doivent s'inscrire dans ce thème et mener à une solution numérique :

- mutualisable – qui soit utile à plusieurs groupes de recherche, éventuellement de disciplines différentes, et soit suffisamment générique pour être utilisée par le plus grand nombre de chercheurs,
- institutionnalisable – qui aspire à devenir un service numérique institutionnel complétant l'offre disponible à l'UNIGE et autres HE, ou à étendre un service existant,
- novatrice – qui se distingue par son originalité, n'existe pas au sein de la communauté de recherche genevoise ou soit développée selon une approche innovante.

Le large périmètre tant technique (développement, intégration, solutions open source, etc.) que thématique de l'appel à projets a pour volonté de favoriser l'innovation et l'expression de besoins au sein des nombreux domaines de recherche constituant la communauté scientifique genevoise.

6. Synergie avec les projets nationaux

Il est intéressant de relever que l'élaboration du PL en été 2011 coïncide avec la mise en place du programme suisse « accès à l'information scientifique », de la Conférence Universitaire Suisse (CUS) – qui est devenue par la suite swissuniversities (SWU) à partir du 1er janvier 2015 – dont les prémisses des idées et concepts remontent à fin 2010. À cette époque, l'explosion de la numérisation était relevée comme un élément disruptif qui contribuait à placer les universités et la communauté scientifique devant de nouveaux défis en matière d'information scientifique dont il n'était pas encore possible de définir le périmètre de manière définitive. Des mesures urgentes étaient donc jugées nécessaires pour que la communauté scientifique suisse puisse avoir durablement accès aux informations scientifiques dont elle avait besoin tout en limitant les coûts pour l'ensemble du système.

En conséquence, durant le premier semestre 2011, l'élaboration du programme fédéral 2013-2016 a sollicité l'apport de propositions de la communauté universitaire. Dans ce contexte d'appel à contributions, j'ai eu l'opportunité de me positionner sur le sujet du cycle de vie des données dans lequel je précisais qu'il devenait primordial de sensibiliser l'ensemble des chercheurs à la problématique de l'explosion de la quantité de données numériques produites quotidiennement. Il convenait dès la création des données de recherche de mettre à leur disposition des services de « data life cycle management » (DLCM), basés sur les meilleures pratiques professionnelles et adaptés aux différents contextes scientifiques. Il était également relevé que l'accessibilité des données devenait une condition posée par certaines agences gouvernementales de financement de la recherche ainsi que par certains éditeurs.

N°DOI : 10.55790/journals/ressi.2019.e2006

Dans cette perspective, j'avais introduit le concept de « Scientific Object Repositories » (SOR) visant à une gestion plus efficiente des données numériques avec des mécanismes de stockage, de partage (data sharing), d'accès sécurisé, de catalogage, d'annotation, ainsi que d'archivage à long terme, selon des pratiques qui correspondent à des standards internationaux établis. Le concept de SOR devait reposer sur une architecture distribuée permettant de mutualiser et de capitaliser les efforts et les connaissances. Ce concept devait également permettre une valorisation des données au-delà de celle initialement pensée par les chercheurs, au travers d'outils de data mining, data visualization, mashup, etc. dont les bénéficiaires seront d'autres communautés exerçant leurs activités dans la recherche et l'enseignement et plus largement la société civile. Même si aujourd'hui le terme SOR n'a pas survécu, le concept a été en bonne partie repris dans le projet DLCM (https://www.dlcm.ch).

Concernant le HPC, il était aussi relevé que pour répondre de manière maîtrisée et sécurisée aux besoins émergents de la science vis-à-vis de la croissance exponentielle des données la mise en place d'un « cloud académique » au niveau national s'avérait nécessaire. Afin de faciliter les collaborations à l'échelle nationale et internationale, tout en optimisant les investissements consentis à différents niveaux, ce cloud devait être conçu selon le principe du fédéralisme en mettant à disposition des communautés scientifiques, en modes IaaS et SaaS, les services requis pour soutenir l'évolution de la recherche.

Aussi, après une période de gestation qui a duré jusqu'en 2012, le programme CUS P-2 a pris sa forme finale et son titre : « Information scientifique : accès, traitement et sauvegarde », pour démarrer en 2013. C'est à ce moment que j'ai concrétisé le projet DLCM (Burgi et al., 2017) en prenant contact en novembre 2013 avec des experts du domaine DLCM dans les universités et écoles polytechniques 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 partenariat entre les deux écoles polytechniques fédérales (ETH-Z et EPF-L), les universités de Zurich, Bâle, Lausanne, et Genève, la Haute Ecole de Gestion (HEG) de la Haute Ecole Spécialisée de Suisse occidentale (HES-SO), et SWITCH s'est finalement établi, 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. Cette première phase du projet s'est terminée en juillet 2018, pour être prolongée dans une deuxième phase qui doit se terminer en décembre 2020, et qui implique un nouveau partenaire, la Zürcher Hochschule für Angewandte Wissenschaften (ZHAW).

La naissance du projet de loi dans ce contexte d'effervescence d'idées au niveau national n'est par conséquent pas un hasard, et a fortement bénéficié de cette synergie. Pourtant, comme mentionné dans la section 2, l'agenda politique fédéral et celui du canton de Genève ont chacun suivi leur feuille de route, et il aura fallu attendre janvier 2018 pour initier le projet de loi. Quant au programme CUS P-2, il est entre-temps devenu SWU P5 couvrant la période 2017-2020. Cette resynchronisation des deux projets a néanmoins permis au canton de Genève de dynamiser le développement de la technologie DLCM qui conduira dès juin 2019 à mettre en production le système d'archivage long terme Yareta, mis à disposition des HE du canton, soit avant l'instance nationale (dénommée « OLOS »), qui devrait voir le jour durant le premier trimestre 2020.

