Atelier Dialogu'IST du 3 décembre animé
par Fabien Borget, enseignant-chercheur à Aix-Marseille Université

Introduction de l'atelier

La dérive actuelle de la science se caractérise par une augmentation quasi exponentielle des retraits d'articles pour fraude. Quelques chiffres : en 1980, on comptait 21 rétractations d'articles en biologie, entre 2001 et 2010, ce chiffre a été multiplié par 11. (1)
On peut lancer un débat pour savoir si ce n'est pas le contexte qui pousse les chercheurs aux limites de l'éthique, voire parfois à dépasser ces limites. Ceci amenant certains à interpréter le « publish or perish » de manière déraisonnable : prenons l'exemple du champion du monde des publications : Alan Katritzky, chimiste anglais décédé qui officiait en fin de carrière à l'Université de Floride et qui a cosigné entre 1953 et 2010, 2215 articles soit une publication tous les 10 jours! (1)
Cet adage « publish or perish » a également été bien compris par les éditeurs privés qui ont verrouillé le système du lecteur payeur qui se trouve être aussi le producteur de la matière qui les enrichit.
Du coup la science a perdu de son sens philosophique. Le chercheur raisonne en PPQP (Plus Petite Quantité Publiable) (1), ce qui l'amène à toujours trouver l'hypothèse qui correspond à ses résultats et non plus le contraire.
On en arrive donc à un cercle vicieux où la Recherche produit de plus en plus de données que plus personne ne sait reproduire et parfois ne peut reproduire et qui ne fait qu'entretenir un système financier qui enrichit le secteur le plus rentable de l'édition.

Ce constat a été fait il y a déjà quelque temps et le financeur public de la recherche s'est même étonné que le fruit de son financement ne soit pas accessible à tout citoyen. Des réactions (ou plutôt des réflexions) sont arrivées. Il y a déjà 20 ans on commençait à parler de publications ouvertes et accessibles.
Tout doucement la communauté scientifique a réagi et a créé des nouvelles choses plus éthiques. Des archives ouvertes sont arrivées, citons une des plus célèbres arXiv hébergé à l'origine au Los Alamos National Laboratory pour ensuite être hébergé par l'Université Cornell (2). Certaines communautés scientifiques étant plus sensibilisées que d'autres pour le partage des informations, car intrinsèque à leur activité. Les véritables pionniers de cette démarche ont été les utilisateurs de Grands Instruments comme au CERN ou bien la communauté des astrophysiciens analysant des données de satellites ou de grands observatoires. Les premiers modèles de partages de données et d'accès libre de publications étaient nés.
Mais l'arsenal législatif n'était pas encore prêt.
Pourtant l'organisation progressive continue d'avoir lieu et, bien heureusement, de ces réflexions naissent des appels pour un changement de système qui se veut une véritable REVOLUTION.
Mais, en France, tout ne devient possible qu'en 2016 avec la loi pour une République Numérique (3) qui place le cadre législatif et c'est là que peut réellement débuter cette REVOLUTION.
Le CoSO (Comité pour la Science Ouverte) (4) suit juste après et une réelle politique peut enfin se mettre en place permettant le véritable développement de la science ouverte. Ce concept qui

permet de redonner du sens à la Recherche, mais qui bouscule, transforme, modifie, bouleverse tous les usages, de la gestion des projets à l'évaluation de la recherche puis des chercheurs.
Ces derniers temps la politique s'accélère, citons la feuille de route du CNRS annoncé aux journées de la science ouverte (18/11/2019), il y a une dizaine de jours présentée par Antoine Petit, président général du CNRS, pour fin 2020 (5). Actuellement au niveau du CNRS 51 % des publications sont en accès fermé, 49% en accès ouvert (23,5% chez éditeurs, 24,5% en archive, 1% inconnu) et l'objectif fixé par cette feuille de route est :

- 100% publications en OA
- Des données de la recherche FAIR-isées
- Promouvoir l'utilisation d'outils pour l'analyse, pour la fouille de textes et de données
- L'évaluation individuelle des chercheurs

