CONSEIL SCIENTIFIQUE

INRA

Rapport du groupe de travail sur la gestion et le partage des données

Juin 2012

Un groupe de travail a été constitué sous la direction de Christine Gaspin (Inra) et de Dominique Pontier (Université Lyon 1). Ce groupe de travail est composé de Laurence Colinet (Inra), Frédéric Dardel (Université Paris Descartes), Alain Franc (Inra), Odile Hologne (Inra), Olivier Le Gall (Inra), Nicolas Maurin (Inra), Guy Perrière (CNRS), Christian Pichot (Inra), François Rodolphe (Inra).

Avis du conseil scientifique

Concernant le rapport « Gestion et partage des données »

Le conseil scientifique a pris connaissance du rapport « gestion et partage des données », dont il partage les analyses. Il remercie Dominique Pontier et l'ensemble des membres du groupe pour leur implication dans ce travail et la remarquable qualité de la réflexion conduite.

Adopté à l'unanimité le 24 mai 2012

(12 votants)

RESUME

Depuis quelques années, la biologie et les sciences humaines et sociales font face à un accroissement exponentiel des données, provenant de l'adoption en masse des nouvelles technologies, et du développement des sciences et techniques de l'information d'une ampleur et à une échelle sans précédents. Une telle rupture nécessite des transformations stratégiques majeures pour assurer le stockage, la préservation, l'exploitation de ces masses de données, mais aussi leur partage. Elle nécessite également une prise de conscience et une modification des pratiques des ingénieurs et chercheurs de l'institut, pour lesquels ces évolutions constituent un défi culturel.

Au terme de son analyse, le groupe de travail propose les recommandations suivantes :

1) Définir la politique de l'établissement et la communiquer.
2) Mettre en place un comité d'évaluation des données produites par l'Inra.
3) S'impliquer dans les comités internationaux de standardisation.
4) Développer un portail d'accès à un ensemble de ressources distribuées.
5) Prendre en compte le cycle de vie des données dès l'élaboration des projets de recherche.
6) Définir un cahier des charges pour les plateformes.
7) Doter l'Inra d'infrastructures dimensionnées pour les stockages et les calculs hautes performances.
8) S'engager dans une politique de gestion des compétences répondant aux besoins en émergence.
9) Conduire une réflexion inter-organismes pour promouvoir une politique nationale et locale en matière de gestion et partage de données.

SOMMAIRE

INTRODUCTION

En septembre 2009, la direction générale de l'Inra a saisi le Conseil Scientifique d'une réflexion prospective relative à la gestion et au partage des données. Cette réflexion se positionne dans un contexte d'évolution extrêmement rapide des technologies d'acquisition de données, dont la production massive qui en résulte a conduit au mouvement du *Big Data*. L'Inra s'est alors interrogé sur ses capacités à gérer, traiter et interpréter les très gros volumes qui sont/seront produits dans le cadre de ses programmes de recherche. Cette réflexion s'inscrit aussi dans le mouvement de l'*Open Data*, qui vise à rendre accessibles les données publiques.

Initialement envisagée en biologie intégrative et systémique, la réflexion demandée a été élargie à l'ensemble des domaines scientifiques de l'Inra pour lesquels le haut débit est (ou sera très vite) une réalité. Cette demande n'intégrait pas le volet portant sur « l'analyse et l'interprétation de gros volumes de données » qui représente aussi en soi un défi et méritera des investigations complémentaires.

Un groupe de travail a été constitué sous la direction de Christine Gaspin (Inra) et de Dominique Pontier (Université Lyon 1). Ce groupe de travail est composé de Laurence Colinet (Inra), Frédéric Dardel (Université Paris Descartes), Alain Franc (Inra), Odile Hologne (Inra), Olivier Le Gall (Inra), Nicolas Maurin (Inra), Guy Perrière (CNRS), Christian Pichot (Inra), François Rodolphe (Inra).

L'objectif de ce groupe était complémentaire de ceux de l'audit d'Ernst & Young commandité par l'Inra sur ses systèmes d'information (portés soit par la Direction des Systèmes d'Information, soit par des départements de recherche). Le groupe a analysé la situation actuelle de la gestion et du partage des données et les enjeux auxquels l'Inra doit faire face. Les principales différences par rapport à cet audit sont :

- L'emploi d'un regard scientifique sur les spécificités thématiques en matière d'acquisition de données, mais aussi de besoins en gestion et partage des données à l'Inra.
- Un regard sur les changements induits par les mouvements du *Big Data* et de l'*Open Data*.
- L'analyse de quelques cadres nationaux et internationaux existants.

Le rapport est organisé autour de quatre chapitres : un état des problématiques, une analyse des évolutions en cours, un état de la situation à l'Inra, une analyse SWOT des forces et des faiblesses de l'institut. Le rapport se conclut par quelques recommandations que le groupe a souhaité formuler.

Pour mener à bien son travail, le groupe s'est appuyé sur : (i) le rapport d'audit d'Ernst & Young et sur un échange avec les personnes qui ont conduit cet audit ; (ii) des interviews d'experts de différentes communautés et de différents organismes ; (iii) des enquêtes auprès d'ingénieurs, chercheurs d'Unités Expérimentales (UE), d'Unités de Service (US), de plateformes, de CATI (Centre Automatisé de Traitement de l'Information) et de métaprogrammes de l'Inra ; (iv) l'analyse d'un ensemble de documents et d'articles scientifiques ; et (v) sur la connaissance qu'avaient certains membres du groupe de quelques grandes initiatives qui se mettent en place à l'échelle nationale ou internationale. Notons que le travail du groupe ne présente pas le caractère exhaustif d'un audit.

1. ÉTAT DES LIEUX

Au cours des toutes prochaines années, on produira dans le monde plus de données de recherche que tout ce qui a été produit dans l'histoire de l'humanité (Beagrie, 2007). Par exemple, dans le cas des séquences d'ADN, la quantité de données générées entre août et octobre 2011, dépasse en nombre de nucléotides tout ce qui a été produit pendant la période 1982-1997[1]. Les nouvelles technologies, en progrès constant, vont permettre d'augmenter considérablement les vitesses et volumes de production des données. Ces données seront générées à partir d'appareils à très haut débit comme les séquenceurs, de simulations numériques de haute performance, de capteurs (*e.g.*, environnementaux), d'imageries scientifiques, de satellites. Ce qui est nouveau, c'est que les équipements servant à la production des données à haut débit deviennent accessibles à tout chercheur, quels que soient son champ disciplinaire et la taille de sa structure d'accueil (*e.g.*, petits séquenceurs de laboratoire, capteurs à enregistrement continu). On peut désormais superposer différentes complexités – moléculaires, individuelles (phénotypes), populationnelles, sociologiques, et environnementales – ceci avec un enjeu qui réside dans la compréhension des interactions entre ces différentes complexités (Christakis *et al.*, 2007).

Ces évolutions technologiques dans le numérique ont conduit dans les années 2000 à l'apparition du mouvement *Big Data* qui inverse le paradigme précédent dans lequel la production de données était gouvernée par la nature des problématiques. A l'heure actuelle la question est donc plus de savoir comment organiser et analyser ces données que de les produire. Plus récemment, l'*Open Data* s'inscrit dans cette dynamique technologique.. L'enjeu au départ était de rendre « transparentes » aux citoyens les grandes quantités de données gouvernementales. Depuis le phénomène, toujours en cours d'évolution, s'est étendu à l'ensemble des données, posant à tous des questions d'ordre technologique, stratégique, éthique et juridique.

L'évolution extrêmement rapide des technologies d'acquisition des données génère aussi dans certains domaines une obsolescence rapide des outils permettant d'en exploiter la production, dès lors que ces outils sont spécifiques d'une technologie particulière. Il n'en reste pas moins que des outils phares peuvent perdurer sur de très longues périodes sans être remplacés. L'exemple type est le logiciel BLAST (Basic Local Alignment Search Tool). En génomique comparative, ce programme de 20 ans d'âge reste encore aujourd'hui la référence incontournable en matière de recherche de similarités dans les banques de données de séquences (Altschul *et al.*, 1990, 1997). Les deux articles décrivant l'algorithme et ses évolutions ont été cités plus de 61000 fois depuis 1990 !

1.1. CONTEXTE

La majorité des données est désormais disponible sous forme numérique. Au plan politique, dans le cadre d'accords récents, il existe une volonté forte pour qu'elles deviennent des objets communautaires. En 2004, 30 pays membres de l'OCDE dont la France, ainsi que quatre autres pays (Afrique du Sud, Chine, Russie et Israël) ont signé la déclaration ministérielle sur l'accès aux données de la recherche financée par des

[1] ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt

fonds publics[2]. C'est ainsi qu'un certain nombre de pays et d'organismes de recherche ont commencé à définir des politiques et à développer des infrastructures de dépôt et d'accès aux données de la recherche financée sur fonds publics[3]. Un exemple de politique prenant en compte le volet conservation des données est celui fourni par un équivalent de l'Inra au Royaume-Uni : le BBSRC[4] (Biotechnology and Biological Sciences Research Council). Il est à noter que la France présente un retard important dans ce type d'initiative par rapport à d'autres pays. La seule obligation au niveau européen est la directive Inspire[5] (Infrastructure for Spatial Information in the European Community ; *cf.* Annexe A.3. du rapport) 2007/2/CE du 14 mars 2007. Cette directive stipule que toutes les données géographiques produites par des financements publics doivent être rendues accessibles.

Au plan scientifique, la nécessité de partage national, international et pluridisciplinaire des données de la recherche augmente, guidée par des questions de portée mondiale comme l'alimentation, le climat de la terre, la santé humaine et la biodiversité. De plus en plus d'initiatives (*e.g.*, Année Polaire Internationale, travaux du GIEC, métagénomes des sols ou du tube digestif humain, projet 1000 génomes[6]) voient le jour pour travailler en commun sur les données et pour les combiner avec d'autres. Les avantages du partage sont considérables pour la recherche et l'innovation, et potentiellement, pour l'Inra :

- Valeur ajoutée créée par la mise en place de nouvelles collaborations nationales et internationales.
- Reproductibilité des analyses et amélioration des méthodes.
- Mise en place de duplications servant de filets de sécurité pour la préservation des données. En effet, actuellement seulement 1 % des données collectées en écologie seraient accessibles publiquement après publication (Reichman *et al.*, 2011), ce qui veut dire que, potentiellement, 99 % pourraient être perdues accidentellement, nonobstant les sauvegardes personnelles ou locales. Le supplément de la revue *Science* du 11 février 2011 est d'ailleurs consacré à ce problème dans différents champs scientifiques.
- Possibilité d'entreprendre de nouvelles recherches sur de grands ensembles de données rassemblés. Quand l'information devient si abondante et diverse, de nouvelles découvertes sont rendues possibles.

Par ailleurs, pour un organisme producteur de données tel que l'Inra, la grande diversité des espèces étudiées et la variété des objectifs finalisés sont sources d'enjeux nouveaux d'efficience, de visibilité et de prestige associés à la mise en place de structures de partage et de contrôle de qualité. Par exemple, une étude a montré que les publications d'essais cliniques sur le cancer pour lesquelles les données étaient rendues publiques ont été citées environ 70 fois plus fréquemment que celles pour lesquelles les

[2] http://www.oecd.org/dataoecd/9/60/38500823.pdf
[3] Voir http://www.dcc.ac.uk/resources/policy-and-legal/institutional-data-policies pour des universités du Royaume Uni.
[4] http://www.bbsrc.ac.uk/ web/FILES/Policies/data-sharing-policy.pdf
[5] http://www.developpement-durable.gouv.fr/La-directive-europeenne-Inspire-de.html
[6] Le projet 1000 Génomes propose de cataloguer les variations génétiques d'au moins un millier d'individus d'origines diverses à partir du séquençage de leur génome (Via *et al.*, 2010).

données n'étaient pas associées, ceci indépendamment du journal (Piwowar *et al.*, 2007). Concernant la qualité, dans la plupart des cas, celle-ci est sous la seule responsabilité du déposant et non du gestionnaire de la base. Par exemple, dans le cas des métagénomes, il est connu que le taux d'erreur d'identification taxonomique dans les bases de données est de l'ordre de 5 % (Ashelford *et al*., 2005).

De manière générale, une production dispersée, non « contrôlée » et sans recours systématique à des métadonnées posera des problèmes d'identification (savoir que la donnée existe), d'accès (localiser l'endroit où est la donnée), d'hétérogénéité et de contrôle qualité.

La décennie 2010–2020 sera celle des infrastructures et celle du partage : partage des données, partage des moyens, partage des infrastructures. Le message est maintenant bien accepté par la grande majorité des organismes de recherche en France, mais aussi en Europe et dans le monde. Par contre, se pose la question de la compatibilité de cette politique de partage avec une autre demande pressante qui est celle du dépôt de brevets par les organismes de recherche publics. Du fait du caractère massif des investissements requis, les deux composantes majeures de cette révolution technologique sont les infrastructures et la mise en place de pratiques adaptées. C'est pourquoi on assiste de plus en plus au développement de consortiums lors du montage de projets de recherche de grande envergure.

1.2. QU'EST-CE QU'UNE DONNÉE, QUELLES DONNÉES CONSERVER ET QUE PARTAGE-T-ON ?

1.2.1. Qu'est-ce qu'une donnée ?

Définitions

En l'absence d'une définition juridique de ce qu'est une donnée, nous appellerons dans le contexte de ce rapport, donnée scientifique, ou donnée de la recherche, l'enregistrement d'une information qui représente le matériau de base d'une activité de recherche ayant bénéficié d'un financement sur fonds publics. Une donnée scientifique peut être le résultat d'une expérience, généralement issue d'un instrument (robot, capteur, enregistrement audio ou vidéo, etc.), ou bien d'une observation humaine sur le monde naturel. Profitant des progrès de l'informatique et des outils du web qui ont élargi les possibilités du travail collaboratif, les données acquises dans le cadre des sciences citoyennes, encore appelées sciences participatives se développent en France depuis les années 90, en associant aux programmes de recherche une participation « citoyenne » de volontaires qu'ils soient amateurs éclairés, spécialistes à la retraite, etc. Observées essentiellement dans les domaines où l'acquisition de l'information se fait sur le terrain (réalisation d'observations, de mesures, de comptages, etc.), elles contribuent aux avancées de la recherche en matière de connaissance et d'inventaires. Les programmes VigiENature[7] et PhenoClim[8] constituent deux exemples français de ce type d'acquisition de données. Les formes sous lesquelles sont actuellement stockées ces données (scientifiques ou sciences participatives) sont multiples (papier, fichier, bases de données).

[7] http://vigienature.mnhn.fr/
[8] http://www.crea.hautesavoie.net/phenoclim/

Deux types de données scientifiques peuvent être clairement identifiés. Le premier est représenté par les données temporaires, liées à des résultats intermédiaires. Leur volume est potentiellement très important et peut être néanmoins régulé (par exemple, une fois les résultats obtenus, les données intermédiaires sont généralement supprimées). Le deuxième type est représenté par les données pérennes, constituées des résultats de recherche (souvent en vue d'une publication) et des données brutes qui ont conduit à ces résultats. Les données brutes revêtent un caractère critique et sont généralement conservées.

La notion de donnée brute est éminemment variable suivant le domaine considéré et peut même évoluer. Dans le cas de la génomique, les images qui représentaient les données brutes, nécessitent d'être traitées en temps réel. La communauté a ainsi reconsidéré leur statut, et désormais les données brutes sont représentées par les séquences (associées à une note de qualité) qui sont issues de ces images. Reliée à cette question figure la problématique de l'obsolescence technologique qui fait qu'une donnée obtenue peut devenir au fil du temps inintéressante, car dépassée, voire inutilisable.

Selon les thématiques, on constate des différences dans les besoins, les pratiques, les organisations et les technologies dans le domaine de la gestion et du partage des données. Si, en génomique, les données sont relativement homogènes (séquences d'ADN produites par des dispositifs standardisés, avec un petit nombre de constructeurs), les données de l'écologie, de l'agronomie et de l'environnement, se caractérisent par leur diversité, leur complexité et leur hétérogénéité.

Les données ouvertes

Selon *l'Open Government Data Group*[9], les données dites ouvertes doivent satisfaire à huit principes :

- Complètes – chaque jeu de données doit comporter toutes les données disponibles à l'exception des données sujettes à des limitations concernant la vie privée, la sécurité ou des privilèges d'accès.
- Primaires – les données ouvertes sont des données brutes, prises directement à la source, aussi détaillées que possible et sans traitement ni modification.
- Opportunes – les données doivent être rendues disponibles aussi vite que possible pour être à jour.
- Accessibles – les données doivent être disponibles pour le plus grand nombre.
- Exploitables – c'est-à-dire prêtes à être traitées par des outils informatiques.
- Non discriminatoires – c'est-à-dire accessibles sans inscription.
- Non propriétaires – c'est-à-dire disponibles dans des formats ouverts.
- Libres de droits.

Ces huit principes définis au départ pour les données ouvertes d'origine publique (État, administration, collectivités locales…) sont généralement repris pour tous les types, quelles qu'en soient les sources, publiques ou privées.

[9] http://opengovernmentdata.org/

Les métadonnées

Les données sont des actifs précieux qui, pour certaines d'entre elles, ont un potentiel de réutilisation infini (*e.g.*, données sur les sols, climatiques, de séquences). Pour qu'elles soient correctement utilisables, les données doivent être caractérisées au travers de métadonnées. Celles-ci sont particulièrement importantes parce qu'elles permettent de rendre les données intelligibles par n'importe quel chercheur qui pourra, le cas échéant, les réutiliser. Les métadonnées fournissent une documentation claire et accessible sur le contenu, la structure, le contexte et les conditions de recueil des données (conditions environnementales, appareils utilisés, etc.). Idéalement, elles constituent un enregistrement de tout ce qui pourrait intéresser un autre chercheur. Introduire des métadonnées n'est pas dans la pratique de tous les chercheurs, d'une part parce que cela n'est pas perçu comme une activité scientifique, et d'autre part parce que cela demande une certaine logistique pour pouvoir documenter les données, les mettre à disposition, les rendre accessibles, et enfin, les rendre visibles.