7. Conclusions

Au-delà de fournir des infrastructures modernes et efficientes aux chercheurs, leur permettant ainsi de pleinement exercer leurs travaux de recherche selon le 4ème paradigme, le projet de loi a potentiellement plusieurs autres impacts. D'une part, la mise en place des trois thématiques, HPC, ADM, et LTP, a permis d'activer des communautés de chercheurs, contribuant ainsi à une mutualisation des pratiques et des outils. D'autre part, du fait que le projet de loi concerne toutes les HE du canton de Genève, des synergies entre des institutions travaillant sur des thématiques similaires sont facilitées. Par exemple, la Haute école de santé avec la Faculté de médecine ; la Haute école de gestion avec la Faculté en sciences économiques ; ou la Haute École du paysage, d'ingénierie et d'architecture avec la Faculté des sciences, etc.

Un autre aspect durable concerne la gestion des données de recherche qui, avec la mise en opération de Yareta, permet de franchir une première étape vers l'Open Science dans laquelle l'accès et le partage des données devrait amener à une plus grande transparence de la recherche, et selon des études, une plus grande citabilité des publications, et la promotion des résultats à une plus grande audience (Popkin, 2019). Ce pas vers l'Open Science est cohérent avec le prochain programme de swissuniversities qui va couvrir la période 2021-2028. Le PL « Infrastructures et services numériques pour la recherche » représente par conséquent une excellente opportunité de préparer les HE genevoises à cette nouvelle orientation que prend la recherche suisse.

8. Remerciements

Je tiens à remercier Alain Jacot-Descombes, directeur de la division des Systèmes d'Information et de Communication (DiSTIC) de l'UNIGE pour ses commentaires constructifs sur une version préliminaire du texte, ainsi que l'équipe du PL12146 et du pôle eResearch de la DiSTIC pour leurs apports (figures, informations, etc.) qui ont contribué à enrichir cet article.

NOTES

(1) Hautes écoles (universitaires et spécialisées)

(2) Un cœur (physique) est un ensemble de circuits capables d'exécuter des programmes de façon autonome.

(3) Charles Beer, conseiller d'Etat genevois de 2003 à 2013, en charge du DIP, avait demandé d'inclure dans le périmètre du PL les Hautes Ecoles Spécialisées (HES) du canton de Genève afin de couvrir également leurs besoins en calcul et gestion des données.

(4) Le rapport de la commission, http://ge.ch/grandconseil/data/texte/PL12146A.pdf est passé devant la parlement le 24 novembre 2017, cf. http://ge.ch/grandconseil/sessions/seances-odj/67/?session=48.

(5) Ce point est important du fait qu'un projet de loi concerne uniquement des investissements, et que des coûts de licences (locations) ne peuvent pas être pris en charge dans ce contexte.

(6) Les GPU (Graphics Processing Unit) permettent d'effectuer des calculs plus rapidement dans les domaines impliquant des algorithmes fortement parallélisables.

(7) Un Peta Byte (PB) correspond à 1'000 Tera Bytes (TB).

(8) Le protocole S3 (Simple Storage Service) a été développé par Amazon. Il consiste à stocker l'information sous la forme d'objets, et non plus selon une organisation en fichiers.

(9) Les principes FAIR ont été rédigés en 2015 lors d'un atelier du Centre Lorentz à Leyde, aux Pays-Bas.

(10) API (Application Programming Interface) est un ensemble normalisé de fonctions, méthodes, etc. qui sert de façade par laquelle un logiciel offre des services à d'autres logiciels.

(11) RDF, développé par le W3C, est le langage de base du Web sémantique destiné à décrire de façon formelle les ressources Web et leurs métadonnées, permettant le traitement automatique de telles descriptions.

(12) Le World Wide Web Consortium, abrégé par le sigle W3C, est un organisme de standardisation à but non lucratif, fondé en octobre 1994 chargé de promouvoir la compatibilité des technologies du World Wide Web.

BIBLIOGRAPHIE

Al-Baik, O. & Miller, J. (2015) The kanban approach, between agility and leanness: a systematic review, Empirical Software Engineering, 20(6), 1861-1897. doi :10.1007/s10664-014-9340-x

Borgman, C.L. (2009). The digital future is now: A call to action for the humanities. Digital Humanities Quarterly, 3(4). Retrieved from http://digitalhumanities.org/dhq/vol/3/4/000077/000077.html

Burgi, P.-Y., Blumer, E., Makhlouf-Shabou, B. (2017). Research Data Management in Switzerland: National Efforts to Guarantee the Sustainability of Research Outputs. IFLA Journal [online], 43(1), 5–21. doi:10.1177/0340035216678238

Hey, T., Tansley, S., & Tolle, K. (Eds.) (2010). The Fourth Paradigm: Data-Intensive Scientific Discovery, Microsoft Research

Moore, G.E. (1965). Cramming more components onto integrated circuits. Electronics, 38(8), 114–117.

Popkin, G. (2019) Data sharing and how it can benefit your scientific career, Nature 569, 445-447. doi: 10.1038/d41586-019-01506-x

Waldrop, M.M. (2016) The chips are down for Moore's law, Nature 530, 144-147.

Wilkinson, M.D. et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship, Scientific Data volume 3, Article number: 160018