Au niveau des universités, c'est la même chose, elles commencent à s'organiser. Des services naissent, des infrastructures évoluent, des nouveaux métiers apparaissent.
Notre Université (Aix-Marseille Université) par exemple est en train de se doter d'une « Maison de la Donnée ». Il est donc bon de constater que dès que le cadre législatif a existé, une réelle politique dynamique a pu s'amorcer. Au niveau national, on peut également citer, par exemple, que maintenant l'ANR impose un plan de management des données pour tout projet déposé. L'Europe le fait depuis longtemps.

Les idées de base permettant la mise en œuvre de cette nouvelle politique sont :
Rendre accessibles les produits de la recherche, mais pas seulement les publications, les données également. Et selon des principes précis que l'on résume par FAIR (Findable, Accessible, Interoperable, Reusable) : Trouvable, Accessible, Interopérable et Réutilisable. Et justement le sujet de notre atelier va être de discuter de ce dernier aspect en particulier. Parce que là on touche un point important de la recherche qui est la reproductibilité des résultats et des expériences pour les sciences expérimentales.
Là se trouve à mon sens, le moyen de réellement réussir cette révolution.

Mais de nombreuses questions suivent.
Quelles données rendre disponibles ? Toutes ? Est-ce que cela doit dépendre du projet ?, combien de temps ? Qu'en est-il des principes RGPD, est-ce une limite ? Certaines de ces questions ont d'ailleurs déjà été abordées lors de précédents ateliers.
Pour répondre à ces questions, des structures s'organisent peu à peu sous forme de services, de cellules, de commissions mettant en relation les chercheurs impliqués, les professionnels de l'IST, les informaticiens, mais aussi les juristes. Des projets naissent, se réalisent… On sent un vent frémir et un bouillonnement annonçant cette fameuse REVOLUTION.

On peut ajouter des questions pragmatiques, quelles infrastructures doit-on créer pour le stockage ? Quel stockage inventer pour permettre une réutilisation ? Quel volume cela représente-t-il ? Comment rendre pérenne des données numériques qui par définition reposent sur des formats « volatils » ?

Les enjeux autour des données sont donc fondamentaux, à mes yeux ; ceux qui gravitent autour de l'intégrité scientifique, de contre-pouvoir face aux éditeurs surpuissants et à une accessibilité des résultats de la recherche à tous sont les plus importants. Mais n'oublions pas que les usages évoluent et d'autres enjeux autour de la carrière des chercheurs, de leurs évaluations seront également intimement liés à cette évolution.
Mais le plus passionnant dans cette évolution, c'est que beaucoup reste à construire, nous verrons et voyons déjà apparaitre des nouveaux métiers, l'évolution de certains autres et au final une recherche qui se retrouvera à créer de la connaissance pour la diffuser, la faire vivre et pourquoi pas la transformer en innovation…

Notre atelier d'aujourd'hui va nous permettre de remettre tout ceci en contexte à travers le prisme de la réutilisation.
En effet qui dit Réutilisation veut implicitement dire Répétable. C'est le point de départ d'une recherche « Reproductible ».
Le fait de rendre ses données de recherches disponibles les rend réutilisables *a priori*, mais pour en faire quelques choses de « nouveau » cela implique qu'elles soient JUSTES (ou FAIR) en anglais.
Et donc la reproductibilité est le prérequis à toute possible réutilisation dans un contexte différent, permettant l'obtention de nouveaux résultats à partir de ce premier jeu de données.

Cette après-midi, nous allons donc dans une première partie discuter du contexte de la reproductibilité en science, en essayant dans un premier temps de définir cette notion. Nous enchainerons alors vers un premier retour d'expériences qui reviendra sur l'aspect technique mais aussi sur son aspect psychologique.