Les métadonnées doivent à la fois fournir les informations de premier niveau décrivant succinctement la nature des jeux de données (date, lieu, thématique, support), ainsi que des informations de second niveau caractérisant finement les données (objets, traits, conditions expérimentales). La génomique a largement bénéficié de la mise en place de standards jusque dans les années 2005, mais l'évolution extrêmement rapide des nouvelles technologies questionne aujourd'hui la capacité à les adapter aux volumes, au type et à la nature des données partageables. Des standards et outils internationaux sont d'ores et déjà disponibles. Dans le domaine du séquençage haut débit, le dépôt dans la SRA (Sequence Read Archive) du NCBI[10] (National Center for Biotechnology Information) définit six types de métadonnées (Leinonen *et al*., 2011). Toujours dans le domaine des omiques, des initiatives émergent pour fédérer les communautés scientifiques à l'échelle internationale (Sansone *et al*., 2012).

En Europe, dans le domaine de l'environnement, la production des métadonnées s'inscrit dans le cadre de la directive Inspire dont l'objectif est d'« établir l'infrastructure d'information géographique dans la Communauté Européenne, aux fins des politiques environnementales communautaires et des politiques ou activités de la Communauté susceptibles d'avoir une incidence sur l'environnement ». Les standards pour les métadonnées de second niveau sont aujourd'hui en cours d'élaboration, par champ disciplinaire. On peut ainsi citer le standard EML[11] (Ecological Metadata Language) pour l'écologie, MRTG (Multimedia Resources Task Group), GBIF (Global Biodiversity Information Facility) et TDWG (Taxonomic Database Working Group) pour la biodiversité, DDI (Data Documentation Initiative) pour les sciences humaines (Vardigan *et al*. 2008), SDMX[12] (Statistical Data and Metadata eXchange) pour les données statistiques.

D'un point de vue technique, des langages de description hypertexte tels que XML (eXtensible Markup Language) permettent d'étendre pratiquement à l'infini les possibilités de description d'un objet conceptuel. Le problème des métadonnées n'est donc plus vraiment d'ordre technologique, mais plutôt une question de moyens. En effet, l'existence de ces outils nécessite une collaboration étroite des chercheurs et des ingénieurs au niveau des étapes de conceptualisation et de formalisation des données.

[10] http://www.ncbi.nlm.nih.gov/
[11] http://knb.ecoinformatics.org/software/eml/
[12] http://sdmx.org/

1.2.2. Quelles données conserver ?

Le monde scientifique est actuellement dans une logique de production massive de données, cette production ayant été rendue possible par la baisse des coûts et l'accroissement des performances technologiques. Le numéro spécial de la revue *Science* de Février 2011 illustre cette réalité dans plusieurs domaines et pose clairement les questions de la gestion et du partage de ces volumes massifs de données. Par exemple, alors que les coûts de séquençage baissent de manière drastique, les coûts de mise en place des infrastructures informatiques pour leur prise en charge ne permettent plus de répondre aux besoins (Figure 1). Il en résulte que si une nouvelle révolution technologique ne s'opère pas sur les solutions de stockage et d'archivage, alors se posera de manière urgente la question du « quoi conserver, et sur quelle durée ? ». Il s'agit d'un changement de paradigme par rapport à l'époque où il était coûteux de produire des données.

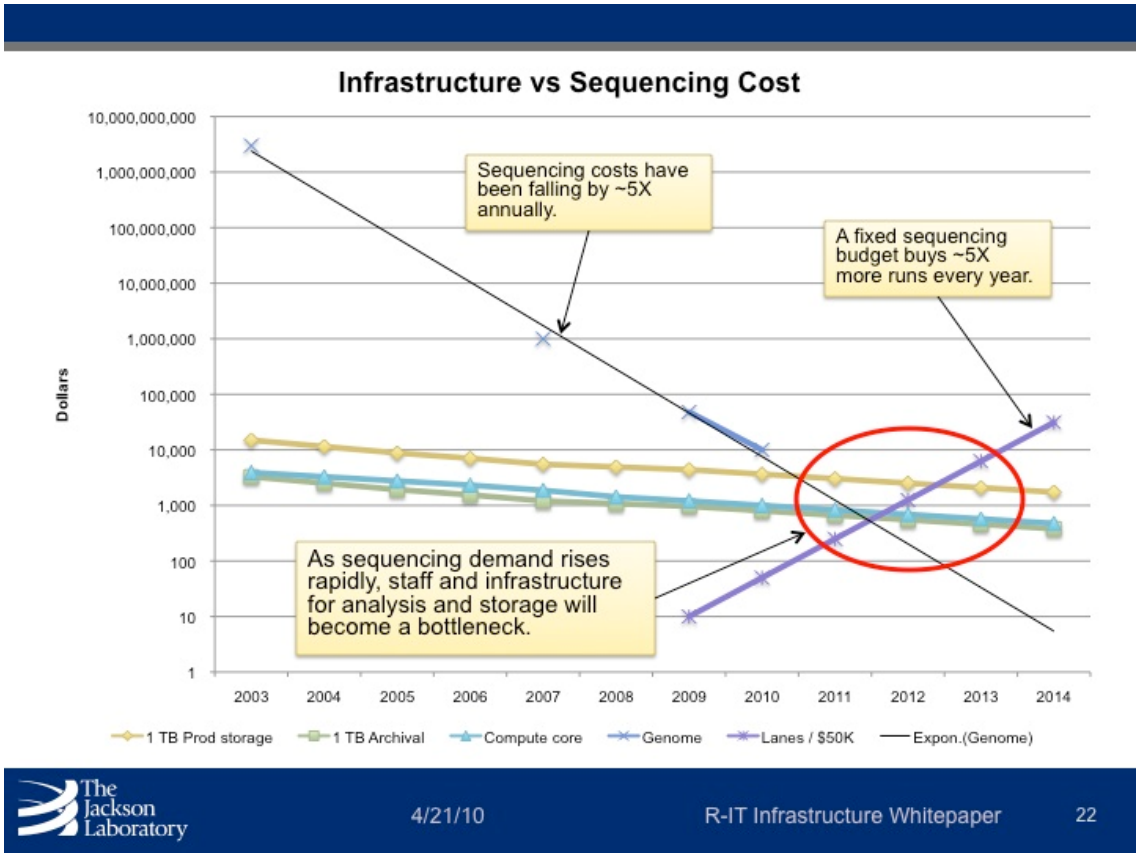

Figure 1. Comparaison entre les coûts associés aux infrastructures et ceux nécessaires au séquençage des données « omiques ». Figure tirée de FY2010 Research Information Technology Strategic Position Whitepaper.

Cette question est particulièrement cruciale pour les données brutes, tout spécialement dans le cas d'expériences non reproductibles (*e.g.*, suivi de cohortes, événements rares, expériences nécessitant la destruction du matériel d'étude).

1.2.3. Que partage-t-on ?

Le caractère partageable des données est généralement déterminé par leur niveau de confidentialité, de documentation, de gestion, d'accès et de préservation. Dans ce cadre, la définition de standards et d'ontologies est considérée comme essentielle, recouvrant des aspects techniques (formats des données, développement logiciel, etc.), mais aussi les métadonnées (Sansone *et al*., 2012). En écologie, les jeux de données sont, pour la plupart, dispersés dans des milliers de laboratoires, « cachés », non référencés (hors publication), non accessibles et sujets à perte. Des initiatives en cours s'attaquent à ce problème, avec le projet DataONE[13] qui fédèrera les initiatives en cours puis le projet GEOSS[14] qui devrait rassembler les grandes fédérations.

1.3. PROPRIÉTÉ INTELLECTUELLE, PROPRIÉTÉ DES DONNÉES, ÉTHIQUE

Le partage des données emporte des questions juridiques et administratives qui vont au-delà des compétences du groupe, mais qui nécessitent d'être mises en avant. Compte-tenu de leur complexité, ces questions importantes et transversales mériteraient à elles-seules un autre groupe de travail.

Eléments juridiques relatifs au partage des données

Le partage national, international et de plus en plus pluridisciplinaire des données issues de la recherche est essentiel pour accélérer la conversion des résultats en savoirs et/ou innovations. Ce partage nécessite de prendre en compte les aspects liés à la propriété intellectuelle pour envisager les conditions de la réutilisation des données. En premier lieu, il est nécessaire d'établir un diagnostic des questions juridiques et de culture juridique à résoudre :

- L'indifférence, la méconnaissance, voire la défiance, de la plupart des chercheurs relativement aux problèmes juridiques.
- Le développement de la recherche au sein de consortiums internationaux regroupant des organismes n'ayant pas les mêmes environnements juridiques ni les mêmes positions stratégiques. Le problème qui se pose est alors de savoir comment respecter les mesures légales et les politiques des différents organismes ou pays.

Rappelons que la réutilisation de l'ensemble des données publiques représente aujourd'hui un impact économique et social potentiel évalué à quelques 140 milliards d'euros par an en Europe[15]. L'importance croissante de l'infrastructure numérique pour l'exploration et l'exploitation des biens communs a accéléré l'adaptation à cette situation nouvelle du régime actuel des droits de la propriété intellectuelle. La liberté d'accès des citoyens aux documents administratifs, qui fondait le cadre ancien de règlementation est devenu, sous l'impulsion du développement d'Internet, obligation de communiquer pour l'administration, et droit à réutiliser pour le citoyen (Décret n° 2011-577 du 26 mai 2011). Dans le cadre de cette mission, nous nous intéresserons exclusivement aux données publiques communicables, c'est-à-dire ce que l'on appelle

[13] http://dataone.org
[14] http://www.earthobservations.org/geoss.shtml
[15] http://ec.europa.eu/information_society/policy/psi/docs/pdfs/opendata2012/open_data_communication/fr.pdf

usuellement des documents administratifs. L'Inra, qui est un établissement public à caractère scientifique et technologique (EPST), est donc directement concerné par cette évolution.

Sont considérés comme documents administratifs – quels que soient leur date, leur lieu de conservation, leur forme et leur support – les documents produits ou reçus, dans le cadre de leur mission de service public, par l'État, les collectivités territoriales ainsi que par les autres personnes de droit public ou les personnes de droit privé chargées d'une telle mission. Constituent de tels documents notamment les dossiers, rapports, études, comptes rendus, procès-verbaux, statistiques, directives, instructions, circulaires, notes et réponses ministérielles, correspondances, avis, prévisions et décisions (*in* Article 1 de la Loi 78-753 du 17 juillet 1978).

L'activité de recherche publique s'exprime sous forme de publications. La liberté de décider d'une publication et d'en fixer le contenu est nécessaire au contrôle, par un chercheur, du degré d'achèvement de sa réflexion et au respect de l'intégralité de celle-ci. Dans ce contexte, depuis 2006, le droit français, au nom du droit d'expression du chercheur, a rendu le chercheur public-auteur seul propriétaire de ses droits de propriété incorporelle sur ses œuvres originales[16].

De plus, les documents sur lesquels des tiers détiennent des droits de propriété intellectuelle font aussi partie des documents non communicables. Ce dernier point mérite d'être relevé, ceci du fait des changements d'échelle qui s'opèrent actuellement dans la recherche académique. Ces changements passent par la mise en place de consortiums nationaux (multi-organismes et/ou public-privé), voire internationaux. Une conséquence de la mise en place de tels consortiums est l'augmentation du nombre de copropriétés qui vient s'additionner à celle déjà existante du fait des multi-tutelles (*e.g.*, les unités mixtes du CNRS ou de l'Inra). La négociation des accords de consortiums doit être l'occasion de formaliser en amont les dispositions permettant aux partenaires de faire face à leurs obligations de rendre publiques leurs données.

En conclusion, les données communicables par l'Inra sont les seules lui appartenant et exemptes d'interdiction de communication. Du fait du développement des recherches partenariales, l'essentiel des données de l'Inra soumises à accessibilité, diffusion et réutilisation sont pour l'essentiel ses bases de données, ses logiciels et les résultats protégés au titre de la propriété industrielle. C'est sur ces sujets qu'il apparaît nécessaire de procéder à un approfondissement quant aux règles de communicabilité.

Il convient de distinguer à l'égard des dispositifs de traitement de données trois composantes : le support logiciel, la base de données et les données proprement dites. Ces trois composantes peuvent (ou non) faire l'objet de développements indépendants. Il est donc nécessaire de déterminer quels en sont le ou les titulaires des droits d'exploitation. Pour la base de données, le producteur titulaire initial des droits est celui qui prend l'initiative et le risque des investissements. Au titre des investissements trois points sont à examiner : qui peut attester de l'investissement financier matériel et humain substantiel ayant permis d'effectuer le développement de la base de données ? Cette question se complique lorsque les dites bases de données sont constituées par une communauté – par exemple de chercheurs – représentant plusieurs organismes répartis dans le monde entier. La protection à mettre en place est notamment caractérisée par

[16] Librement adapté de l'amendement n° 152 présidant à l'adoption de L111-1 dans la loi du 01-08-2006.

une définition du titulaire des droits et de l'exploitation pouvant en résulter tant en dépôt que lors d'une réutilisation.

Ethique et sécurisation des données

Dans le cas de données concernant des personnes physiques, il existe une demande forte de la part des SHS pour sécuriser au maximum l'anonymisation des données, en particulier au travers du cryptage. En effet, il est d'ores et déjà possible de lever l'anonymat sur certaines personnes en procédant par recoupement d'informations disponibles sur Internet (*cf.* section 2.). En 2008, pour prévenir certains abus dans le domaine médical, la loi américaine GINA[17] (Genetic Information Nondiscrimination Act) a été votée afin de pallier l'usage impropre d'informations génétiques sur les particuliers. L'anonymat n'est toutefois pas la seule dimension éthique ; et d'autres droits fondamentaux sont assurés en France par la loi « informatique et libertés » qui est le cadre d'interrogation par essence. En Europe, la directive 95/46/EC constitue la base de la protection des droits de la personne. A une époque où le numérique simplifie les transferts d'informations, ce n'est pas tant le partage des informations qui pose problème mais bien de savoir à qui, et surtout à quoi, un individu veut bien donner accès. Depuis, de nombreuses initiatives et consultations ont eu lieu pour la révision de la directive européenne[18].

Au-delà de la sécurisation des données couramment prise en charge par les administrateurs système, d'autres solutions, tel le cryptage de gros volumes de données, posent problème pour les traitements ultérieurs. Dans le cas des données de la génomique, seule la possibilité de masquer certaines portions du génome semble pouvoir être effective. La nécessité de demander, *via* un formulaire d'enquête, une autorisation d'accès aux données sur la personne semble émerger comme alternative dans les grands centres médicaux ou consortiums (Greenbaum *et al*., 2011). Dans un contexte de production de données sur les personnes, il apparaît indispensable de clarifier précisément plusieurs niveaux. Ces niveaux incluent (A. Cambon-Thomsen, comm. pers.) : (i) la source des données et le consentement des personnes ; (ii) une identification claire des propriétaires des données et de la propriété intellectuelle liée aux bases de données et outils générés ; (iii) de décider des données qui seront en accès libre ou sous contrôle (et de préciser sous quelles formes) ; (iv) les procédures de vérification des outils rendus disponibles en regard des contraintes précédentes.

1.4. COÛTS INDUITS

Les données deviennent de moins en moins coûteuses à produire, et leur volume et leur caractère fréquemment innovant entraînent des coûts de traitement de plus en plus élevés, à la fois en termes de ressources informatiques et humaines.

Planification expérimentale

Du fait de la facilité à produire des données en grandes quantités, les biologistes peuvent avoir tendance à négliger la démarche indispensable de planification expérimentale. Ces deux aspects sont liés et leur connexion est connue depuis longtemps, mais cette difficulté est exacerbée dans un contexte de massification des

[17] http://www.genome.gov/pages/policyethics/geneticdiscrimination/ginainfodoc.pdft
[18] http://ec.europa.eu/justice/data-protection/index_en.htm

données. Dans le pire des cas, l'absence d'interactions entre le biologiste et le modélisateur peut conduire à l'obtention de données inutiles et à la publication de résultats statistiquement non robustes et non avérés, c'est-à-dire un gaspillage de temps et de moyens.

Coûts matériels : stockage, calcul

Certaines problématiques requièrent le recours à des moyens de stockage et de calcul massifs (*e.g.*, génomique, biologie systémique, modélisation). A ce jour, la plupart des infrastructures ouvertes offrent gratuitement ces services. Cependant, l'évolution rapide des capacités de production, et donc des besoins, rendent incontournables à terme une participation financière des utilisateurs (*cf.* rapport du groupe de travail DGRI[19] « e-infrastructures pour la Génomique et la Biologie à Grande Echelle », 2012) et la définition d'une politique de préservation des données.

Le problème de l'analyse des données peut aussi être rendu complexe du fait de l'évolution des logiciels et des systèmes d'exploitation. En effet, les fichiers au format texte traversent le temps, mais ce n'est pas le cas des données binaires. Il existe de nombreux exemples de programmes scientifiques qui sont apparus, sont devenus des standards dans leur domaine, et ont ensuite disparu au bout de quelques années. De nombreux fichiers contenant des données dans le format de ces programmes sont ainsi devenus illisibles. En liaison avec cette évolution technologique perpétuelle, un coût induit indirect est la nécessité qu'ont les personnels de se former tout au long de leur carrière. Il s'agit là d'un processus inhérent à la recherche scientifique, mais qui s'est beaucoup accéléré au cours de cette dernière décennie.

Enfin, il existe un coût important lié à la pérennisation des données. Ces coûts proviennent : (i) de la nécessité d'effectuer des mises à jour régulières dans le cas des bases de données ; (ii) de suivre l'évolution des supports logiciels et des systèmes d'exploitation (mise à niveau des versions, compatibilité ascendante, passage à des programmes concurrents) ; (iii) de suivre l'évolution matérielle (changement de processeur, de technologie ou de constructeur).

