Préface

De la mesure à la connaissance, l'importance de la donnée et de son cycle de vie…

Michaël Chelle[1], Odile Hologne[2], Gilles Aumont[3]
Délégué(e) auprès de la Directrice Générale Déléguée Science
à la transition numérique[1], à l'IST[2], aux infrastructures scientifiques collectives[3]
INRA-Codir, 147, rue de l'Université, Paris

Les sciences agronomiques au sens large font grand usage de données de toutes sortes en raison de la complexité des agro écosystèmes étudiés et des transitions qu'ils rencontrent. Comme pour les autres domaines, ces données sont essentielles pour quantifier, qualifier, comparer, représenter, prédire… Elles sont de différentes origines (observation, expérimentation, enquête, simulation, etc.), sous des formes très variées (données chiffrées, texte, son, image, questionnaire d'enquête, logiciel…). Les données, souvent issus de processus d'acquisition spécialisés, fournissent après des étapes d'analyses et de contextualisation, des informations qui alimentent nos différentes communautés scientifiques, mais aussi des acteurs de la société civile. Traiter des données dans notre domaine nécessite d'aborder des processus d'acquisition, de transformation, de curation, de partage, d'analyse, de stockage, d'archivage, sans oublier une étape essentielle mais difficile et souvent oubliée, de destruction. Ces processus constituent le cycle de vie (Fig. 1), qui est de plus en plus décrit formellement dans un plan de gestion de données. Les articles de ce numéro spécial aborderont ces différents processus qui suivent l'acquisition.

Dans ses domaines thématiques, l'Inra produit depuis toujours des données caractérisant l'état chimique, biologique, physique, voire sociologique, de ses objets d'étude. Les progrès en mesure (automatisation, drone, imagerie, etc.) font que le volume de ces données s'accroît de façon conséquente. Par ailleurs, les progrès en biologie moléculaire ont donné naissance aux "omics" (génomique, transcriptomique, métabolomique…), qui produisent un volume croissant de données du fait d'une « industrialisation » volontariste de la mesure. Plus récemment, la transformation numérique de notre société est à l'origine de nouvelles sources de données, comme les objets connectés, le *crowd-sourcing* avec le développement des sciences participatives, l'accès illimité à des données partagées, associées à des méthodologies et outils nouveaux comme la fouille de textes et de données que permettent les progrès en intelligence artificielle, sans oublier la simulation. L'émergence de ces volumes de données, de type et origine variés nécessite de nouvelles pratiques pour les manipuler, les transformer par le calcul, les stocker, les partager. Ces pratiques peuvent s'appuyer sur les nouvelles technologies des données massives (*big data*).
Le volume et la diversité des données disponibles pour le chercheur proviennent également d'un essor de la réutilisation de données produites par d'autres avec le développement de la politique d'ouverture des données de la recherche. Corollaire de l'ouverture de la science, la recherche se veut plus reproductible, et impose une plus grande traçabilité des données dans leur cycle de vie. Cette science ouverte et reproductible suppose que les données partagées soient trouvables, accessibles, interopérables et réutilisables, le fameux concept « F.A.I.R ».[1]

[1] Findable Accessible Interoperable Re-usable

Ces évolutions se sont accompagnées de la mise en place de nouvelles organisations afin d'aider le chercheur à mettre en œuvre ces pratiques innovantes grâce à une offre renouvelée de services et de supports. Cela a conduit au développement d'e-infrastructures de recherche. Une e-infrastructure (cyber-infrastructure aux Etats-Unis) est une organisation aux missions similaires aux infrastructures de recherche (service, développement, formation-diffusion) qui permet de combiner des technologies numériques (matériel et logiciel), des ressources (données, services, bibliothèques numériques, qualification des informations), des dispositifs de communication (protocoles, droits d'accès et réseaux, sécurité), des compétences, et une gouvernance permettant une gestion rassemblée, collective, ouverte et transparente des données. Ces e-infrastructures s'appuient sur des infrastructures de recherche dont les technologies sont de plus en plus « digitales » et produisent de grandes quantités de données, ou sur des données produites « ailleurs » et mises en ligne selon des modalités diverses. Ces organisations sont au cœur de l'Agenda numérique de la Commission européenne pour soutenir la vision « Science ouverte » et le lien entre chercheurs, citoyens ou entreprises privées. Citons notamment la construction de *l'European Open Science Cloud* (EOSC). Le développement des e-infrastructures est variable selon les communautés ; ces e-infrastructures, centrales en sciences physiques, en bioinformatique, majeures en écologie, sont encore en construction dans d'autres communautés.

Ces changements autour de la donnée questionnent également d'un point de vue épistémologique les relations données-modèles, mais aussi plus profondément le changement de paradigme entre approche hypothético-déductive et approche descriptive intégrale et massive, souvent nommée « *data-driven science* ».

L'émergence de la « data science » et des capacités technologiques autour des données bouleverse les rapports de la science à la société et les bases éthiques de pans entiers de recherche, et renouvèle les questions d'intégrité et de reproductibilité, voire de déontologie du chercheur, problématiques qui ne sont pas abordées ici. Ce numéro spécial a été d'abord conçu pour vous apporter un éclairage sur les outils sous-jacents à ces évolutions : contexte science ouverte, gestion et stockage, traitement et analyse, en combinant synthèse et exemples concrets.

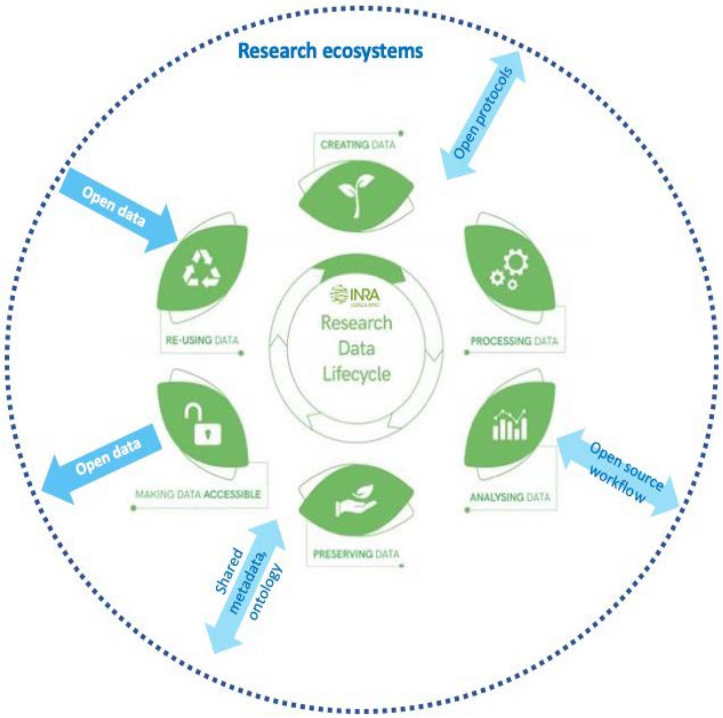

Figure 1 : Une vision des différents processus du cycle de la donnée, ouverts sur l'écosystème de recherche (*d'après datatree.uk.org*)