Nous ferons alors une courte pause pour ensuite enchainer sur une deuxième partie traitant des compétences nécessaires et des apprentissages afférents. Pour cela nous nous appuierons sur 3 retours d'expériences, un premier sur un projet international sur la biodiversité, un deuxième dans l'enseignement en se basant sur les apprentissages délivrés dans le Master Gestion des données scientifiques et le dernier sur un MOOC qui revient sur les principes méthodologiques pour une science transparente.

Conclusion de l'atelier

En introduction à cet atelier, Sabrina Granger est revenue sur ce que veut dire le R de FAIR. Il existe un vrai problème de méthodologie qui se traduit, dans le contexte actuel, par de la malscience. C'est-à-dire qu'on se retrouve à la limite de l'interprétation possible, voire à la non répétabilité. Ce qui nous emmène à la survalorisation des données « positives », publiables.
Daniel Jacob nous a prévenu sur le fait qu'il fallait faire attention à la novlangue, derrière la Science Ouverte, il faut mettre de la consistance. C'est le message déjà délivré à travers les besoins méthodologiques mentionnés plus haut.
Du coup, il faut faire attention à ce que veut dire FAIR pour que cela ne soit pas seulement du marketing ponctuel mais une réelle avancée conceptuelle.

De plus, il ne faut pas oublier dans les principes FAIR que la notion d'ouverture doit être considérée comme : « OUVERT autant que possible, fermé autant que nécessaire ». L'agenda du chercheur doit être le moteur de cette démarche, et il faut éviter toute injonction pour éviter que cela ne tombe dans le pur marketing. Le cycle de la donnée doit être présenté assez tôt, dans les cursus universitaires, pour que cette sensibilisation soit adossée à des pratiques méthodologiques adaptées. Ce sont des problématiques d'enseignement comme présentés par Véronique Ginouves qui a bien insisté sur l'importance de la sensibilisation numérique. Mais qui dit formation, implique le développement de nouvelles compétences qui ne peuvent pas être figées car la thématique même est en train d'évoluer, que ce soit au niveau technique, technicité mais aussi finalement en projection dans le futur en relation directement avec le I de FAIR.
L'ensemble des concepts liés aux lettres « FAIR », est déjà utilisé comme principe de fonctionnement méthodologique et est déjà en déploiement. On a pu le voir avec la présentation de Anne-Sophie Archambeau et le travail énorme qui a été réalisé dans le cadre du GBIF (Global Biodiversity Information Facility). Le GBIF représente une masse de données très importante et en accès totalement FAIR. On remarquera que la communauté l'utilisant et le développant a mis en place des formations autour de la Mobilisation de données et de la Réutilisation de données. Ces formations permettent le développement de nouvelles compétences.

Ce sont l'ensemble des nouvelles compétences à décolérer des outils à utiliser qui nous montrent qu'il est possible de faire quelque chose des données. Mais si on se focalise sur le R, il est important d'inspecter les méthodes utilisées, d'être capable de développer les outils permettant de garantir cela pour un jeu de données. Ceci nous a été présenté par Laurence Fahri et Marie Hélène Comte dans le cadre d'un MOOC développé à cet effet, et disponible sur la plateforme FUN.
Mais finalement, la numérisation du labbook (il existe même des possibilités techniques dont nous avons parlé telles que JUPYTER NOTEBOOK) est sans doute une piste fondamentale à explorer. En effet, ce cahier de labo, qui est la première donnée de la Recherche, est une des pierres angulaires du « R » de « FAIR ». De la qualité de ce « cahier de labo » (numérique ou non) dépendra le R réel des résultats obtenus à travers des jeux de données produites par les chercheurs.

(1) MALSCIENCE, de la fraude dans les labos, N. Chevassus-au-Louis, 2016, Ed. Le Seuil
(2) https://arxiv.org, consulté le 13 mars 2020
(3) LOI n° 2016-1321 du 7 octobre 2016 pour une République numérique
(4) https://www.ouvrirlascience.fr/presentation-du-comite/ consulté le 20 mars 2020
(5) https://webcast.in2p3.fr/container/journees-nationales-de-la-science-ouverte-fr, consulté le 20 mars 2020