Coût humain

De nombreuses données nécessitent d'être structurées sous la forme de bases de données afin de pouvoir être pleinement utilisables et partageables. Du fait de l'accumulation et du manque de moyens humains, il est désormais de plus en plus fréquent que les données soient à la fois mal analysées et sous-employées (Richardson et Watson, 2012).

La diversité des applications rendues possibles mais aussi la diversité des données acquises et des technologies d'acquisition rendent nécessaire le développement de méthodologies appropriées et de nouveaux standards. Les problèmes soulevés par l'accroissement des volumes de données portent non seulement sur leur analyse statistique, mais aussi sur leur analyse « tout court » (extraction d'information, traitement algorithmique). Ces traitements doivent pouvoir être adaptés, optimisés, de façon à pouvoir supporter le changement d'échelle. Par ailleurs, le volume des données

[19] Direction Générale de la Recherche et de l'Innovation au MESR (Ministère de l'Enseignement Supérieur et de la Recherche).

change aussi la nature des questions qui peuvent être traitées, et pose des problèmes de modélisation en amont de tout traitement, algorithmique ou statistique.

2. QU'EST-CE QUI EST EN TRAIN DE CHANGER ?

2.1. LA DÉMOCRATISATION DU TRÈS HAUT DÉBIT

L'exemple de la génomique est particulièrement marquant à cet égard. En effet, trois générations de séquenceurs se sont succédé en cinq ans multipliant pour les deux premières les capacités de production par deux ordres de grandeur. La production d'un *run* de séquenceur HiSeq 2000 est aujourd'hui de l'ordre de 600 Gigabases. La production des données est actuellement centralisée sur des plateformes bien identifiées, cependant, la baisse des coûts des équipements devrait permettre l'arrivée rapide du haut débit dans de nombreux laboratoires de biologie.

Récemment, Oxford Nanopore Technologies®[20] a annoncé la commercialisation de deux nouveaux types de matériel qui devraient une nouvelle fois révolutionner le domaine du séquençage. Le premier type de matériel, GridION, permet de mettre en grille un ensemble de séquenceurs dont les ressources pourront être utilisées selon les besoins du projet. Par exemple, 20 de ces unités utilisées simultanément permettraient de séquencer l'équivalent d'un génome humain en 15 minutes seulement, pour un coût de 1000 dollars. Le second type de matériel, MinION, est une unité de séquençage intégrée dans une clef USB. Cet outil permettrait d'analyser un échantillon d'ADN en milieu aqueux et de directement transférer les séquences lues sur l'ordinateur, et ceci pour moins de 1000 dollars. Cette généralisation du haut débit questionne les grands centres internationaux sur leurs capacités à continuer à rendre le service du stockage et de la mise à disposition des données pour l'ensemble de la communauté scientifique[21].

De la même manière, dans le domaine de l'écologie et de l'environnement, des techniques d'acquisition de données en flux continu (capteurs, vidéo, etc.) et/ou à plus hautes résolutions (*e.g.*, LIDAR – Light Detection and Ranging) sont apparues et sont désormais facilement accessibles. Ces méthodes conduisent également à une forte augmentation des volumes produits.

Ces (r)évolutions technologiques ont entraîné une diversification des capacités expérimentales qui modifie la nature des questions posées. On commence à aborder des problématiques aussi diverses que les variations génomiques entre individus, les conséquences des conditions environnementales sur l'expression des génomes et des phénotypes, la biodiversité. Ces questions intègrent de plus en plus des sources d'informations différentes, plusieurs niveaux d'organisation biologiques et échelles spatiales et temporelles. Elles exigent de nouvelles recherches en mathématiques et en informatique fondamentale pour leur analyse, le développement de logiciels de traitement des données, car l'augmentation exponentielle des volumes de données exclut un traitement manuel.

2.2. LE BESOIN D'INFRASTRUCTURES DIMENSIONNÉES ET SÉLECTIVES

Les volumes de données produits sont aujourd'hui tels que l'on peut souligner plusieurs

[20] http://www.nanoporetech.com/
[21] http://www.genomeweb.com/node/962092/

limitations aux infrastructures TI (Technologies de l'Information) existantes. Une première limitation est constituée par les débits réseaux disponibles lorsqu'on souhaite rapatrier des données d'une structure de production ou de stockage vers une infrastructure de traitement (Kahn, 2011). Une deuxième limitation est posée par la dimension des moyens de stockage et de calcul de la plupart des laboratoires. En effet, la multiplicité des jeux de données de grande dimension rend nécessaire la présence d'infrastructures de proximité capables de les stocker et de les analyser. Cependant, du fait de leur croissance exponentielle, la préservation de l'ensemble des données ne peut plus être garantie sur le long terme.

Plusieurs solutions émergent pour faire face à ces limitations. L'une d'elles consiste à sélectionner les données stockées et mises à disposition sur la base de leur qualité et donc de la visibilité qu'elles sont en mesure d'apporter. Une autre solution consiste à demander une participation financière aux producteurs de données. Le *cloud* (privé ou académique) constitue une alternative pour, d'une part, répondre aux besoins des laboratoires mais aussi, d'autre part, pour partager les jeux de données de grande dimension, tout en évitant leur duplication. Les architectures orientées services constituent aussi une solution visant à rapprocher le calcul des données, ceci en lien avec des grilles de calcul. Concernant les analyses, des technologies spécifiques de gestion de fichiers (mémoire de grande taille partagée par plusieurs cœurs de calcul) deviennent de plus en plus nécessaires, ces solutions commençant à être implémentées de façon courante.

2.3. LE *CLOUD* OU INFORMATIQUE EN NUAGE

Pour faire face aux besoins, l'informatique en nuage offre aujourd'hui un concept opérationnel qui permet d'accéder, *via* le réseau et à la demande, à des ressources informatiques de stockage et de calcul virtualisées et mutualisées. Ce concept est aujourd'hui de plus en plus en vogue dans les contextes du *Big Data* et de l'*Open Data*. Dans le cas de *clouds* mis en place par des prestataires privés (*e.g.*, Google, Amazon, Microsoft, Apple), les inconvénients attribués à ce concept sont : (i) la soumission au cadre légal en vigueur dans le pays qui accueille les données ; (ii) les temps et coût de transfert (à relier avec les débits en vigueur et les garanties offertes sur ces débits) ; (iii) la dépendance au prestataire de service ; (iv) la sécurisation des données (répartition unique ou distribuée, tests d'intrusion, sécurisation des locaux) ; et (v) la pérennité du service. Des expériences telles que celles de DigitalOne (saisie des serveurs par le FBI) ou Megaupload (fermeture du site) constituent deux illustrations concrètes de ces inconvénients.

2.4. LA « DÉSANONYMISATION » DES INDIVIDUS

Les capacités offertes aujourd'hui dans le domaine du séquençage (génotypage à grande échelle, etc.) ainsi que l'accès libre à certaines données sur la personne ont imposé de revoir les recommandations et les règles de sécurisation des données et de la protection de l'anonymat. Qui plus est, il a également été nécessaire de revoir les conditions d'utilisation des données obtenues par consentement (Greenbaum *et al.*, 2011). La possibilité de lever l'anonymat des personnes a ainsi conduit à la fermeture temporaire de l'accès à certaines bases de données du NIH (National Institute of Health). Cette fermeture a été maintenue jusqu'à ce qu'il soit possible de déterminer si toutes les conditions permettant de lever le risque de désanonymisation des individus étaient

vérifiées (Kahn, 2011).

Dans le domaine des Sciences Humaines et Sociales (SHS), l'accès aux données digitalisées disponibles dans des organisations diverses représente une source encore peu exploitée de manière « transparente » et ouverte. La préservation de l'anonymat des personnes y reste un problème difficile, comme dans le domaine médical (Greenbaum *et al*., 2011) et les solutions de préservation doivent y évoluer. On notera qu'en SHS, des jeux de données de la taille de ceux manipulés dans le domaine des omiques existent, mais certains ne sont pas accessibles gratuitement.

2.5. LA QUESTION DU PARTAGE SELON LA DISCIPLINE

Les méthodes et les usages en cours pour la gestion et le partage des données issues du processus de recherche varient beaucoup d'une discipline à l'autre. Ceci s'explique par l'histoire et les caractéristiques des contraintes propres à chaque domaine de recherche. Traditionnellement, la physique des particules, l'astrophysique et la génomique ont une pratique de stockage et de mise en commun des données. Ainsi en génomique on ne peut pas déposer de publication sans un numéro de dépôt (*accession number*) que l'on n'obtient que si on complète les métadonnées. Pour la physique des particules, le stockage, l'archivage et le partage des données sont conçus en même temps que l'expérience.

Dans d'autres disciplines, comme l'écologie, l'agronomie, ou encore les SHS, il n'y pas de partage des données par habitude. En effet, il existe dans ces domaines une connexion très forte entre obtention et traitement des données. Un chercheur peut publier sur les données qu'il a lui-même recueillies, sans avoir besoin de faire appel à des données d'autres chercheurs. De ce fait il peut être réticent à les rendre accessibles. Les raisons de ce refus sont multiples : désir de garder le contrôle de son matériel de production scientifique, mauvaise évaluation de la valeur ajoutée du partage, ou encore manque de temps et absence d'environnement adéquat pour rendre accessibles ses données.

Dans le même temps, on observe une pression croissante pour publier les données, notamment en écologie. Depuis 2011, outre les revues *on-line* telles que les *PLoS*, d'autres journaux comme *Ecological Monographs*, *Evolution*, *American Naturalist*, ou *Molecular Ecology* obligent désormais à publier les données. Enfin, beaucoup d'autres encouragent fortement cette pratique (*e.g.*, *Ecology*, *Ecology Letters*). Ces revues proposent de déposer les données sur le dépôt Dryad[22] mais il est possible de les déposer sur des serveurs autres, avec la possibilité de les conserver privées durant une période pouvant aller jusqu'à une ou deux années. L'idée principale est de permettre la réplication des résultats. La revue *Nature* indique ainsi clairement sur son site que si les auteurs ne mettent pas à disposition leurs données, ils doivent le signaler à la revue qui informera alors ses lecteurs que les résultats ne sont pas reproductibles – une publicité que les auteurs voudront à coup sûr éviter ! Il s'agit là d'un mouvement culturel de fond, du fait de l'accroissement des exigences de transparence à tous les niveaux.

Par ailleurs, la forte croissance des données, liée à l'explosion du nombre de laboratoires produisant des données haut débit, sature les collections centralisées (*e.g.*, GenBank, EMBL et DDBJ dans le cas des séquences biologiques), à tel point que la

[22] http://datadryad.org/

survie de ce modèle est remise en question. Une conséquence de ceci est la prolifération de petites bases de données spécialisées. C'est ainsi que Cochrane et Galperin (2010) ont recensé 1230 bases de données ou collections dans le domaine de la biologie cellulaire et moléculaire. Cette dispersion pose de façon aiguë la question de la récupération des données pertinentes pour les chercheurs. La question du choix entre la poursuite vers un modèle centralisé (type GenBank par exemple) ou distribué (qui est en train de se mettre en place) est inévitable.

2.6. DES SYNERGIES À RENFORCER POUR LA GESTION ET LA VALORISATION DES DONNÉES

De nouvelles synergies entre différentes compétences sont nécessaires pour la gestion et la valorisation des données de recherche. Le rapport de Swan et Brown (2008) mentionne quatre types de profils ou compétences :

- Créateur de données (*data creator*): chercheurs ayant un domaine d'expertise conduisant à la production des données.
- Modélisateur (statisticien, bioinformaticien, etc.) (*data scientist*) : chercheurs impliqués dans le traitement et l'analyse des données, ainsi que dans la modélisation conceptuelle des bases de données.
- Gestionnaire de données (*data manager*) : ingénieurs en charge du stockage, de l'archivage et, plus généralement, de la préservation des données.
- Curateur de métadonnées (*data librarian*) : personnes venant du monde de la documentation en charge de la qualité des métadonnées, de la maintenance des référentiels.

Ces deux rôles *data manager* et *data librarian* sont à rapprocher du rôle de *data curator*[23] (curateur), nouveau métier du web aux définitions multiples. Leur mission est de mettre en relation des données non seulement avec l'amont (métadonnées sur le recueil et le contexte d'acquisition) mais aussi avec l'aval (mise en relation avec des champs de connaissance, des communautés de chercheurs, des attentes sociétales, etc.)[24].

Au Royaume Uni, dans le cadre du programme RDM (Research Data Management) du JISC (Joint Information Systems Committee, *cf.* Annexe A.1.4.), des formations ou des outils d'autoformation sont proposés aux chercheurs et aux étudiants comme par exemple à l'université d'Edinburg avec le projet MANTRA[25]. On peut aussi trouver de nombreux exemples de formations diplômantes aux États-Unis comme par exemple à l'Université d'Illinois[26]. En Europe, la France est en retard sur ces questions.

En Australie, l'exemple du dispositif mis en place à la Monash University[27] est

[23] http://pro.01net.com/editorial/529624/le-guide-de-la-curation-%281%29-les-concepts/ : « La pratique qui consiste à sélectionner, éditorialiser et partager du contenu a été baptisée par les Américains *curation* ou *content curation*, par analogie avec la mission du *curator*, le commissaire d'exposition chargé de sélectionner des œuvres d'art et de les mettre en valeur pour une exposition. »

[24] Une figure illustrant l'organisation de ces nouvelles compétences est disponible à l'adresse http://data-forum.blogspot.com/2008/12/rdmf2-core-skills-diagram.html

[25] http://datalib.edina.ac.uk/mantra/

[26] http://www.lis.illinois.edu/academics/programs/ms/data_curation

[27] http://www.jisc.ac.uk/media/documents/programmes/mrd/28and29March/

particulièrement intéressant car il illustre la place des compétences en information scientifique et technique, ceci dans un projet d'établissement dédié à la gestion des données de la recherche.

La bibliothèque est plus particulièrement impliquée dans :

- L'élaboration de la stratégie, de la politique et de la gouvernance.
- L'information, la communication et les conseils, la promotion des bonnes pratiques.
- La formation et le développement des compétences pour les scientifiques et les professionnels en information scientifique et technique.
- La valorisation : diffusion des métadonnées, promotion des ressources produites, acquisition d'autres ressources, maintenance des référentiels, gestion des identifiants.
- L'appui aux scientifiques pour l'élaboration des plans de gestion des données.
- Le développement des partenariats dans ce domaine.

Au niveau d'un établissement, il semble nécessaire de créer les conditions politiques et organisationnelles pour une synergie entre les différents corps de métier.

2.7. EXTERNALISATION DE LA PRODUCTION DES DONNÉES HORS PLATEFORMES

Certains domaines comme le séquençage font de plus en plus appel à des sociétés privées dans la production de données. Les externalisations de ce type permettent d'obtenir des données de qualité industrielle, d'avoir accès à des technologies non disponibles localement et libèrent les personnels d'un certain nombre de tâches routinières et consommatrices de temps. Ces avantages sont contrebalancés par le fait que les utilisateurs deviennent « captifs » d'une certaine offre technique (*cf.* section 3.).

2.8. PUBLICATION DES DONNÉES ET REVUES CIBLÉES

En février 2010, l'INIST (Institut de l'Information Scientifique et Technique du CNRS) a officiellement rejoint le consortium DataCite, en tant que représentant français. DataCite est issu d'une initiative de plusieurs grandes bibliothèques scientifiques européennes, visant à améliorer l'accès aux données de la recherche, en proposant de nouveaux services. Le consortium comprend aujourd'hui 12 membres, à travers le monde. Il a un statut d'agence DOI (Digital Object Identifier) dont les membres sont, à ce titre, habilités à attribuer des identifiants à diverses ressources numériques, rendant ainsi ces ressources citables et plus facilement accessibles. Dans cette mission, DataCite devrait collaborer étroitement avec des centres producteurs de données et avec des éditeurs scientifiques. Pour accompagner ces collaborations et pour orienter et harmoniser ses activités, le consortium a mis en place plusieurs groupes de travail (métadonnées, objets dynamiques, préservation, nouvelles compétences requises, etc.).

En 2011, la revue GigaScience[28], née d'une collaboration entre le BGI (Beijing

04A_Searle_Research_Data_Planning.pdf

[28] http://www.gigasciencejournal.com/

Genomics Institute) et BMC (BioMedCentral) vient révolutionner le monde des données. La particularité de cette revue – en dehors de publier des articles décrivant des études menées sur de grands jeux de données eux-mêmes référencés par un DOI et rendus disponibles *via* une base de données hébergée au BGI – est de proposer une organisation qui permet de partager non seulement les résultats d'études mais aussi les données et les outils qui ont conduit à ces résultats. Pour atteindre ces objectifs, le BGI met à disposition une base de données et un *cloud*. Ce journal couvre non seulement les données omiques mais aussi tout type de données issues de la biologie à grande échelle et incluant par exemple les domaines de l'imagerie, des neurosciences, de l'écologie, des cohortes, etc. D'autres journaux, tels que Datasets International[29], suivent ce mouvement en proposant, à côté de la publication des articles, la prise en charge des grands jeux de données associés. A ce jour, il reste difficile de prévoir quelles seront, à moyen terme, les évolutions des conditions de prise en charge des données.

3. SITUATION À L'INRA

3.1. NATURE ET DYNAMIQUE DE PRODUCTION DES DONNÉES À L'INRA

Les caractéristiques des données à l'Inra sont leur diversité, leur hétérogénéité, leur production éclatée géographiquement, et le caractère privé (cas des données d'élevage, *cf.* Annexe A.4. du rapport), ou confidentiel (cas des cohortes), de certaines d'entre elles. On observe également dans le secteur de la biologie un problème relatif à la qualité des données produites (variabilité des protocoles employés, manque de standardisation, absence de certification ISO ou de label qualité, etc.). Les chercheurs en général ont rarement les compétences nécessaires pour bien archiver et gérer leurs données[30] car leur activité est concentrée sur la production de publications. Dans le cadre de l'analyse de la situation à l'Inra, les chercheurs et les ingénieurs de l'Institut ont été interrogés par le biais de plusieurs enquêtes. Dans le cadre de ces enquêtes, les plateformes et les CATI nous ont paru constituer des lieux propices à l'évaluation de l'existant en matière de gestion et de partage de données. En effet, les CATI, tels qu'ils ont été mis en place à l'INRA, regroupent thématiquement et souvent par grand domaine, la majorité des ingénieurs de l'Institut qui sont impliqués dans les projets scientifiques. Ils sont en charge de (i) l'accès à des données extérieures nécessaires à la réalisation d'un projet ; (ii) la gestion et l'analyse des données produites dans le cadre d'un projet (et bien souvent au-delà de la durée du projet) ; mais aussi de (iii) la mise à disposition de ces données dans des banques internationales dans un but de partage avec la communauté (mise à disposition imposée par la revue où sont publiés les résultats en lien avec l'analyse des données) ; ou (iv) leur valorisation par la mise à disposition de la communauté (base de données, service web, etc.) en utilisant des moyens propres à l'équipe ou au laboratoire.

