Logiciels libres : à la recherche du bien commun

Edlira Nano
Associations April,
La Quadrature Du Net,
informaticienne indépendante
eda@mutu.net / eda@laquadrature.net

Olivier Langella
Ingénieur au CNRS
Plateforme PAPPSO
Laboratoire GQE-Le Moulon
Ferme du Moulon
91190 Gif-sur-Yvette
France
olivier.langella@universite-paris-saclay.fr

Filippo Rusconi
Chercheur au CNRS
Plateforme PAPPSO
Laboratoire GQE-Le Moulon
Ferme du Moulon
91190 Gif-sur-Yvette
France
filippo.rusconi@universite-paris-saclay.fr

Résumé

Le logiciel libre et la recherche publique partagent un objectif : le bien commun, au service de tous. Cette présentation revient sur ce que sont le bien commun et la science ouverte pour essayer de les analyser à l'aide de l'exemple de la plateforme scientifique Analyses protéomiques de Paris Sud-Ouest (PAPPSO).

PAPPSO s'est dotée d'une infrastructure informatique complète basée exclusivement sur du Logiciel libre : réseau, serveurs, stockage, calcul et postes personnels. Elle développe plusieurs logiciels sous licence libre, dont ceux qui forment la chaîne de traitement des données de spectrométrie de masse. Ce choix naturel facilite la reproductibilité des traitements, apporte la maîtrise des logiciels et permet l'intégration de code source tiers.

L'apport majeur du Logiciel libre à la recherche publique permet l'utilisation des réseaux et systèmes informatiques. Scientifiquement, l'ouverture du code source et la liberté d'utilisation des logiciels garantissent l'échange des données, leur réutilisation et leur vérification par les pairs. Cet ensemble forme un bien commun protégé par des licences. De nombreux laboratoires y contribuent en utilisant ou en produisant des logiciels libres, comme en témoigne en partie la forge du code source du secteur public (https://code.gouv.fr).

Comment contribuer efficacement ? Quelles sont les recommandations et obligations pour les établissements publics ? Quelle licence choisir ? Comment une licence copyleft peut-elle aussi séduire les partenaires privés ? Nous apporterons des réponses et des éléments de réflexion pour corriger quelques fausses croyances et promouvoir la construction collective d'une culture libre, au service du bien commun.

Mots-clefs

logiciel libre, logiciel scientifique, biens communs, recherche scientifique, science ouverte, licences, open source, copyleft, protéomique

1 Une plateforme pour la recherche sous Logiciel libre

1.1 Contexte scientifique

La protéomique est l'étude de l'ensemble des protéines produites par une cellule, un organe ou un organisme . Elle permet d'analyser sans a priori les changements qualitatifs ou quantitatifs dans la composition en protéines en fonction de traitements, du développement ou de variations génétiques, et de comprendre par exemple comment un organisme répond à un stress ou à une maladie.

L'apparition du terme « protéomique » en 1995 coïncide avec le début de l'utilisation de la spectrométrie de masse pour l'identification des protéines puis pour leur quantification. Aujourd'hui la méthode utilisée quasi-exclusivement est l'analyse de peptides issus de la digestion tryptique des protéines, par un spectromètre de masse couplé à une chaîne de chromatographie liquide. Les appareils n'ont pas cessé d'évoluer mais au final il s'agit toujours de mesurer la masse moléculaire des peptides ayant pénétré dans le spectromètre et/ou de leurs fragments et de quantifier leur intensité.

Les analyses par spectrométrie de masse produisent de grandes quantités de données, directement en sortie de l'instrument. Chaque fabricant a son propre format de données, qui est le plus souvent propriétaire. Ordinairement, les données aux formats propriétaires sont traitées par les logiciels fournis avec l'instrument, eux aussi propriétaires. Comme souvent dans le domaine des formats de fichiers, ces formats ne sont pas pérennes et les licences d'utilisation des logiciels propriétaires sont très coûteuses. Il y a donc ici un problème majeur d'interopérabilité.

Depuis 2005, la plateforme scientifique "Analyses protéomiques de Paris Sud-Ouest" (PAPPSO) a fait le choix du Logiciel libre pour garantir la pérennité de ses chaînes de traitement, la reproductibilité des expériences et la capitalisation de son savoir faire.