Dispersion des dispositifs

Au contraire des données de la physique des particules, qui sont homogènes et produites

[29] http://www.datasets.com/journals/

[30] Par exemple, une enquête réalisée au Royaume-Uni (Lyon, 2007) a montré que : « Bien qu'il y ait des écarts entre les disciplines, la compétence des chercheurs en matière de gestion des données est en général mauvaise ».

de façon centralisée, les données issues de la biologie sont très hétérogènes et éclatées sur différents sites de production. La très grande variabilité des types et formats de données dans l'ensemble des disciplines ajoute à leur complexité. Enfin, l'intensification des collaborations nationales et internationales entre communautés différentes, issues d'organismes également différents, rend particulièrement sensible la question de l'interopérabilité entre les systèmes d'information portant les données.

Résultats des enquêtes concernant la dynamique de production

Afin de connaître à l'Inra la nature et la dynamique de production des données des différentes communautés (omiques, écologiques, sciences humaines et sociales) une enquête sur les modalités de production des données a été conduite à l'automne 2011. Les informations ont été recueillies par l'intermédiaire d'un formulaire accessible sur Internet. Le contenu de cette enquête figure dans l'Annexe du rapport. Ce formulaire comprenait 14 champs correspondant à des questions majoritairement à réponse fermée et permettant de caractériser en nature, volume, discipline et perspective d'évolution, les jeux de données produits par l'Inra. Trois groupes de personnes ont été sollicités : les responsables de CATI, les directeurs des unités expérimentales et des unités de services et les responsables de plateformes. Des entretiens ont été conduits pour compléter les résultats des enquêtes.

Sans être exhaustive cette enquête permet d'identifier et de caractériser les lieux de production, la nature des données et leur mode de gestion. Elle ne permet pas d'estimer le volume global des données produites en raison du trop faible nombre de réponses. Le volume total correspondant aux 111 réponses reçues est de 51 Téraoctets par an. L'ordre de grandeur du volume de données produites annuellement par l'Inra, en 2011, serait donc de quelques centaines de Téraoctets.

Les producteurs de données identifiés sont majoritairement les unités puis les plateformes. A titre d'exemple, un laboratoire comme le LIPM (Laboratoire des Interactions Plantes-Microorganismes), à Toulouse, cumule sur trois années un volume de 2,5 Téraoctets (données compressées), avec une moyenne de 162 Gigaoctets par mois en 2011. La plateforme bioinformatique de Toulouse, qui stocke les données de la plateforme de séquençage, cumule près de 11 Téraoctets de données dont 10 générés en 2011. La dynamique d'évolution de la production de séquences y est difficilement prévisible et planifiable sur le long terme tellement les technologies évoluent vite (production estimée à plus de 2 Téraoctets/mois d'ici fin 2012). Enfin dans un contexte d'évolution moins rapide, et dans une dynamique encore loin de la démocratisation, on peut aussi observer des productions massives de données dans le domaine de la protéomique. Par exemple, dans le cas de la plateforme GenoToul, les données produites en spectrométrie de masse étaient de l'ordre de 27 Téraoctets en 2011, avec 120 Téraoctets prévus pour 2014.

Peu de jeux de données restent encore sous forme papier (<3 %), la plupart étant informatisés et gérés dans des fichiers (55 %) ou des bases de données (25 %), et 10 % en fichier SIG.

En matière de domaine d'activité, un tiers des réponses (34/111) relève de l'environnement. Un deuxième tiers porte sur le phénotypage animal ou végétal. Seules huit réponses concernent le génotypage. Toutefois, en volume de données produites, celui-ci est prépondérant (34 %). D'autre part, l'ensemble des activités omiques

représentent 66 % du volume total (dont 50 % pour les séquences génomiques), le phénotypage 22 %, et l'environnement 11 %.

Les réponses aux questionnaires proviennent de 19 centres différents. Malgré l'hétérogénéité des taux de réponses, la production en volume ne semble pas concentrée sur quelques sites mais se répartirait de façon relativement équilibrée sur la moitié d'entre eux. Les centres à forte activité omique se placent en tête. Ainsi Jouy-en-Josas et Toulouse représentent (en cumulé) plus du tiers, sept autres centres (Avignon, Dijon, Nantes, Rennes, Versailles) représentent de 2 à 10 % chacun. Les autres (Angers, Bordeaux, Clermont-Ferrand, Nancy, Orléans) n'apparaissent que très peu en volume.

Les accroissements prévus pour les trois à cinq ans à venir sont fréquemment estimés comme étant d'un ordre de grandeur. Les domaines d'activité génotypage et phénotypage sont les premiers concernés, avec parfois un facteur 100 pour des activités omiques. En matière de phénotypage, la généralisation du recours à l'imagerie, et encore plus aux vidéos, devraient conduire à une augmentation très sensible des volumétries.

Les tendances affichées dans le cadre de cette enquête sont conformes aux tendances internationales observées dans la littérature. Le phénotypage commence à se révéler un producteur important de données, mais l'écologie et l'environnement sont aussi des domaines émergents.

Etude du centre de Jouy-en-Josas

Dans le but d'avoir une photographie instantanée d'un site pilote, nous avons conduit une enquête avec les chercheurs et directeurs d'unités sur le centre de Jouy-en-Josas. Cette enquête a été réalisée sous la forme d'entretiens. Elle a permis de mettre en évidence que, même dans les petites unités, la diversité des techniques mises en œuvre, et par conséquent des types de données qui sont produites ou manipulées, est extrême.

Le premier résultat de cette enquête montre que les unités sont responsables du stockage des données qu'elles produisent et en supportent les coûts. L'espace alloué par la DSI (Direction des Services Informatiques) à chaque utilisateur n'est pas destiné à cet usage et n'est pas dimensionné pour cela. Les données qui sont produites par un équipement commun au sein d'une plateforme sont stockées par celle-ci et sous sa responsabilité. En revanche, les données qui, après un traitement de premier niveau, sont issues de ces données brutes, ne sont en général pas maintenues par la plateforme. Leur stockage est donc sous la responsabilité de l'unité et le coût correspondant lui est imputé.

Actuellement les unités font face avec des solutions trouvées souvent en urgence et qui ne supporteront pas une augmentation importante des flux de production. De plus, ces solutions (achat de serveurs, de mémoire vive, de disques durs) sont financés par des réponses à des appels d'offre qui couvrent le coût d'un équipement par petites tranches successives. Cette stratégie ne favorise pas l'émergence d'un ensemble cohérent, ne finance pas son fonctionnement, et ne permet pas sa pérennisation. Face à ce problème, certaines unités ont pour politique − dans le cadre de leurs projets financés en réponse à des appels d'offre − de demander de l'équipement destiné à équiper la plateforme avec laquelle elles travaillent et auquel elles ont, en retour, un accès réservé ou privilégié.

La situation est différente pour les unités qui sont impliquées dans des projets qui reposent essentiellement sur une production massive de données. La gestion et le traitement de ces données sont d'emblée perçues comme une partie centrale et stratégique du projet. Des solutions propres à chaque projet sont recherchées, le problème est qu'elles ne possèdent pas de caractère mutualisé. Elles ne sont pas toujours dimensionnées à une hauteur suffisante, en particulier pour l'analyse des données.

Les services rendus par les plateformes en termes de calcul ne sont pas payants. Il n'est pas garanti que cette situation perdure, une tendance à l'autofinancement des plateformes se faisant jour. Les services rendus deviendraient payants, avec une tarification différenciée selon l'origine du demandeur. L'essentiel du coût des calculs est en réalité représenté par du temps de travail ingénieur et chercheur. Ce travail concerne la maintenance et la gestion des données et de l'environnement qui les supporte (administration, maintenance des bases de données et des logiciels) et surtout leur analyse. Les données n'acquièrent de valeur que si elles ont été traitées et validées et seulement si elles sont stockées de façon structurée et maintenues sur des supports qui en garantissent l'accessibilité dans la durée.

La nécessité de la mutualisation est évidente, non seulement par les économies d'échelle qu'elle permet, mais surtout comme unique moyen de regrouper les compétences nécessaires et de constituer des ensembles de masse critique suffisante. L'Inra a fortement favorisé la constitution de plateformes et s'est doté d'instruments d'arbitrage et de décision.

Les questions de propriété intellectuelle

Comme nous l'avons vu précédemment (paragraphe 1.3.), les données communicables par l'Inra sont pour l'essentiel ses bases de données, ses logiciels et les résultats protégés au titre de la propriété industrielle. C'est sur ces sujets qu'il apparaît nécessaire de procéder à un approfondissement quant aux règles de communicabilité.

En tant qu'établissement public de recherche, l'Inra peut fixer librement les conditions de réutilisation des dites données (à la différence d'organismes publics non liés à la recherche). Pour les résultats valorisables économiquement, l'Inra a mis en place une politique active présidant aux transferts de technologies. Pour les données à caractère non commercial, une politique de réutilisation spécifique pourrait (aussi) être mise en place.

Il n'existe pas, pour l'instant, de politique officielle concernant la réutilisation des documents administratifs et les pratiques empiriques prévalent. Ce problème est aggravé par la méconnaissance qu'ont les chercheurs du cadre juridique concerné et des obligations qui s'y rattachent, ceci tout particulièrement dans le cas de documents multipropriétaires.

Sécurisation des données

L'audit d'Ernst & Young sur la sécurité a mis en évidence à d'importantes lacunes dans l'infrastructure de collecte et de préservation des données de recherche. Une grande ligne directrice est qu'il n'existe pas de site de repli, ceci y compris pour des données d'importance stratégique. Bien souvent, les dispositifs sont localisés sur le même site que celui de production et de stockage des données. Le processus de sauvegarde

représente une vraie problématique, en effet plus de 40 % des serveurs ne sont pas sauvegardés, et les postes de travail sont rarement sauvegardés automatiquement. La mise en œuvre de tels dispositifs s'avère à la fois complexe et délicate. Dans ce dernier domaine, les pratiques individuelles prévalent donc. Par ailleurs, les supports utilisés ont une durée de vie limitée, chose dont les utilisateurs n'ont pas forcément conscience.

Si les plateformes ont été soumises à des audits afin d'avoir le label qualité, par contre, nous avons relevé que rien n'a été effectué dans le domaine de la sécurité. Il n'existe à notre connaissance aucune PSSI (Politique des Services de Sécurité Informatique) au niveau national pour l'Inra, ceci alors que d'autres institutions ont mis en place des groupes de travail sur le sujet (ENS, CNRS, INRIA). Dans ces structures, les responsables informatiques ont reçu des recommandations quant à la nécessité de mettre en place de telles politiques.

Le maillon faible qui est apparu concernant la sécurisation des applications se situe au niveau du recours de plus en plus massif à des CDD de courte durée (de six mois à un an) pour effectuer les développements. En effet, du fait de la durée limitée des missions qui leur sont confiées, il n'est que très rarement possible de former ces personnes aux bonnes pratiques en matière de sécurité informatique. Il est également fréquent que des identifiants correspondant à du personnel temporaire soient longuement conservés pour des raisons pratiques.

Il existe plusieurs exemples d'attaque en règle de serveurs Inra par des *hackers*. Aucun système informatique ne peut être sécurisé à 100 %, ce problème étant exacerbé par l'hétérogénéité matérielle et logicielle généralement présente sur un site.

3.2. ACCÈS AUX DONNÉES PRODUITES HORS INRA

Les travaux de recherche conduits à l'Inra s'appuient fortement sur des données produites en dehors de l'institut. Tous les domaines de recherche sont plus ou moins concernés : (i) l'écologie avec les données environnementales sur le milieu physique, le climat, la végétation ; (ii) la génétique, notamment animale avec les données de performance collectées par l'interprofession (*cf.* Annexe A.4. du rapport) ; (iii) la génomique, avec les séquences produites et mises à disposition au niveau international ; (iv) les sciences humaines et sociales, avec des données issues du ministère et des structures agricoles.

Origine des données

Le panorama des « fournisseurs » et les types de données sont ainsi très variés. L'accessibilité va de la gratuité complète (*e.g.*, données de séquences publiques, données climatiques *via* Agroclim, certaines données IGN), à des solutions payantes. C'est le cas des données satellitaires qui vont, de surcroît, générer des coûts de stockage croissants du fait des gains en résolution. C'est la raison pour laquelle, à l'initiative de collectivités territoriales, se créent des fédérations régionales pour l'acquisition et la mutualisation des données concernant les territoires (*e.g.*, CRIGe PACA[31]), ou des instituts de recherche (*e.g.*, Equipex GeoSud ; *cf.* Annexe A.1.6.).

[31] http://www.crige-paca.org/

Résultats des enquêtes

Une enquête sur l'origine et la nature des données produites à l'extérieur de l'Inra et utilisées par l'Institut a donc été conduite à l'automne 2011. Les informations ont été recueillies par l'intermédiaire d'un formulaire accessible par Internet. Le contenu de l'enquête est disponible dans l'Annexe A.2. du rapport. Ce formulaire comprenait 16 champs correspondant à des questions majoritairement à réponse fermée et permettant de caractériser en origine, nature, volume, discipline et perspective d'évolution des jeux de données acquis par l'Inra. Les personnes sollicitées sont identiques à celles listées dans la section 3.3.

Du fait du faible nombre de réponses (57), les résultats obtenus ne sauraient être totalement représentatifs de l'ensemble des données externes utilisées à l'Inra. Ces 57 réponses ont été apportées par 19 personnes provenant de 13 centres ou implantations Inra. Les jeux de données sont acquis et utilisés par 19 structures, majoritairement des unités (10), puis des départements (5), et enfin des équipes (2) et centres (2).

Le volume total de données (20 Téraoctets pour 48 des 57 réponses) est très inégalement réparti entre les différents centres Inra, celui de la région PACA en représentant la moitié. Sur ce centre, l'Unité EMMAH (Environnement Méditerranéen et Modélisation des Agro-Hydrosystèmes) apporte à elle seule 8,5 Téraoctets. Le centre de Nancy vient en seconde position avec 4 Téraoctets, les autres centres ne dépassent pas 1 Téraoctet. Près des deux tiers des réponses (34/57) correspondent à des jeux de données sur l'environnement. Environ 60 % du tiers restant relèvent de données de génotypage. Par contre, en volume, les données relatives à l'environnement représentent 90 % des 20 Téraoctets, la deuxième place étant occupée par le génotypage (8 %).

Les données proviennent de très nombreux partenaires (37 pour les 57 réponses). L'IGN apparaît, dans cette enquête, comme premier fournisseur (8 réponses). Viennent ensuite des fournisseurs d'images satellites et de données climatiques ainsi que le ministère de l'Agriculture et les organismes agricoles. Vingt-huit fournisseurs ne sont cités qu'une fois. Les jeux de données en provenance de l'IGN représentent 44 % du volume (8,5 Téraoctets). Les données d'images satellites déclarées par l'Unité EMMAH proviennent de sept fournisseurs différents (Spot Image, Eurimage, MODIS, VITO, POSTEL, GEOLAND, LTDR) et atteignent le même volume. Les données de génotypage proviennent essentiellement de trois partenaires : BGI (Beijing Genomics Institute), CEA et BCM (Baylor College of Medicine). La moitié des jeux de données sont des images, 23 % des séquences génomiques, et 19 % des données alphanumériques. Les volumes respectifs correspondants sont de 89 %, 11 % et 0,4 %.

L'évolution pressentie à 3-5 ans est soit stable (42 % des réponses), soit en augmentation d'un ordre de grandeur (23 % des réponses).

3.3. ENQUÊTES SUR LA SITUATION DES PERSONNELS INRA

Dans le but d'établir un état des lieux de la perception des changements en cours qu'ont les personnels Inra, deux enquêtes (« ingénieurs » et « chercheurs ») ont été mises en ligne courant décembre 2011. Les objectifs de ces enquêtes étaient de :

- Déterminer l'évolution du contenu de l'activité professionnelle des personnes contactées.

- Lister les changements perçus dans les modes de fonctionnement et de collaboration.
- Déterminer quels étaient les moyens de calculs requis par les changements de pratique et la massification des données.
- D'estimer la fréquence du recours à des structures de type plateforme et centres de calcul.

Le questionnaire a été envoyé à un sous-ensemble de 985 ingénieurs et de 1371 chercheurs que le groupe estimait être concerné par la question.

Le contenu des deux enquêtes est donné dans plusieurs annexes de ce rapport (A.2.3. et A.2.4.). A la différence de celles décrites dans les sections 3.3. et 3.4., elles contenaient plusieurs questions à réponse ouverte, ce qui a rendu le dépouillement complexe. Du fait de cette complexité, seuls les résultats les plus saillants sont présentés dans ce rapport. Une synthèse générale s'avère également difficile, du fait du caractère fréquemment contradictoire des réponses apportées aux questions ouvertes. Un certain nombre de commentaires nous ont toutefois paru suffisamment signifiants pour être évoqués. Cette sélection résulte d'un choix subjectif, et les personnes intéressées par un examen plus exhaustif peuvent se reporter au dépouillement complet de l'enquête, disponible sur le site du conseil scientifique.[32]

Par ailleurs, la quasi-totalité des ingénieurs et chercheurs qui ont été échantillonnés dans le domaine des SHS n'ont pas répondu au questionnaire ou ont répondu qu'ils ne se sentaient pas concernés par la production de données à haut débit. Pour rendre compte des enjeux autour de la gestion et du partage de ces données, quatre entretiens téléphoniques ont été conduits auprès d'ingénieurs ou de chercheurs de plateformes SHS.

Enquête ingénieurs

Le pourcentage de participation a été important puisque 31 % (soit 307 personnes sur 985) des ingénieurs consultés ont répondu. Par ailleurs, 44 d'entre eux ont rédigé des commentaires libres. Une première constatation est que la conscience de l'émergence du haut débit est inégalement répartie en fonction de la structure d'appartenance. Les ingénieurs des plateformes sont en proportion (77 %) légèrement plus concernés que ceux rattachés aux UR/UMR (55,3 %) et aux équipes (60,1 %). Les ingénieurs dans les unités expérimentales se sentant peu concernés (29 %).

Un sentiment partagé par plusieurs ingénieurs est la nécessité qu'il faut désormais faire beaucoup plus avec autant – si ce n'est moins – de moyens. Du fait de la multiplication des projets à traiter, ainsi que de leur plus grande variété, on constate une dispersion plus importante des tâches. Les ingénieurs des plateformes et des plateaux techniques, c'est-à-dire des structures les plus collectives, observent une tendance au raccourcissement des délais impartis. La complexification des tâches à accomplir s'accompagne en même temps d'une intensification du travail administratif à fournir (*e.g.*, contrôle qualité, labellisations). Enfin, depuis l'apparition du haut débit, 52 % des ingénieurs ont vu leur temps consacré à la veille technologique augmenter. Toutes ces nouvelles contraintes pourraient avoir un effet sur la qualité du travail produit, et donc de la recherche scientifique correspondante.

[32] https://intranet.inra.fr/conseil_scientifique/rapports_du_conseil

Les catégories les plus concernées par l'avènement des méthodes haut débit se situent principalement en bioinformatique et en biostatistique. Ces deux catégories se distinguent cependant au niveau de l'implication des personnels dans le processus décisionnel. En effet, si seulement 45 % des biostatisticiens interrogés travaillent sur des projets sur lesquels ils n'ont eu que peu de prise au cours de l'élaboration, cette proportion monte à 80 % pour les ingénieurs bioinformaticiens. Cette différence résulte probablement de l'historique des disciplines correspondantes, l'informatique étant encore largement perçue comme un domaine essentiellement technique.

Le mode de collaboration prédominant reste encore l'intégration des ingénieurs dans des projets de recherche. Dans les UR/UMR et équipes, près de 60 % des ingénieurs concernés sont sur ce mode de collaboration, quand seulement 20 % réalisent des prestations de service et 20 % ont une participation seulement technique. La tendance est différente pour les ingénieurs de plateformes, UE et plateaux techniques, puisqu'ils sont intégrés dans des projets de recherche ou participent à des prestations de service dans des proportions équivalentes comprises entre 33 et 44%.

Concernant la nature des données haut débit, celles-ci se partagent essentiellement entre omique (61.1 %), phénotypage (36.6 %) et imagerie (18.9 %) (total > 100% car réponses multiples possibles). Ces proportions sont cohérentes avec celles observées dans les enquêtes présentées aux sections 3.3. et 3.4. Le faible nombre de réponses dans d'autres domaines (*e.g.*, données climatiques ou satellitaires) ne permet pas de tirer des conclusions quant à une éventuelle tendance.

Une grande majorité (80 %) des personnes disant utiliser des moyens de calcul se déclarent satisfaites. Les ressources utilisées proviennent essentiellement des mésocentres membres du ReNaBi[33] (Réseau National des plateformes en Bioinformatique, *cf.* Annexe A.1.1) (52,1 %). Par contre, seulement 6,8 % des personnes ayant répondu font appel à des centres spécialisés pourvoyeurs de très gros moyens de calculs (CCRT, IDRIS, CINES, CC-IN2P3). Les bioinformaticiens constituent la majorité des utilisateurs de ces moyens. Parmi les 20 % de personnes non satisfaites, on constate une grande variété dans les propositions de solutions à apporter (*e.g.*, création d'un centre national Inra, mise en place de moyens locaux).

Pour ce qui est des projections à cinq ans, les changements les plus cités concernent l'augmentation des volumes de données (25,1 %), le développement des techniques et des méthodes de traitement (21,7 %), les besoins en stockage, en temps et en puissance de calcul (13,7 %), l'échange de données entre communautés (10,9 %). Les enjeux sur la propriété et le partage des données ne sont cités que dans 1,1 % des réponses ! Les bioinformaticiens restent les personnes les plus concernées par les questions de stockage et de calcul (problèmes soulevés par 22 % des ingénieurs rattachés à cette catégorie). Cette proportion est à comparer avec celle, beaucoup plus faible, observée chez les biologistes (8 %).

Une autre tendance qui ressort est un renforcement de la spécialisation des ingénieurs. Les ingénieurs spécialisés participent de plus en plus à des prestations de service ou à des missions techniques. Par ailleurs, dans les équipes, les projets de recherche qui intègrent des ingénieurs sont en hausse.

[33] http://www.renabi.fr/

Enquête chercheurs

Le pourcentage de participation s'est également révélé important, avec 29 % des chercheurs (soit 397 sur 1371). De même, 76 chercheurs ont développé des commentaires libres.

Un résultat global est que les chercheurs les plus concernés par l'avènement des techniques à haut débit appartiennent aux domaines de la bioinformatique et des biostatistiques. Il est d'ailleurs notable que les demandes de formation exprimées par les biologistes se situent principalement dans ces deux domaines. Les types de données concernées relèvent en priorité du domaine des omiques (80 %), du phénotypage (37.7 %) et de l'imagerie (26,8 %) (total > 100 % car réponses multiples possibles). Les changements ressentis sont récents, avec un pic en 2008. Ce résultat n'est pas surprenant, et il correspond effectivement à la date d'apparition de ces technologies haut débit, tout du moins dans le domaine du séquençage.

Seulement un tiers des chercheurs ont constaté des changements dans le nombre de projets à traiter, ce changement étant très majoritairement perçu (90 %) comme ayant une tendance à l'augmentation. Les raisons avancées à cette augmentation sont variées, l'accessibilité à de nouvelles technologies n'étant pas la seule en cause (*e.g.*, multiplicité des appels d'offres, management par projet).

Près de 40 % des chercheurs concernés par les données à haut débit font appel à des moyens de calcul. Cette proportion est similaire à celle observée chez les ingénieurs (42 %). Encore une fois, les personnes faisant le plus appel à ces moyens sont les bioinformaticiens et les biostatisticiens. Tout comme dans le cas des ingénieurs, les moyens les plus couramment utilisés sont ceux des mésocentres ReNaBi (38 %). Ceci n'est pas surprenant puisqu'un des rôles assignés à ces centres était de fournir des moyens de calcul à la communauté des biologistes. Là encore, le taux de satisfaction est particulièrement élevé (74 %). Pour les personnes non-satisfaites, les solutions à apporter sont du même ordre que celles proposées par les ingénieurs.

Concernant les plateformes, il apparaît que celles-ci sont largement utilisées puisque 58,3 % des chercheurs ayant répondu les ont utilisées. Le questionnaire envoyé ne permettait pas de déterminer s'il existait une tendance à la croissance dans l'utilisation des plateformes, mais il s'agit là d'un élément qui pourrait faire l'objet d'un suivi par l'Inra. Au sujet de la question des changements observés dans la façon d'élaborer les projets, les principaux points positifs relevés sont la possibilité de monter des projets de plus grande envergure, l'accès à des moyens techniques et méthodologiques appropriés et la présence d'interlocuteurs compétents et centralisés. Parmi les points négatifs, le principal problème semble être l'introduction de contraintes supplémentaires (réactivité moindre : disponibilité des personnes, prise de décision, etc.) dans le montage et la réalisation des projets.

De manière attendue, concernant les solutions que l'Inra pourrait apporter, les suggestions faites se situent essentiellement au niveau de l'augmentation des moyens (humains et matériels). Les projections à cinq ans faites par les chercheurs sont proches de celles des ingénieurs (augmentation des volumes de données, besoins en termes de stockage et de puissance calcul, etc.). Quelques points spécifiques aux chercheurs sont la nécessité accrue de s'approprier les outils permettant de traiter les données (en particulier au moyen de formations) et la crainte d'une valorisation insuffisante. La

question de l'interdisciplinarité, même si elle n'est pas nouvelle, apparaît également comme devenant cruciale.

Entretiens avec les responsables de plateformes SHS

A l'INRA, de nombreux chercheurs SHS ne se sont pas sentis directement concernés par le haut débit. Il nous a paru cependant important d'enquêter sur cette thématique pour laquelle le haut débit apparaît en émergence (King, 2011). Cinq entretiens ont été conduits. Les quelques réponses reçues font ressortir que le problème n'est pas lié à l'accumulation ou à la fugacité des données, mais à leur accès et leur confidentialité.

- Entretiens avec des responsables de plateformes SHS

Quatre entretiens ont concerné des responsables de plateformes SHS. Ces personnes ont fait part d'une évolution dans la nature ou l'accès aux données. Les deux personnes interviewées travaillant sur des plateformes dédiées à l'analyse des politiques publiques françaises n'ont pas ressenti un changement dans la nature de leur travail, mais bien dans l'accès aux données : l'avènement dans le cadre de la directive Inspire de la démarche *Open Data* permet un accès facilité aux données administratives. Les deux personnes travaillant sur de gros corpus de données moissonnées sur le web font part d'évolutions très sensibles, dans la nature et le volume des données, tout comme la nature des partenariats et le modèle économique dans lequel s'exerce la recherche.

Ainsi le travail, qui se faisait il y a quelques années sur de grosses bases de données, accessibles dans la durée, a dû s'adapter pour cueillir un flux continu et fugace d'informations sur des réseaux sociaux, blogs scientifiques, ou *tweets*, qui doivent être collectées, analysées et archivées en temps réel. Parallèlement à cette évolution, le standard international, qui était en 2006 l'analyse de quelques centaines de blogs durant quelques semaines, soit une masse d'information de quelques Gigaoctets, nécessite aujourd'hui l'analyse d'un corpus de données de l'ordre du Téraoctet. Le foisonnement des entreprises du web s'est accompagné du développement d'entreprises dont le seul objet est de collecter des données web dans un format standardisé pour les revendre (Spin3r[34], Linkfluence[35]). Il existe un véritable enjeu sur la qualité des données en question.

L'accès aux données, souvent propriété des entreprises du web (Yahoo, Google, Facebook, etc.), est de moins en moins gratuit (Twitter par exemple vend ses données depuis cette année), et constituent un véritable enjeu pour les communautés de recherche. Les partenariats avec les propriétaires des données se développent (exemple de Google Ngram Viewer[36], issu d'une collaboration entre Google et des chercheurs de Harvard, du MIT, de l'Encyclopeadia Britannica et de l'éditeur Houghton Mifflin Harcourt ; Michel *et al.*, 2010).

Les aspects logistiques sont perçus comme réels, mais ne constituent pas le problème majeur. Les quatre personnes se disent contraintes par les ressources humaines plus que par les ressources physiques. Ainsi, le stockage n'est pas le principal problème. Comme dans les autres domaines, l'organisation et le traitement des données textuelles avant de les stocker est un enjeu, en termes de débit et de puissance de calcul,

[34] http://spinn3r.com/
[35] http://fr.linkfluence.net/
[36] http://books.google.com/ngrams

mais des solutions concrètes sont en cours de mise en œuvre. Le principal défi est l'organisation d'interfaces pour rendre les données accessibles aux chercheurs. Les besoins accrus en personnel sont perçus comme déterminants. Les besoins en ingénierie sont en particulier nécessaires pour concevoir des interfaces facilitant l'utilisation des outils, et la visualisation des résultats.

Des questions éthiques et juridiques émergent. Un chercheur a mentionné les questions éthiques posées par la moisson, à l'insu des personnes, des informations les concernant sur internet, tout en soulignant que le problème n'était pas particulier à la recherche, et que les gros corpus étaient stockés sous forme de contenu indexé et non de données personnelles. Un autre chercheur a évoqué la question de la confidentialité des données personnelles collectées par la recherche, et a souligné le besoin d'une sensibilisation des chercheurs à cet aspect. Deux personnes ont évoqué le besoin d'une meilleure information sur des aspects juridiques. Cela concerne par exemple le stockage temporaire des données de sources variées sur un serveur pour produire un produit de recherche. Se pose la question du cadre juridique permettant d'enregistrer et de conserver ces données, et avec quelles limites.

Des enjeux de formation sont perçus dans certains domaines comme déterminants pour le maintien de la production scientifique. L'évolution toujours plus rapide des technologies et des applications internet exige une mise à niveau constante sur les services et les outils de développement. Les formations professionnelles dans ces domaines ne sont pas disponibles dans le temps de la recherche, qui nécessite une très forte réactivité. De ce fait, le personnel concerné a beaucoup recours à l'autoformation ou la formation/information par les pairs. Confrontés à cette obsolescence très rapide des compétences, les grands laboratoires de recherche comme le MIT qui ont un réseau de chercheurs plus important, disposent d'un avantage comparatif à la fois en termes de formation et de capacité à produire.

Enfin, la recherche fait face à de nouveaux enjeux. La mutation dans l'analyse des données est un point critique pour les SHS ; la réussir est un enjeu crucial d'instrumentation, mais également épistémique qui marque une mutation par rapport à la sociologie traditionnelle. Des champs entiers de sociologie ne seront plus accessibles si on ne la réussit pas. Par exemple l'analyse des controverses nécessite de s'équiper de moyens de suivi des controverses *in vivo* mais aussi des moyens de traiter / analyser / naviguer dans ces données. Un autre enjeu est celui de la co-construction : il est nécessaire, pour un travail de qualité, que le sociologue comprenne les hypothèses de travail qui vont fonder la collecte et l'analyse des données. Il n'est pas possible de situer les plateformes dans une pure optique de prestation de service. Le pari est que cette co-construction produira un ensemble cohérent de données pour la problématique envisagée, ce qui devrait aboutir *in fine à* une analyse plus riche et plus pertinente, et permettre une recherche de meilleure qualité.

- La situation des autres chercheurs en SHS :

Deux éclairages ont été recherchés : les enjeux concernant la protection des données individuelles, l'accessibilité, et les barrières émergentes concernant le partage des données.

Le recours à des données individuelles (individus, ménages, entreprises, etc.) est en augmentation forte dans les travaux des chercheurs, quelle qu'en soit la thématique. En

ce qui concerne le département SAE2 (Sciences sociales, Agriculture et Alimentation, Environnement et Espace), entre les deux tiers et les trois quarts des thématiques nécessitent des données concernant les exploitations et ménages agricoles, les firmes agroalimentaires, les consommateurs, les administrations publiques ou privées, pour lesquels se pose le problème de l'anonymisation des données. Cette évolution se traduit par une mise en ligne de ces données où la protection de l'anonymat se fait en supprimant les données géographiques (code commune) ou par le brouillage de certaines informations. Les données masquées ou brouillées sont malheureusement celles qui deviennent essentielles pour le développement des recherches en économie, notamment pour gérer l'inter-opérationnalité entre ces données économiques et sociales et les données biogéophysiques d'impacts ou de contraintes localisées au cœur des enjeux pluridisciplinaires. La mise en place de Centres d'accès sécurisé aux données (CASD), comme le propose l'INSEE à certains organismes ou laboratoires, permettrait de lever ces contraintes.

Dans le même temps, la raréfaction des moyens attribués aux organismes publics de production de données entraîne une suppression de certaines enquêtes jugées peu centrales et un resserrement sur les opérations lourdes, stratégiques, ou obligatoires (Recensement de population ou recensement agricole, etc.) La suppression d'une enquête engendre la recherche de sources de substitution qui peuvent avoir trois origines : producteurs privées de données (type World Panel Kantar[37]), recours à des données administratives, mise en œuvre d'une opération de production de données alternatives. Chacune de ces trois solutions est plus onéreuse que le recours à des données de producteurs publics (qui sont de fait mutualisées entre de multiples utilisateurs). Une réflexion devrait pouvoir être lancée sur l'opportunité à construire des opérations lourdes et communes à plusieurs organismes de recherche de production de données spécifiques. Un bon exemple actuel est celui de la Cohorte ELFE, initiée par l'INED (Institut National des Etudes Démographiques) et à laquelle l'INRA (et le département SAE2) est associé pour, notamment, son volet alimentation-nutrition

Une des questions centrales pour le futur des recherches socio-économiques de l'Inra est celle de l'appariement de données entre sources économiques et sociales (très lourde interdiction et légitime inquiétude) et l'inter-opérationnalité entre données recueillies à des échelles différentes, notamment les données économiques et sociologiques, et données biogéochimiques. Un travail spécifique est souvent invoqué au sein des organismes (ou interorganismes) mais reste bien incantatoire.

4. ANALYSE SWOT DE LA SITUATION À L'INRA

Au terme de l'analyse des principaux documents étudiés et des résultats des enquêtes, le groupe a identifié les problèmes et leviers principaux. Nous les présentons sous la forme d'un diagnostic SWOT (Strengths, Weaknesses, Opportunities and Threats) pour préparer les recommandations :

- Forces (*Strength*) :
  - Prise en compte de l'importance de la question « Gestion et partage des données » au plus haut niveau (Direction Générale).