1.2 Transition des logiciels propriétaires au Logiciel libre

Le passage progressif au Logiciel libre a permis une rationalisation de l'utilisation des ressources informatiques. Nous sommes passés de postes dédiés à licence unique pour usage unique à une infrastructure collective combinant stockage, calcul et enchaînement des traitements depuis n'importe quel poste de travail.

Durant la période 2005–2015, notre travail a été facilité par la définition de formats standards en protéomique, l'émergence de nombreux logiciels libres dans le domaine [1] et le développement de

nouvelles solutions logicielles sous licence libre (PROTICdb [2], mineXpert2 [3], X! TandemPipeline [4], MassChroQ [5])[1].

1.3 Choix du système d'exploitation

Le choix de PAPPSO s'est porté d'abord sur la distribution GNU/Linux Ubuntu, puis sur Debian. Un groupe de développeurs officiels Debian, dont un auteur de ce rapport, s'attache à fournir dans la distribution de nombreux logiciels pour la chimie, et en particulier pour la spectrométrie de masse ("team debichem"). L'intérêt principal de la distribution Debian est la richesse de son offre logicielle qui permet de disposer d'un socle de fonctionnalités robuste couvrant les exigences "serveur" et les impératifs "bureautique". Tous les logiciels développés par PAPPSO sont disponibles sous forme de paquets Debian dans des dépôts publics. Le déploiement des logiciels, dépendances comprises, sur les serveurs et les postes de travail de l'équipe est ainsi simplifié et totalement automatisé.

1.4 Stockage des données

Les besoins en stockage de la plateforme évoluent constamment en fonction des progrès techniques des spectromètres de masse. Chaque nouvelle génération d'instruments apporte des améliorations, en particulier sur la précision de mesure de masse, qui provoquent une augmentation significative du volume des données générées. Cette progression des besoins en volume de stockage sur les quinze dernières années est illustrée dans la figure ci-dessous.

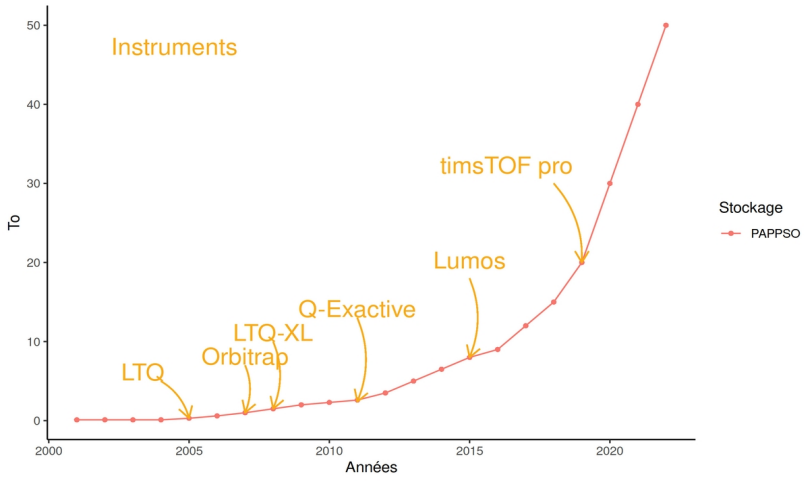

Figure 1 – *Evolution des besoins de stockage de la plateforme en To/an, comparée à l'évolution des disques durs*

Le système de stockage des données de spectrométrie de masse doit permettre une adaptation en continu de la volumétrie d'espace disque disponible ainsi que les meilleures performances en lecture et écriture. Les solutions classiques de type NAS ont été écartées pour éviter la dépendance matérielle et les problèmes liés au renouvellement des équipements.

1 http://pappso.inrae.fr/

Dès 2011, nous avons été parmi les premiers à faire confiance à une solution nouvelle de stockage distribué: Ceph[2]. La principale caractéristique de ce système de stockage est de ne requérir que des serveurs standard. La flexibilité et l'adaptabilité de ce système à des besoins perpétuellement en évolution en ont fait la solution la plus robuste que nous connaissions.

1.5 Calcul scientifique

La plateforme PAPPSO est spécialisée dans les traitements en protéomique haut débit (nombreux échantillons à traiter dans les plus brefs délais). Pour assurer la disponibilité de nos moyens de calcul à l'ensemble des utilisateurs, nous utilisons le gestionnaire de processus HTCondor[3].