[37] http://www.kantarworldpanel.com

- Fonction IST Inra mobilisée sur la question. Le référencement des données par DOI est aujourd'hui une réalité, à l'exemple de GigaScience.
- Identification de grandes familles de SI (Systèmes d'Information) via l'audit. L'identification de ces grandes familles doit permettre de faciliter la prise en charge de la réflexion au regard de l'existant en matière de communauté et de partage pour chacune d'entre elles.
- L'existence des CATI et de la cartographie thématique associée doit permettre de faciliter la communication, de mettre en place et de mutualiser des moyens pour développer les méthodes de recherche, les outils/infrastructures indispensables à la construction nécessaire de l'interopérabilité des données et des outils.
- Existence d'une organisation pour le PEPI (Partage d'Expériences et de Pratiques en Informatique). Cet instrument devrait constituer un vecteur des échanges de pratiques et d'expériences entre CATI (inter-CATI et inter-thématiques) et favoriser la construction collective.
- Existence et bonne connaissance des collectifs d'ingénieurs dont beaucoup sont mobilisés sur la question de la gestion et du partage des données et en capacité d'anticiper des besoins (*e.g.*, au sein des plateformes, CATI).
- Intégration des collectifs (Recherche/Ingénierie) Inra dans les grands projets nationaux (*e.g.*, Infrastructures, Equipements d'excellence) en lien avec les domaines de recherche Inra.

• Faiblesses (*Weaknesses*) :

- Couverture large des champs du haut débit, dispersion et multiplicité des infrastructures de production haut débit.
- Diversité des objets étudiés.
- Gouvernance complexe des CATI de la deuxième vague dans un processus en cours de construction.
- Absence de prise de conscience des scientifiques par rapport au haut débit et à l'importance du partage des données.
- Absence de politique incitative vis-à-vis de la gestion et de la valorisation des données.
- Caractère transversal peu priorisé dans la chaîne d'arbitrage des postes. Le processus de sélection des profils donne la priorité aux postes thématisés au détriment des profils transversaux.
- Pas de moyens attribués ni d'engagement vis-à-vis des mouvements de l'*Open Data* et du *Big Data*.

• Opportunités (*Opportunities*) :

- Construction d'une vision globale et collective des données scientifiques à l'Inra.
- Restructuration en cours de la composition des CATI au regard des grandes familles de SI.

- Mobilisation effective de personnels IST sur la question de la gestion et du partage des données.
- Démarrage de projets pour lesquels la production et l'analyse de données sont au cœur des recherches. Il est essentiel pour l'institut d'être en capacité d'évaluer les coûts sur la durée des projets et au-delà de leur durée.

• Menaces/risques (*Threats*) :

- Perte des données.
- Non visibilité des données de qualité (non accessibles, non référencées).
- Interopérabilité des données et des outils associés impossible.
- Coût supplémentaire exorbitant (financier, humain) pour l'institut des projets ambitieux qui ont mal évalué les questions de la gestion, du partage et du traitement des données.

5. RECOMMANDATIONS

A l'heure actuelle, la grande majorité des données de recherche produites par l'Inra ne sont pas répertoriées. Celles qui n'ont pas été stockées dans une base de données (ou archivées dans un entrepôt) sont probablement perdues ou méconnues et donc peu utilisées en dehors de l'équipe qui les a produites. De ce fait, un grand ensemble de données qui pourraient être réutilisées ne le sont pas – voire sont inutilement dupliquées – par méconnaissance de l'existant. La faible mise à disposition des données constitue aussi un handicap au développement de projets collaboratifs, d'analyses comparatives et plus globalement de méta-analyses. Une modification de ces pratiques, pour accorder aux données l'attention nécessaire, dépendra non seulement de la politique suivie par l'institut, mais également de l'implication active des chercheurs et des ingénieurs. Les recommandations qui suivent ne couvrent pas le volet « analyse » des gros volumes de données. Ce volet est tout autant stratégique que le volet « gestion et partage », et mériterait des investigations pour définir les compétences, les moyens et les dispositifs nécessaires à une science numérique (ou *e-science*) (Hey *et al*., 2009).

5.1. DÉFINIR LA POLITIQUE DE L'ÉTABLISSEMENT ET LA COMMUNIQUER

Il est nécessaire au niveau institutionnel de définir une politique générale en matière de gestion et de partage des données. Cette politique définira les grandes lignes des modalités de publications des données (quelles données publier, sous quelles conditions et dans quelles limites, moyens et lieux de publication), précisera les modèles économiques possibles (gratuité, payante) et cadres juridiques de réutilisation. Elle devra être incitative pour encourager la formalisation de bonnes pratiques et l'utilisation de standards. Elle sera instanciée par les différents départements en fonction des domaines de recherche et du caractère stratégique associé, puis sera mise en œuvre sous la responsabilité des Directeurs d'Unités en regard des SI hébergés concernés.

Cette politique de l'établissement devrait prendre en compte les points suivants et être déclinée aux niveaux de gouvernance appropriés :

• Définir une politique d'investissement équilibrée entre gros et petits projets. Le

terme petit projet n'impliquant pas pour autant que les volumétries et la complexité des données soient nécessairement faibles. Le fait même que la production (et ensuite l'accès) à des volumes importants de données soit ouvert à un plus grand nombre de chercheurs implique aussi de donner aux équipes qui le demanderaient les moyens pour les traiter.

- Communiquer les règles, sensibiliser les chercheurs et les ingénieurs à l'*Open Data*, aux thématiques de la protection des données, de la propriété intellectuelle et du droit d'auteur, au cadre juridique. Des formations internes pourront être organisées au cours desquelles seront expliqués de manière simple les différents dispositifs juridiques. Après coup, transmettre aux chercheurs et ingénieurs le contenu de la négociation des contrats en langage non juridique.
- Veiller à la réalisation d'une contractualisation avant le démarrage des projets. Cette contractualisation étant de plus en plus nécessaire du fait de la judiciarisation de la société et des risques que ce phénomène fait peser sur le partage des données.
- Expliquer où et comment calculer, stocker et archiver.
- Donner des consignes et effectuer des arbitrages sur les données qui doivent être conservées et rendues accessibles. A ce niveau, il n'est pas possible d'effectuer des recommandations trop générales, mais plutôt spécifiques aux domaines concernés.
- Valoriser les activités liées à la gestion et au partage des données.

Concernant ce dernier point, le dépôt numérique de connaissances fiables repose sur la mise à jour et la validation des données. Le développement et la gestion courante des bases de données ne sont pas considérés aujourd'hui de la même façon que les publications d'articles de recherche. Pour les domaines où il n'est pas possible de publier un article scientifique décrivant une base de données, l'utilisation de ces bases est alors ignorée des indicateurs de citation. *A contrario*, lorsqu'il est possible de publier de tels articles, leur valorisation est généralement limitée à une seule publication, ceci alors que le travail de maintenance peut se poursuivre sur de nombreuses années. Il est donc important d'inciter les concepteurs de bases de données (et les personnes impliquées dans leur maintenance) à mettre en place des systèmes de qualité, ouverts et structurants, plutôt que de multiplier les développements *ad hoc* au fil des projets. Une politique de soutien actif devra impérativement prendre en compte la visibilité (ou la potentialité) des dites bases.

5.2. METTRE EN PLACE UN DISPOSITIF D'ÉVALUATION DES DONNÉES PRODUITES PAR L'INRA

Les structures et le financement de la recherche ne prévoient pas la gestion des données au-delà d'une certaine période. Une évaluation scientifique régulière des données (échéance à définir et variable selon la communauté d'origine) pourra suggérer des pistes quant à la nécessité de leur maintien et de la forme sous laquelle il doit être effectué. Ce dispositif mettrait en place des outils d'évaluation de l'existant (bases de données, logiciels, etc.). Dans l'optique de cette évaluation, il serait nécessaire de disposer d'un répertoire à jour de l'existant (données, métadonnées, logiciels permettant leur exploitation, etc.) avec son statut (en développement, en production, en sommeil,

arrêté). Pour ce faire, il faudrait désigner des référents locaux, ayant à charge d'alimenter ce répertoire.

5.3. S'IMPLIQUER DANS LES COMITÉS INTERNATIONAUX DE STANDARDISATION

La standardisation des données progresse de façon constante dans la plupart des champs disciplinaires. L'utilisation de standards et la définition d'ontologies est fondamentale parce qu'elle facilite le travail des chercheurs et l'interopérabilité des données. Il serait donc nécessaire d'encourager l'implication de personnels Inra (biologistes, informaticiens) dans les comités internationaux de standardisation. Cette implication permettrait aux chercheurs et ingénieurs de contribuer au développement des standards, de prendre connaissance et d'adopter les standards existants, d'être en capacité de les communiquer, voire de les imposer. Par exemple, la position dominante actuelle de l'URGI[38] (Unité de Recherche en Génomique et Informatique) en matière de diffusion des données acquises sur le blé lui a permis d'influencer les standards employés au niveau international dans ce domaine. Cette position a été atteinte du fait de l'existence d'une communauté importante d'utilisateurs à l'Inra et, simultanément, de la capacité de cette équipe à mettre rapidement à disposition des données d'intérêt. Dans les autres cas, le degré d'investissement devra être soigneusement pesé en fonction des enjeux et des forces en présence.

5.4. DÉVELOPPER UN PORTAIL D'ACCÈS À UN ENSEMBLE DE RESSOURCES DISTRIBUÉES

Les ressources identifiées comme étant stratégiques – aussi bien les bases de données que les outils de traitement – sont actuellement atomisées, et la mise à disposition d'un portail intelligent qui permette d'y accéder de façon conviviale et transparente est incontournable. En effet, il s'agit là d'un besoin impérieux lié à la croissance de la disponibilité locale de technologies « nouvelle génération » (*e.g*, nouvelles technologies de séquençage). La mise en place de ce portail donnant accès à l'information sur les ressources, et à ces ressources en tant que telles, permettra à la fois une meilleure exploitation des données mais aussi leur conférera une plus grande visibilité. L'existence du rapport « Bases de Données » (Christophe, 2009) et la liste des actifs stratégiques obtenus suite à l'audit d'Ernst & Young pourraient fournir les éléments de départ pour l'élaboration d'une première version de ce portail. Les solutions techniques permettant son développement sont néanmoins à étudier. Les éléments déterminants pour le succès de l'initiative seront l'accès facilité à l'information mais aussi l'utilisation par ces ressources de standards disponibles à l'échelle internationale. Un autre point est l'écriture de la documentation et de tutoriels de haut niveau permettant d'optimiser l'utilisation des ressources.

5.5. PRENDRE EN COMPTE LE CYCLE DE VIE DES DONNÉES DÈS L'ÉLABORATION DES PROJETS DE RECHERCHE

Compte tenu de l'évolution rapide des technologies et de la diversité des applications rendues possibles, il est nécessaire de promouvoir que les chercheurs s'assurent *a priori* de la pertinence des protocoles en regard des questions posées, en associant des

[38] http://urgi.versailles.inra.fr

méthodologistes (biostatisticiens, bioinformaticiens, modélisateurs) dès la phase de conception des projets. Il est tout autant nécessaire que ces mêmes chercheurs se mobilisent dès la construction de leur projet sur les modalités de gestion (stockage, archivage) et de mise à disposition (portail, banques de données institutionnelles, nationales ou internationales, standards pour les métadonnées) des données.

Les métadonnées et la documentation doivent être mises en place dès le début d'un projet de recherche. Chaque appel d'offre Inra devrait exiger que, dans les projets présentés, figure le volet « gestion et partage des données ». Cette exigence pourrait être également promue dans le cadre des projets partenariaux. Chaque projet, pour être éligible, aurait à construire un plan de gestion et de partage des données et à en prévoir les coûts : élaboration d'une base de données, stockage, archivage sur une période plus ou moins longue en fonction de l'intérêt des données. Les règles mises en place par l'Inra seront adaptées en fonction des contraintes portées par les partenaires hors Inra du projet. Le comité d'évaluation de l'appel d'offres examinerait la proposition des chercheurs et s'assurerait qu'il est conforme à la politique de l'Inra. C'est une condition indispensable qui permettra de ne pas perdre de données, de réduire de beaucoup le temps, mais aussi le coût financier, nécessaires pour assurer la diffusion et l'accessibilité des données à long terme.

Il serait également nécessaire de mettre à la disposition des chercheurs des outils qui leur permettront de renseigner au mieux des métadonnées de qualité pendant tout le cycle de vie des données, ceci selon des normes reconnues nationalement ou internationalement. En effet, l'adoption de normes « locales », si elle peut permettre un gain immédiat, entraîne souvent une marginalisation à long terme. Par ailleurs, le renseignement de ces métadonnées doit se faire conformément aux directives pour les données privées et confidentielles.

Les professionnels de l'IST (Information Scientifique et Technique) pourraient assister les chercheurs dans tous les processus relatifs à la construction des métadonnées, la maintenance des référentiels (liste de termes, ontologies) et les questions de propriété intellectuelle (propriété des données, des bases de données, etc.). Les documentalistes disposent d'un socle de compétences et d'outils acquis dans la gestion des publications qui est transposable à la gestion des données, mais il est important de souligner qu'ils n'ont généralement pas une connaissance intime de la nature des recherches, et des jeux de données associés. Ils ne sont donc pas en mesure de définir les métadonnées correspondantes. Une implication forte des chercheurs, producteurs de données, et des personnels en charge du traitement des données, est donc nécessaire. La typologie des compétences liées à la gestion des données (*cf.* section 2.) est généralement méconnue par les chercheurs alors que les personnes en charge des traitements sont de plus en plus mobilisées sur ces questions. Une sensibilisation à la réalité et à l'utilité de ces problématiques doit être effectuée pour mieux organiser l'adéquation entre besoins et compétences ainsi que leur articulation. On peut envisager que des documentalistes se mobilisent pour acquérir ces compétences, voire que l'Inra recrute sur les profils correspondants.

Pour une prise en charge opérationnelle, ce soutien pourrait être organisé à l'intérieur des départements par des personnels ou des unités de service qui seraient garants de la mise en place opérationnelle des standards. Ces structures auraient pour mission de gérer l'information liée à l'utilisation de ces standards, les aspects

documentation et propriété intellectuelle liée à la valorisation des données. Elles aideraient à la mise en place de bases de données lorsqu'il s'agit de projets d'envergure impliquant différents acteurs ou des données hétérogènes. Enfin, elles participeraient à l'écriture des métadonnées et interviendraient dans la création de documents d'accompagnement clairs et d'utilisation facile. A ce niveau, il est important d'impliquer le plus possible les chercheurs des différentes communautés dans le processus de modélisation conceptuelle de leurs problématiques et de leurs besoins.

5.6. DÉFINIR UN CAHIER DES CHARGES POUR LES PLATEFORMES

Le rôle des plateformes est à réfléchir avec soin : outre la gestion des données, ces structures illustrent la dualité entre production massive (qui conduit à une certaine standardisation des protocoles) et une contribution spécifique aux projets (un problème ne se représente jamais deux fois à l'identique). Par conséquent, la définition d'un cahier des charges pour les plateformes doit satisfaire à deux contraintes : une production à la fois individualisée (chaque projet étant par essence unique) tout en respectant des délais permettant de traiter un maximum de demandes. Cantonner les plateformes à des opérations de routine les conduit à ne plus prendre part à la production de résultats innovants. En conséquence, les ingénieurs et chercheurs rattachés aux plateformes doivent garder le contact avec la recherche dans leur domaine. C'est à cette condition qu'ils seront en mesure de conduire des développements méthodologiques et de traiter des projets non canoniques, projets qui sont potentiellement les plus intéressants du point de vue scientifique.

Dans le cas des CATI, les évolutions actuelles au sein de l'Inra visent à mutualiser les ressources humaines autour d'un projet commun d'outils techniques de production et de gestion de données. Une coordination intra et inter CATI, au-delà des PEPI (Partage d'Expérience et de Pratiques en Informatique), permettrait de favoriser un partage sur l'interopérabilité dans tous les domaines scientifiques représentés dans les différents CATI. Se pose la question du pilotage de telles structures qui doivent rester au service d'une diversité de stratégies scientifiques.

Les enquêtes réalisées ont montré la satisfaction des utilisateurs Inra des plateformes soutenues par l'institut. Pour accroître leurs capacités de réponse à l'accompagnement des projets scientifiques, ces plateformes devront être évaluées très régulièrement pour les maintenir au meilleur niveau d'accompagnement. Ce suivi régulier devrait permettre d'avoir une plus grande réactivité face aux évolutions nécessitées par les ruptures technologiques.

5.7. DOTER L'INRA D'INFRASTRUCTURES DIMENSIONNÉES POUR LES STOCKAGES ET CALCULS HAUTES PERFORMANCES

Le caractère non prévisible des (r)évolutions technologiques de production des données dans tous les domaines demande une grande réactivité dans l'organisation des infrastuctures. Il devient indispensable de mettre en place en lien avec les partenaires scientifiques et régionaux de l'Inra une politique d'équipement pour le stockage et le traitement des données, ceci dans un nouveau contexte de dispersion et de volumes massifs (changement d'échelle).

L'Inra doit proposer *a minima* des solutions techniques transparentes et pérennes garantissant à l'ensemble des scientifiques de l'Institution un accès efficace au

stockage, à la pérennisation, au partage et au calcul à grande échelle dans le cadre de ses projets et programmes stratégiques.

L'offre de calcul se décline traditionnellement en trois niveaux pour des stratégies d'équipements en un lieu unique : (i) un niveau local, avec des machines dont l'achat et la maintenance sont à la portée d'une unité de recherche (investissement de l'ordre de quelques dizaines de milliers d'euros) ; (ii) un niveau intermédiaire, typiquement des mésocentres de calcul d'équipements mutualisés, du coût de l'ordre de quelques centaines de milliers d'euros ; (iii) un niveau national (quelques millions d'euros, typiquement le Grand Equipement National de Calcul Intensif – GENCI[39]).

Une stratégie alternative multisites fondée sur une évolution technologique a permis de relier les clusters locaux en un réseau maillé afin d'avoir accès à des calculs massivement distribués. Cette technologie, qui a débuté vers l'établissement de grilles, évolue actuellement vers la technologie du *cloud*, et pour ce qui concerne la recherche, vers les *clouds* académiques. Toutes ces technologies et équipements proposent une offre tant en calcul qu'en stockage, mais pas en archivage.

Aussi, il convient d'allouer au mieux les moyens de l'Inra entre ces différentes solutions, notamment entre : (i) le partenariat avec des mésocentres existants, utile pour les calculs massivement parallèles ou nécessitant de grandes capacités mémoires (comme l'assemblage en génomique) ; (ii) l'investissement vers une (ou des) plateformes Inra qualifiées de « stratégiques », d'échelle intermédiaire (comme les mésocentres) ; (iii) un investissement en partenariat dans un *cloud* académique – typiquement le GIP (Groupement d'Intérêt Public) France-Grille dans lequel l'Inra est partenaire – pour des calculs massivement distribués.

Concernant le stockage/archivage des données et leur mise à disposition, une organisation centralisée devrait permettre une économie de moyens. La cartographie des systèmes d'informations maintenant disponible au sein de l'Inra doit permettre, en lien avec les capacités réseaux nécessaires, de proposer une (ou plusieurs) localisation(s) géographique(s) stratégique(s) pour la mise en place de centres de données. Un *cloud* académique à l'échelle de l'institut (ou de plusieurs instituts), s'appuyant sur un (ou plusieurs) centre(s) de données présentera alors un intérêt majeur.

Quelle que soit la solution retenue, l'Inra doit prévoir les investissements lui permettant de se doter d'un maillage à haut débit de l'ensemble de ses sites, notamment pour les sites isolés.

5.8. S'ENGAGER DANS UNE POLITIQUE DE GESTION DES COMPÉTENCES RÉPONDANT AUX BESOINS EN ÉMERGENCE

La question de la formation continue est consubstantielle à la recherche scientifique. Cette dimension devient cruciale dans un contexte où de nouvelles technologies sont accessibles à un nombre croissant de chercheurs et où leur maîtrise est nécessaire pour la production d'une recherche au meilleur niveau international.

L'enquête qui a été conduite auprès des chercheurs et des ingénieurs a fait apparaître la difficulté d'appréhender les besoins. En premier lieu, du fait du décalage entre les besoins perçus par les acteurs et leurs besoins réels. Et en second lieu, du fait

[39] http://www.genci.fr/

de l'obsolescence de l'offre de formation sur le marché, au regard de la rapidité des évolutions techniques.

Face au risque d'ores et déjà perceptible d'émiettement des compétences, il est nécessaire de construire une offre de formation interne coordonnée et structurante, qui s'appuie sur une vision stratégique nationale des compétences à acquérir et des technologies prometteuses à l'échelle de l'institut. Cette vision stratégique devra être dûment explicitée et prise en compte dans la gestion prévisionnelle des compétences.

Il faudrait promouvoir la création de formations initiales universitaires pour développer les métiers liés au partage des données encore inexistants en France (mais présents dans les pays anglo-saxons, *cf.* section 2.). L'Inra pourrait être force de proposition et s'investir dans la définition des cursus. Un soutien fort de l'institut doit être mis en place pour faciliter l'acquisition de ces nouvelles compétences par les personnels de la recherche et de la fonction IST en lien avec les besoins de la recherche.

Une mutualisation inter-organismes (potentiellement déclinable en région) des ressources IST, par exemple avec le JISC (*cf.* Annexe A.1.4.), permettrait la fédération des initiatives tout en représentant une économie d'échelle.

Enfin, l'Inra devrait participer à la mise en place de modules traitant du partage et de la gestion des données, de la propriété intellectuelle, de l'éthique, au niveau des Ecoles doctorales.

5.9. CONDUIRE UNE RÉFLEXION INTER-ORGANISMES

L'afflux massif des données est potentiellement source d'initiatives multiples non coordonnées. L'Annexe A.1. fait un rapide tour d'horizon de la façon dont d'autres organismes se saisissent de ce type de problème. Compte tenu de l'ampleur des enjeux – enjeux partagés par l'ensemble de la recherche française et internationale dans le domaine – il est évident que les réflexions sur les structures ne sauraient être limitées au périmètre de l'Inra seul.

En bioinformatique, la réflexion dans le cadre du projet déposé par ReNaBi et des propositions qui en découlent (IFB – Infrastructure Française de Bioinformatique), pourrait rapidement aboutir, en lien avec les grands projets d'infrastructure nationaux, au caractère opérationnel du partage de compétences en matière de traitement, de gestion et de partage des données. Sur les autres thématiques finalisées (agronomie, écologie, SHS), l'Inra doit recenser les initiatives en cours et s'y impliquer (*e.g.*, infrastucture ANAEE-Service), voire être force de proposition.

De manière générale, les investissements d'avenir devraient permettre de fédérer des initiatives inter-organismes centrées sur des objectifs ou des communautés scientifiques. Il nous semble nécessaire, à la clôture de ces appels d'offres et au démarrage de ces projets, de mobiliser les porteurs autour des questions de gestion et de partage des données. Le recensement des initiatives devra être utilisé pour sélectionner les champs prioritaires à développer en propre pour l'Inra.

Par ailleurs, l'Inra en tant qu'organisme de recherche finalisé, producteur de données, est en position de participer à l'orientation d'une politique nationale dans le domaine de la gestion des données. Cette question pourrait émerger à l'issue d'une réflexion entre les organismes et l'ANR pour définir des modalités qui doivent permettre au chercheur de mesurer immédiatement le bénéfice qu'il retire du dépôt de ses données dans l'entrepôt prévu.

RÉFÉRENCES

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403-410.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST : A new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389-3402.

Ashelford, K.E., Chuzhanova, N.A., Fry, J.C., Jones, A.J. and Weihtman, A.J. (2005) At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. *Appl. Environ. Microbiol.*, **71**, 7724-7736.

Beagrie, N. (2007) *E-infrastructure strategy for research : Final report from the OSI preservation and curation working group*. Beagrie Publishing. http://www.nesc.ac.uk/documents/OSI/preservation.pdf

Cochrane, G.R. and Galperin, M.Y. (2010) The 2010 nucleic research database issue and online database collection: A community of data resources. *Nucleic Acids Res.*, **38**, D1-4.

Christakis, N.A. and Fowler, J.M. (2007) The spread of obesity in a large social network over 32 years. *New Engl. J. Med.*, **357**, 370-379.

Christophe, C. (2009) *Analyse de l'inventaire des bases de données scientifiques*. Rapport de mission – Système d'Information, Document interne Inra.

Greenbaum, D., Sboner, A., Mu, X.J. and Gerstein, M. (2011) Genomics and privacy : Implications of the new reality of closed data for the field. *PLoS Comput. Biol.*, **7**, e1002278.

Hey, T., Tansley, S. and Tolle, K. (2009) *The Fourth paradigm : Data-intensive scientific discovery*. Microsoft Research, Redmond. http://research.microsoft.com/en-us/collaboration/fourthparadigm/

Kahn, S.D. (2011) On the future of genomic data. *Science*, **331**, 728-729.

King, G. (2011) Ensuring the data-rich future of the social sciences. *Science*, **331**, 719-721.

Leinonen, R., Sugawara, H. and Shumway, M. (2011) International nucleotide sequence database collaboration. The sequence read archive. *Nucleic Acids Res.*, **39**, D19-21.

Lyon, L. (2007) *Dealing with data : Roles, rights, responsibilities and relationships*. UKOLN, University of Bath. http://www.jisc.ac.uk/media/documents/programmes/digitalrepositories/dealing_with_data_report-final.pdf

Michel, J.-B., Shen, Y.K., Aiden, A.P., Veres, A., Gray, M.K., The Google Books Team, Pickett, J.P., Hoiberg D., Clancy, D., Norvig P., Pinker, S., Nowak, M.A. and Lieberman Aiden, E. (2010) Quantitative analysis of culture using millions of digitalized books. *Science*, **331**, 176-182.

Piwowar, H.A., Day, R.S. and Fridsma, D.B. (2007) Sharing detailed research data is

associated with increased citation rate. *PLoS ONE*, **2**, e308.

Reichman, O.J., Jones, M.B. and Schildhauer, M.P. (2011) Challenges and opportunities of open data in ecology. *Science*, **331**, 703-705.

Richardson, E.J. and Watson, M. (2012) The automatic annotation of bacterial genomes. *Brief. Bioinformatics,* E-pub 9 March.

Sansone, S.A. *et al*. (2012) Toward interoperable bioscience data. *Nature Genet.*, **44**, 121-126.

Swan, A. and Brown, S. (2008) *The skills, role and career structure of data scientists and curators: An assessment of current practice and future needs*. School of Electronics and Computer Science, University of Southampton. http://eprints.ecs.soton.ac.uk/16675/

The 1000 Genome Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061-1073.

Vardigan, M., Heus, P. and Thomas, W. (2008) Data documentation initiative: Toward a standard for the social sciences. *Int. J. Digital Cur.*, **3**, 107-113.

Via, M., Gignoux, C. and Burchard, E.G. (2010) The 1000 genome project: New opportunities for research and social challenges. *Genome Med.*, **2**, 3.

REMERCIEMENTS

Jean-Pierre Castelli (Inra) pour nous avoir présenté la problématique de la propriété intellectuelle. Renaud Albigot (CNRS), Mireille Cadou (ARSOE de Bretagne), Anne Cambon-Thomsen (CNRS), Philippe Auclair (Inra), Maurice Barbezant (UNCEIA), Dominique Boutigny, Pascal Calvat (CC-IN2P3), Philippe Breucker, Eric Cahusac, Jean-Philippe Cointet, Jérôme Gouzy, Nicolas Guinet (Inra), Pascal Kosuth (Irstea), Jean Lobry (Police Scientifique de Lyon), Bernard Monsarrat (Inserm), Bertrand Schmitt (Inra), et Thierry Simon (CREAVIA), pour le temps consacré à nous présenter leurs activités et à répondre à nos questions. Vincent Miele, Bruno Spataro, Stéphane Delmotte (Université Lyon 1) pour avoir bien voulu tester nos questionnaires. Dominique Fournier (Inra) pour avoir pris en charge la mise en ligne de nos questionnaires. Capucine Gallouët (Inra) qui a réalisé un travail exceptionnel pour nous aider à exploiter l'information contenue dans les réponses aux questionnaires, Elisabeth de Turckheim (Inra) pour ses conseils. Catherine Christophe (Inra), Christian Gautier (Université Lyon 1), Muriel Mambrini-Doudet (Inra), Catherine Oudin (Lyon Ingénierie Projet), Franck Picard (CNRS), Anne-Françoise Schmid (INSA), pour les échanges sur la problématique « gestion et partage des données ». Emmanuelle Maguin, Patrick Etievant (Inra) pour les discussions sur les métaprogrammes Inra. Les membres de l'URGI (Inra) pour la présentation de la problématique « blé ». Enfin, nous remercions Christine Schrive (CNRS), Jean-Louis Verrel (IGREF, en retraite) et Marie-France Sagot (INRIA) pour la relecture du document.

DOCUMENT ANNEXE

A.1. EXEMPLES D'INFRASTRUCTURES EN FRANCE ET À L'ÉTRANGER

A.1.1. Elixir et ReNaBi

Concernant les données du vivant, l'initiative européenne Elixir (European Life Sciences Infrastructure for Biological Information) est à prendre en considération. L'objectif de cette initiative est la mise en place d'une infrastructure bioinformatique européenne. Elixir sera constitué d'une station centrale hébergée par l'EBI[40] (European Bioinformatics Institute) à Hinxton (Royaume-Uni), et d'entités scientifiques nationales (les « nœuds ») réparties dans des centres d'excellence des pays européens qui ont accepté de participer au réseau. L'EBI est situé sur le même campus que le Sanger Institute[41] et les thématiques des deux structures sont en articulation forte. La future infrastructure aura pour but de mettre en réseau les ressources bioinformatiques européennes de pointe. L'appel d'offre Elixir est paru fin avril 2010 et la France réfléchit actuellement à la meilleure solution technique pour mettre en place un nœud qui sera raccordé au réseau.

Il est à noter que la solution technique proposée est très fortement centralisée, avec l'EBI se situant au centre du dispositif, et les nœuds nationaux correspondant eux-mêmes à des grands centres de bioinformatique. Le projet ReNaBi-IFB – porté conjointement par le Génoscope et l'Inra de Jouy-en-Josas – dans le cadre des Infrastructures Nationales, devrait développer un tel nœud pour la France.

La mise en place d'une structure européenne aussi centralisée pose le problème de la transmission de très grands volumes de données, sachant que les débits des réseaux ne sont pas extensibles à l'infini. En effet, il est désormais courant que le temps nécessaire pour faire transiter les données vers un serveur soit supérieur au temps requis pour en générer de nouvelles. Une conséquence de ce fait est que les trois banques généralistes (*i.e.*, GenBank, EMBL et DDBJ) ne sont d'ores et déjà plus exhaustives comme elles ont pu l'être pendant près de trois décennies. De plus, les investissements massifs et récurrents qu'il faut effectuer afin de se doter des capacités de stockage nécessaires, font que la politique de certains organismes fluctue d'une année sur l'autre (exemple du NCBI avec la question de l'archivage des lectures courtes produites par les séquenceurs « nouvelle génération »). On peut donc s'attendre, au regard d'initiatives telles que celle développée par la revue GigaScience ou bien le projet 1000 génomes (The 1000 Genome Project Consortium, 2011), à la mise en place rapide de solutions de partage s'appuyant sur le *cloud*.

A.1.2. GBIF – LifeWatch – ANAEE

En matière de biodiversité et d'environnement des projets fédérateurs se mettent en place au niveau international. Le projet GBIF[42] (Global Biodiversity Information Facility) se positionne en tant que système d'information mondial. Il vise à mettre à disposition les données primaires de biodiversité en s'appuyant sur les métadonnées et l'interopérabilité des bases. GBIF France est son point nodal pour la France.

[40] http://www.ebi.ac.uk/
[41] http://www.sanger.ac.uk/
[42] http://www.gbif.org/

Le projet LifeWatch[43] est une infrastructure à l'échelle européenne concernant la biodiversité et les écosystèmes. Au travers d'un réseau de bases de données LifeWatch permettra de renforcer les capacités de modélisation en biodiversité en offrant l'accès aux ressources informatiques sur les données de génomique, d'inventaire des espèces et des spécimens. LifeWatch devrait être opérationnel en 2015. La contribution française au projet est gérée par le MNHN (Muséum National d'Histoire Naturelle). Ces deux infrastructures coopèrent pour le développement et le partage des ressources informatiques en biodiversité.

Tout comme LifeWatch, le projet ANAEE[44] (ANAlysis and Experimentation on Ecosystem) est inscrit sur la feuille de route du Forum pour la Stratégie Européenne en matière d'Infrastructures de Recherche (ESFRI[45]). Il permettra la mise en place de plateformes d'observation et d'analyse et modélisation des écosystèmes européens, continentaux terrestres et aquatique.

La composante française au projet ANAEE-Europe fait l'objet de Infrastructure ANAEE-S(ervice) retenue en 2012 dans le cadre de l'appel d'offres « Infrastructures Biologie et Santé » et co-portée par le CNRS et l'Inra. En associant sciences du vivant et sciences de l'environnement, ANAEE-S fournira les outils nécessaires à l'analyse intégrée des écosystèmes, objets complexes présentant des interactions entre les compartiments biotiques et abiotiques. L'infrastructure expérimentale s'articule en trois nœuds de services ouvrant l'accès aux meilleurs dispositifs expérimentaux et à leurs ressources biologiques ainsi qu'aux données produites :

- Les deux Ecotrons de la Très Grande Infrastructure du CNRS.
- Trois plateformes en milieu semi-naturel pour des écosystèmes terrestres et aquatiques.
- 21 sites pour l'expérimentation *in situ* sur le long terme (forêts, cultures, prairies et lacs).

ANAEE-S investit également dans l'instrumentation partagée et les plateformes de microbiologie environnementale. Enfin ANAEE-S développe des services de bases de données et métadonnées ainsi que des plateformes de modélisation pour l'analyse de la dynamique des écosystèmes et du rôle des gestionnaires des milieux.

La mise en place de ces infrastructures nécessitera de réaliser un travail considérable de standardisation dans l'acquisition et la caractérisation des données, ainsi que d'interopérabilité entre les systèmes d'information pour la gestion des données et la modélisation. Au niveau national, cet enjeu constitue, de fait, une opportunité de coopération renforcée entre les organismes de recherches.

A.1.3. CC-IN2P3

Une structure ayant l'habitude de gérer de très grands volumes de données est le CC-IN2P3[46] (Centre de Calcul de l'Institut National de Physique Nucléaire de Physique des Particules). Le CC-IN2P3 est l'un des onze centres membres d'un réseau mondial de

[43] http://www.lifewatch.eu
[44] http://www.anaee.com/
[45] http://ec.europa.eu/research/infrastructures/index_en.cfm?pg=esfri
[46] http://cc.in2p3.fr/

distribution et de traitement des données. Bien que les données de la physique des particules présentent une grande spécificité, avec une homogénéité bien plus importante que ce qu'on peut trouver en biologie, l'expérience du CC-IN2P3 est à prendre en compte par rapport aux solutions que pourrait mettre en œuvre l'Inra.

En physique des particules, les données brutes sont systématiquement conservées, le filtrage s'effectuant au niveau des instruments de mesure. Les données sont centralisées et les centres de traitement vont les récupérer dès l'instant où elles sont nécessaires (pas de duplication sur un ensemble de sites comme dans le cas des données de la génomique). Ces données sont temporalisées, ce qui implique la nécessité de leur conservation, une donnée ne pouvant prendre parfois tout son sens que plusieurs années après son obtention. Elles peuvent également devenir obsolètes.

Le CC-IN2P3 est l'exemple type d'une solution où le stockage, l'archivage et les calculs sont effectués sur place. Comme toutes les solutions de ce type, elle a nécessité de lourds investissements en termes de personnels, d'infrastructure, d'équipement et de fonctionnement. A l'heure actuelle, le centre emploie 81 personnes dont 37 CDD de longue durée (2×3 ans en moyenne). La construction d'un nouveau bâtiment servant à abriter les moyens de calcul a coûté 7,6 M€ (dont 4,8 versés par la région et 2,8 par le CNRS). Par ailleurs, la consommation électrique est de l'ordre de 1,8 MW, avec un maximum fixé à 7,8 MW[47]. Le coût de fonctionnement est de l'ordre de 700 k€/an, financés par le TGIR (Très Grandes Infrastructures de Recherche) et par une convention avec le CEA. Enfin, ce sont près de 20000 cœurs qui sont disponibles, ainsi qu'une capacité de stockage de 200 Pétaoctets, dont 10 seulement sont actuellement utilisés. Cet investissement va permettre de disposer d'une capacité d'archivage et de calculs suffisante pour absorber les développements scientifiques en physique des particules sur une dizaine d'années.