Les besoins en calcul évoluent eux aussi en fonction des instruments utilisés. Avec l'évolution des techniques, de nouvelles possibilités sont apparues dans le traitement de données en protéomique, exigeant elles aussi des capacités de calcul supplémentaires. De la même manière que pour les capacités de stockage, les machines dédiées au calcul doivent être ainsi renouvelées régulièrement et intégrées au fur et à mesure.

2 Retour d'expérience sur 10 ans

2.1 Matériel

L'intégration de nouvelles machines de calcul ou de stockage s'est faite de manière transparente. Nous sommes passés d'une capacité de stockage initiale de 18To (3 serveurs R515, disques de 3To) en 2011 à une capacité de 917To (8 serveurs hétérogènes). Le réseau est passé du 1Gb cuivre au 10Gb SFP+. Il n'y a pas eu de transfert de données/migration, pas de modification de l'architecture logique pour les utilisateurs. Le système de fichiers cephfs permet un accès direct aux données depuis chaque nœud de calcul. Globalement, les performances ont suivi les évolutions matérielles (augmentation du débit, augmentation des capacités de calcul). La résistance aux pannes a été mise à rude épreuve (panne électrique, disques ou erreurs humaines) et nous n'avons jamais eu de perte de données.

2.2 Logiciel

Les systèmes pour les serveurs et pour les postes utilisateur ont été migrés en 2013 de Ubuntu GNU/Linux vers Debian. Nous y avons gagné en stabilité et en simplicité lors des mises à jour de version. La stratégie consiste à maintenir le parc informatique sous Debian "stable" et effectuer le passage à la version successive dans les mois qui suivent sa publication officielle. Le stockage centralisé est disponible pour tous les postes dans une arborescence commune, via un montage automatique sur les nœud de calculs (cephfs via systemd sur les serveurs, sshfs sur les postes clients). Les logiciels sont les mêmes sur les serveurs et les postes utilisateurs. L'accès distant au cluster de calcul se fait avec x2go, via une clé publique ssh.

2.3 Scientifique

Les analyses de la plateforme ont évolué pour passer de la technique des gels d'électrophorèse 2D vers des analyses uniquement basées sur la spectrométrie de masse. Le traitement des images des

2 https://ceph.io/

3 https://htcondor.org/

gels 2D était majoritairement effectué avec des logiciels propriétaires sous Windows, sur des postes dédiés, ce qui limitait les capacités de traitement. Le passage progressif à des processus analytiques qui faisaient l'économie de l'étape d'électrophorèse a coincidé avec les apports massif du Logiciel libre dans le domaine scientifique, au milieu des années 2000. Nous avons alors pu commencer la transition vers certains logiciels libres qui émergeaient à cette époque. Cependant, ces logiciels étaient principalement des librairies encore imparfaitement dotées des fonctionnalités requises dans notre domaine. Nous avons alors entrepris le développement de nos logiciels sur la base des besoins scientifiques particuliers à notre plateforme. Le logiciel MassChroQ est né ainsi, de nos besoins en protéomique quantitative. Notre indépendance vis-à-vis des formats de données propriétaires des fabricants nous a permis de produire un logiciel évolutif et pérenne dès le départ, évitant l'effet « boîte noire ». Ainsi, notre offre logicielle a pu être adaptée au fur et à mesure aux nouvelles techniques, à des instruments significativement différents de génération en génération, absorbant ainsi les « chocs » technologiques : doublement des fréquences d'acquisition à chaque génération (3 ans), doublement du pouvoir résolutif (précision des mesures des masse).

La dernière « rupture technologique » a été l'apparition du timsTOF Pro du fabricant Bruker. Cet appareil dispose d'une fréquence d'acquisition 10 fois supérieure à celle de la génération précédente ainsi que d'une nouvelle dimension de séparation des peptides (mobilité ionique). Fait unique depuis les débuts de la protéomique, Bruker nous a confié les spécifications techniques de son format de fichier de données. Cela nous a permis d'adapter MassChroQ pour pouvoir utiliser de manière native les données obtenues sur cet instrument. Notre savoir faire en développment C++ nous a ainsi permis d'obtenir des gains de performances remarquables par rapport aux logiciels commerciaux, ainsi que de meilleurs résultats scientifiques.

2.4 Bilan

Le passage au Logiciel libre pour tous les besoins informatiques de la plateforme PAPPSO a permis une maîtrise totale de ses outils, depuis la production des données brutes jusqu'à l'interprétation biologique. Les sommes importantes économisées en licences de logiciels propriétaires (20k€ par an) ont été investies dans la maintenance des ressources de calcul et de stockage. Le savoir faire développé par PAPPSO dans l'analyse protéomique à haut débit est reconnu au niveau international (129 articles citant MassChroQ depuis 2011, publication d'un article de référence en métaprotéomique [6]). Toutes les analyses sont complètement vérifiables et reproductibles, les logiciels étant tous librement téléchargeables, sous licence GPLv3+, sans demande préalable.

3 Importance et apports du libre dans la recherche scientifique

3.1 Affinités entre culture libre, communs et recherche publique

La recherche publique vise à produire et à développer des connaissances, les améliorer sans cesse, puis les faire connaître et les transmettre au public, à travers les générations, constituant ainsi un savoir et une connaissance qui caractérisent notre civilisation. Dans ce travail incrémental, il est crucial de pouvoir partir de l'existant, de ne pas réinventer sans cesse. Pour ce faire, le partage de la connaissance, la possibilité de réutiliser le travail déjà effectué, de pouvoir modifier et adapter ceux-ci, les améliorer, àfin de les préciser ou de les rejeter. Et de pouvoir ensuite à son tour partager et diffuser ces nouveaux résultats, qui viendront s'ajouter à la connaissance existante et ainsi de suite.

Vous avez là la définition d'un logiciel libre ou plus largement d'un bien libre. La culture du libre n'est pas unique aux logiciels. Elle peut concerner un logiciel, un service informatique en ligne, un

document (article scientifique, livre, poème, rapport) un produit artistique (oeuvre numérique, production graphique, reproduction numérique d'une œuvre physique, par exemple les tableaux d'un musée, les photos de monuments etc), une semence agricole (les graines libres de droit), un médicament libre, etc. Une œuvre ou bien libre est quelque chose qu'on peut utiliser librement, partager librement, distribuer librement autour de nous, mais aussi en modifier le contenu, l'adapter, l'améliorer, pour pouvoir ensuite librement le repartager, et le redistribuer etc.

Cette culture du libre est inhérente à la recherche publique également, et plus largement à ce qu'on appelle parfois les communs. Les communs sont des ressources (naturelles ou culturelles) partagées et gérées collectivement, accessibles et disponibles pour tous, et qui n'appartiennent pas à qui que ce soit au sens de la propriété, qu'elle soit publique ou privée (Elinor Ostrom, prix Nobel en économie pour ses travaux sur la théories communs, parle par exemple de « ressources de propriété commune » qu'elle définit et explique dans [7]).

3.2 Recommandations institutionnelles et obligations légales en terme de publications de code source

La recherche publique en France est financée (sources 2018) au moins à 75 % par des fonds publics français (MIRES, hors MIRES, Administrations), et parmi les 25 % restants on ne retrouve que 5 % de ressources contractuelles provenant d'entreprises, les autres 20 % sont un mélange de fonds publics étrangers, notamment européens, ressources propres tels que les prestations de services des structures de recherche elle même, et enfin des ressources privées d'entreprises étrangères) [8].

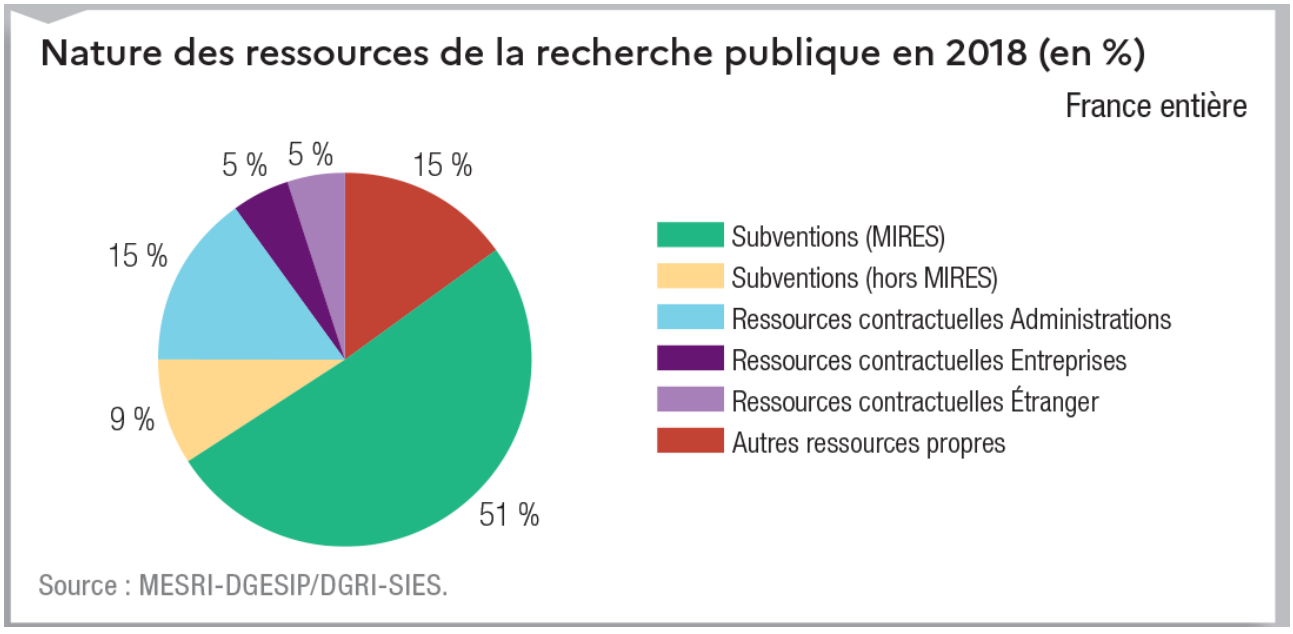

***Figure 2** - Nature des ressources de la recherche publique en 2018 (en %)*

…. Version non terminée ...

Bibliographie

[1] Rusconi F. Free Open Source Software for Protein and Peptide Mass Spectrometry- based Science. Curr Protein Pept Sci, 2 (22) 134-147, 2021 ; https://doi.org/10.2174/1389203722666210118160946

[2] Langella O. , Valot B., Jacob D., Balliau T., Flores R., Hoogland C., Joets J., Zivy M.. (2013) Management and dissemination of MS proteomic data with PROTICdb: Example of a quantitative comparison between methods of protein extraction. Proteomics, 9 (13) 1457-66

[3] Langella O, Rusconi F. mineXpert2: Full-Depth Visualization and Exploration of MSn Mass Spectrometry Data. J. Am. Soc. Mass Spectrom., 4 (32) 1138-114, mars 2021 ; https://doi.org/10.1021/jasms.0c00402

[4] Langella O, Valot B, Balliau T, Blein-Nicolas M, Bonhomme L, Zivy M. X!TandemPipeline: A Tool to Manage Sequence Redundancy for Protein Inference and Phosphosite Identification. J. Proteome Res., 2 (16) 494-503, décembre 2016 ; https://doi.org/10.1021/acs.jproteome.6b00632

[5] Valot B, Langella O, Nano E, Zivy M. MassChroQ: a versatile tool for mass spectrometry quantification. Proteomics, 17 (11) 3572-3577, juin 2011 ; https://doi.org/10.1002/pmic.201100120

[6] Van Den Bossche T, Kunath BJ, Schallert K, Schäpe SS, Abraham PE, Armengaud J, Arntzen M, Bassignani A, Benndorf D, Fuchs S, Giannone RJ, Griffin TJ, Hagen LH, Halder R, Henry C, Hettich RL, Heyer R, Jagtap P, Jehmlich N, Jensen M, Juste C, Kleiner M, Langella O, Lehmann T, Leith E, May P, Mesuere B, Miotello G, Peters SL, Pible O, Queiros PT, Reichl U, Renard BY, Schiebenhoefer H, Sczyrba A, Tanca A, Trappe K, Trezzi JP, Uzzau S, Verschaffelt P, von Bergen M, Wilmes P, Wolf M, Martens L, Muth T. Critical Assessment of MetaProteome Investigation (CAMPI): a multi-laboratory comparison of established workflows. Nature Communiations, 1 (12) 7305, décembre 2021 ; https://doi.org/10.1038/s41467-021-27542-8

[7] Elinor Ostrom, Governing the Commons : the evolution of institutions for collective actions, 1990, Cambridge University Press ; https://archive.org/details/ElinorOstromGoverningTheCommons

[8] MESRI, L’état de l'Enseignement supérieur, de la Recherche et de l'Innovation en France n°14, 2021 ; https://publication.enseignementsup-recherche.gouv.fr/eesr/FR/T622/le_financement_des_activites_de_recherche_et_developpement_de_la_recherche_publique/