Une des caractéristiques de ce centre est qu'il s'est ouvert depuis 2002 aux biologistes et aux sciences sociales et humaines. Il permet aux chercheurs de ces disciplines d'accéder à une fraction de ses ressources de calcul et de stockage. Pour l'ensemble de ces deux domaines, l'effort consenti se situe à environ 5 % des capacités totales du CC-IN2P3. Qui plus est, un effort particulier d'ouverture a été fait, comme en témoigne la création d'un poste d'ingénieur dédié à la tâche d'allocation des ressources à ces deux communautés. Cette ouverture du CC-IN2P3 lui assure une bien plus grande visibilité, et donc des opportunités supplémentaires de financement.

Il est à noter que les solutions de type *cloud* sont suivies de près par la communauté des physiciens des particules. En effet, la virtualisation d'environnements complets étant maintenant facilement réalisable, il est possible de s'affranchir de toutes les contraintes liées à la disponibilité d'un environnement matériel et logiciel. De cette façon, effectuer un calcul dans un environnement très fortement dispersé ne pose plus de problème particulier. Cette utilisation du *cloud* devrait permettre d'optimiser l'utilisation des machines du centre. Par contre, ce genre de solutions n'est pas adapté pour de nombreux problèmes, en particulier ceux nécessitant un accès intensif à de grandes bases de données.

[47] Un système en cours de réalisation permettra de récupérer les calories générées par le fonctionnement des serveurs du centre de calcul qui seront réutilisées pour chauffer le campus.

A.1.4. JISC

Le JISC est une structure transversale créée par les agences de financement anglaises en charge de la recherche et de l'enseignement. Il regroupe des compétences dans le domaine de l'information numérique nécessaires à la recherche et à l'enseignement. Par contraste, le paysage français est beaucoup plus morcelé puisque les fonctions prises en charge par le JISC se retrouvent éclatées entre le RENATER[48] (Réseau National de télécommunications pour la Technologie, l'Enseignement et la Recherche), l'INIST et le CCSD[49] (Centre pour la Communication Scientifique Directe), le Couperin[50] (consortium servant de réseau de négociation et d'expertise des ressources documentaires électroniques), l'ABES[51] (Agence Bibliographique de l'Enseignement Supérieur), le CINES[52] (Centre Informatique national de l'enseignement supérieur) plus quelques autres organismes de moindre importance. L'objectif du JISC est de piloter des programmes transversaux tels que ceux autour de l'*open access* et le programme RDM (Research Data Management).

Lancé en 2009, le programme RDM finance des projets dans les domaines suivants :

- Politique, soutien et infrastructure.
- Formation et compétences.
- Projets disciplinaires.
- Publications, données et citations.

Le JISC s'appuie en particulier sur le DCC[53] (Digital Curation Center) pour développer sa politique de soutien aux universités et aux structures de recherche. La France manque cruellement de ce type d'organisation.

A.1.5. ANDS

L'Australie conduit depuis plusieurs années une politique en faveur de la gestion et la valorisation des données de la recherche. L'Australian Code for the Responsible Conduct of Research[54], a été mis en place par le NHMRC[55] (National Health and Medical Research Council). Ce code, publié en 2007, inclut un chapitre « Management of Research Data and Primary Materials » qui définit les responsabilités des chercheurs et des institutions. Ces dernières ont la responsabilité de conserver les données, d'offrir un système sécurisé de stockage de données, de gérer les questions de propriété intellectuelle.

En appui à cette politique, l'ANDS[56] (Australian National Data Service) a été créé comme une plateforme collaborative. En 2007, le rapport « Towards Australian Data

[48] http://www.renater.fr/
[49] http://www.ccsd.cnrs.fr/
[50] http://www.couperin.org/
[51] http://www.abes.fr/
[52] http://www.cines.fr/
[53] http://www.dcc.ac.uk/
[54] http://www.nhmrc.gov.au/_files_nhmrc/publications/attachments/r39.pdf
[55] http://www.nhmrc.gov.au/
[56] http://www.ands.org.au/

Commons »[57] précise que, dans dix ans, l'ANDS sera un succès si les conditions suivantes sont réunies :

- Un réseau cohérent des données produites par la recherche australienne existe.
- Les chercheurs et les gestionnaires de données australiens sont perçus comme les meilleurs dans les domaines de la gestion et du partage des données et par la pertinence de leur politique dans ce domaine.
- La quantité de données déposées dans des environnements techniques stables et pérennes augmente.
- Le nombre de personnes ayant une expertise dans la gestion des données augmente, tant dans les communautés scientifiques que dans le management de la recherche.
- Les chercheurs australiens peuvent facilement accéder aux données produites en Australie.
- Les chercheurs australiens peuvent découvrir, échanger, utiliser les données produites ailleurs, éventuellement dans d'autres domaines pour explorer de nouvelles voies.
- L'Australie est en mesure de partager des données facilement et de façon transparente pour faciliter les recherches multidisciplinaires sur le plan national et international.

Aujourd'hui, le site de l'ANDS propose des guides et des outils pour tous les aspects de la gestion des données, et son portail[58] donne accès à la description de 35000 collections de données.

A.1.6. GeoSud

L'EquipEx GeoSud[59] (Infrastructure d'Information Spatiale sur les Territoires et l'Environnement) est un consortium regroupant 14 partenaires (dont l'Irstea, le Cirad, l'IRD, AgroParisTech, l'IGN, l'Université Montpellier 2, le Cines, des laboratoires d'informatique, de calculs hautes performances, des startups). L'idée de cette infrastructure nationale est née du constat de sous-utilisation des données satellitaires par le domaine public, tant pour la recherche que pour les politiques publiques, et de l'échec de la stratégie *technology push* et du modèle commercial adopté. Les objectifs de ce consortium sont : (i) d'acquérir chaque année la couverture satellitaire du territoire français, ainsi que d'autres données satellitaires, et de les mettre à disposition gratuite pour les organismes publics ; (ii) de mettre à disposition des moyens logiciels, de calcul haute performance et de gestion de données ; (iii) de développer des méthodes et des algorithmes autour du traitement de l'image et de l'analyse spatiale ; (iv) de favoriser la mise en réseau et la formation des acteurs publics et privés intéressés par la problématique.

Le dispositif est déjà opérationnel et met à disposition différentes couches images dont la couverture satellitaire France 2010. En six mois (septembre 2011-février 2012)

[57] http://www.pfc.org.au/pub/Main/Data/TowardstheAustralianDataCommons.pdf
[58] http://services.ands.org.au/
[59] http://geosud.teledetection.fr/

une centaine d'entités publiques y ont adhéré. Le financement EquipEx obtenu permettra d'acquérir cinq années (2011-2015) de couverture satellitaire haute résolution du territoire national ainsi que de l'imagerie Pléiades très haute résolution. Pour cela, l'Irstea (anciennement Cemagref) signe des contrats juridiques de multilicence avec des fournisseurs comme Spot Image et Geosys. Ces deux entreprises privées exploitent les images obtenues respectivement par des satellites français et allemands. Au-delà des images proprement dites, l'Irstea sera également le responsable juridique de contrats multi-licences d'utilisation de logiciels propriétaires de traitement d'images. Ces contrats excluent les exploitations de nature commerciale qui doivent faire l'objet de négociations spécifiques avec les fournisseurs. Il s'agit là d'une pratique courante (*cf.* les licences multi-utilisateurs de Microsoft Office® dans les organismes publics) qui est appliquée ici. Des scénarios d'élargissement à des licences complètes (public, privé, exploitation commerciale) pour l'imagerie satellitaire sont à l'étude.

Le cas du projet GeoSud est particulier en ce sens qu'il ne requiert pas de grandes capacités de stockage. En effet, les volumétries utilisées sont actuellement de l'ordre de 1 Téraoctet, avec une perspective de croissance à 10 Téraoctets d'ici 2015. L'enjeu se situe en fait beaucoup plus dans l'exploitation des données, cette exploitation nécessitant l'emploi de moyens de calcul haute performance. C'est cette nécessité qui a rendu l'implication du Cines indispensable. Une autre originalité de ce projet réside dans une mutualisation inter-organisme des moyens et une animation du partage d'expérience et de la valorisation.

A.1.7. Expérience de partenariat en région

La Région Aquitaine a depuis plusieurs quadriennaux une politique de soutien à l'accès au calcul intensif et au stockage des données pour l'ensemble des laboratoires de recherche. Cette politique passe par une mutualisation *via* le Mésocentre de Calcul Intensif Aquitain (MCIA), qui est une structure au sein de laquelle se retrouve l'ensemble de la communauté scientifique, toutes disciplines confondues, dont l'Inra. Deux outils sont mis en place : (i) un cluster local offrant 110 millions d'heures de calcul à la communauté sur quatre ans ; (ii) un point d'entrée à la grille de calcul European Grid Infrastructure (EGI), dont certains composants donnent accès à 30000 CPU distribués en Europe. Les laboratoires Inra ont accès à cette ressource, sous réserve de conformité à la charte des utilisateurs, et l'utilisent actuellement sans saturation des besoins pour les domaines de la bioinformatique, de l'écologie et de l'évolution. L'accès est actuellement, et pour le quadriennal qui vient, gratuit.

Cette politique partenariale permet le développement de liens avec l'Institut des Grilles et du Cloud, (IdGC) *via* une association dans des dépôts de projets à l'ANR, avec notamment en perspective une connexion avec les structures européennes en Distributed Computing Infrastructures (DCI) pour la biodiversité : LifeWatch, dix projets préparatoires, et de proposer le développement à Bordeaux d'un partenariat triple Inra, MCIA, IdGC dans cette perspective intégrant la gestion et le partage des données en biodiversité.

A.2. QUESTIONNAIRES

A.2.1. Enquête sur la « Production des données à l'INRA »

Champs du questionnaire

CHAMPS	DESCRIPTIF	VALEURS
email	identité de l'interlocuteur	Texte libre
entite_de_production	'Structure' (Inra) à l'origine des données	inconnue département unité(UR-UE) centre équipe cati plateforme
identite_de_l'entite	Nom de cette structure	Texte libre
domaine_d'application		inconnu phénotypage animal phénotypage végétal phénotypage autre génotypage animal génotypage végétal génotypage autre environnement socio-éco alimentation-biotechno autre (voir commentaires)
nature_des_données		inconnue alpha-numériques images texte sequences autre(voir commentaires)
acquisition		manuelle automatisée
mise_à_jour		inconnue aucune 5 ans annuelle mensuelle hebdoOUjournalière
support_d'origine	Support physique	papier fichier(txt,tabl.) fichierSIG

		SGBD
accessibilite	Droit d'accès	inconnue producteur inra partenaires totale
volume_annuel_gigaoct_ou_m3papier	En GigaOctets (ou m3 pour papier)	Texte libre
évolution_du_volume_à_3_5_ans	A court-moyen terme (3-5 ans)	inconnue stable(=x1) x1000 x100 x10 /10 /100 /1000inconnu stable(=x1) x100 x10 /10 /100
pérennite		inconnue aucune temporaire permanente
gestion_stockage	Prise en charge de la gestion des données	inconnu dans la structure à l'INRA hors INRA
commentaires		Texte libre

A.2.2. Enquête accès aux données hors Inra

Champs du questionnaire

CHAMPS	DESCRIPTIF	VALEURS
email	identité de l'interlocuteur	Texte libre
source		externe partenaires
identite_du_fournisseur		Texte libre
domaine_d'application		inconnu phénotypage animal phénotypage végétal phénotypage autre génotypage animal génotypage végétal génotypage autre environnement socio-éco alimentation-biotechno autre(voir commentaires)
nature_des_donnees		inconnue alpha-numériques images texte sequences autre(voir commentaires)
support_dorigine	Support physique	papier fichier(txt,tabl.) fichierSIG SGBD
volume_en_gigaoctets_ou_m3papier		Texte libre
frequence_acquisition		une fois hebdomadaire mesuelle annuelle 5 ans
cout_en_keuros	En k€	Texte libre
utilisateur	'Structure' (inra) utilisant les données	inconnu département unité(UR-UE) centre équipe cati
identite_de_utilisateur	Nom de cette structure	Texte libre
transformation_des_donnees		aucune enfichier(txt,tabl.) enfichierSIG enSGBD
evolution_du_volume_a_3_5_ans	A court-moyen terme (3-5 ans)	inconnu stable(=x1)

		x1000 x100 x10 /10 /100 /1000
mise_à_jour		inconnue aucune 5 ans annuelle mensuelle hebdoOUjournalière
gestion_stockage	Prise en charge de la gestion des données	inconnu dans la structure à l'INRA hors INRA
commentaires		Texte libre

A.2.3. Enquête ingénieurs

QUESTIONNAIRE

« QU'EST-CE QUI EST EN TRAIN DE CHANGER ? »

VOTRE STATUT, SPÉCIALITÉ ET STRUCTURE DE RATTACHEMENT

AI ☐ IE ☐ IR ☐ CDD ☐ Autre ☐ précisez :

Bioinformatique ☐ Informatique ☐ Biostatistique ☐ Autre ☐ précisez :

Plate-forme (autonome) ☐ UR/UMR ☐ Équipe au sein d'une UR/UMR ☐

Date de recrutement :

Vos nom, prénom, adresse mél (facultatif) :

ÉVOLUTION DU CONTENU DE VOTRE ACTIVITÉ PROFESSIONNELLE

- Etes-vous concerné/e par l'avènement des nouvelles techniques à haut débit de production des données : oui ☐ non ☐

  Si oui, dans quel domaine :
  – « Omiques » ☐
  – Imagerie ☐
  – Données satellitaires ☐
  – Données climatiques ☐
  – Phénotypage ☐
  – Autre ☐, précisez :

- Votre pratique professionnelle a-t-elle changé du fait de l'avènement du haut débit : oui ☐ non ☐
  Si oui, depuis quand avez-vous constaté ce changement :

  Ce qui a changé :
  - Dans quel(s) domaine(s) scientifique avez-vous constaté un changement:
  - Avez-vous été confronté/e à des nouvelles questions scientifiques : oui ☐ non ☐
    Si oui, pouvez-vous citer des exemples :
  - Consacrez-vous plus de temps à une activité de veille technologique : oui ☐ non ☐
  - Avez-vous eu besoin de vous former : oui ☐non ☐
    Si oui, dans quel domaine :

    Quelles ont été les durées des formations suivies :

MODE DE FONCTIONNEMENT ET COLLABORATION

- Qui vous sollicite principalement : biologistes ☐ autres ☐, précisez :
- Qui sélectionne les projets sur lesquels vous travaillez :
  - Êtes-vous consulté/e au cours du processus décisionnel : oui ☐ non ☐ parfois ☐

Si parfois, précisez la proportion :
Si oui, à quel niveau intervenez-vous :

- Quels sont vos modes de collaboration et d'interaction :
  - Intégration dans des projets de recherche oui ☐ non ☐
    Si oui, tendance à la stabilité ☐, *à l'augmentation* ☐, *à la diminution* ☐
  - Participation à des prestations de service oui ☐ non ☐
    Si oui, tendance à la stabilité ☐, *à l'augmentation* ☐, *à la diminution* ☐
  - Participation d'ordre purement technique oui ☐ non ☐
    Si oui, tendance à la stabilité ☐, *à l'augmentation* ☐, *à la diminution* ☐
- Avez-vous constaté des changements :
  - Dans l'élaboration des projets avec les biologistes : oui ☐ non ☐
    Si oui, donnez quelques exemples :
  - Dans le nombre de projets traités : oui ☐ non ☐
    Si oui, donnez quelques exemples :
  - Dans les délais impartis pour l'obtention de résultats : oui ☐ non ☐
    Si oui, donnez quelques exemples :

ASPECTS CALCULATOIRES

Vous est-il nécessaire de faire appel à de gros moyens de calcul : oui ☐ non ☐
Si oui, précisez d'où viennent ces moyens (mésocentre local, grilles, gros centre de calcul (CCRT, IN2P3, IDRIS)) :

Si oui, le moyen de fonctionnement choisi vous satisfait-il : oui ☐ non ☐

Si non, que comptez-vous faire pour améliorer la situation :

REGROUPEMENT DES MOYENS

L'afflux de données se traduit souvent par la mise en place de plates-formes, c'est-à-dire le regroupement d'équipement et de compétences dans une problématique bien ciblée (bioinformatique, microscopie électronique, imagerie, etc.) qui sont ouvertes vers les utilisateurs extérieurs. Ainsi le traitement des données peut passer d'une vision « cas par cas » à une vision « haut débit » (pôle de compétence, mutualisation des expériences, ressources matérielles).

- Avez-vous constaté des éléments de ces évolutions : oui ☐ non ☐
  Si oui, qu'est-ce qui a changé en termes d'organisation du travail (recrutements, achat matériel, collaborations inter-équipes, pluridisciplinarité, etc.) :

  Si oui, qu'est-ce que le regroupement des moyens en plate-forme induit de changements dans la façon d'exercer votre métier :

QUESTIONS DIVERSES

- Quels changements anticipez-vous dans un horizon à cinq ans :
- Comment comptez-vous y faire face :
- Qu'attendez-vous de l'Inra? :

COMMENTAIRES LIBRES

A.2.4. Enquête chercheurs

QUESTIONNAIRE

« QU'EST-CE QUI EST EN TRAIN DE CHANGER ? »

VOTRE STATUT, SPÉCIALITÉ ET STRUCTURE DE RATTACHEMENT

CR ☐ DR ☐ Post-doc ☐ Autre ☐ précisez :

Biologie ☐ Bioinformatique ☐ Biostatistique ☐ Autre ☐ précisez :

Plate-forme (autonome) ☐ UR/UMR ☐ Équipe au sein d'une UR/UMR ☐

Date de recrutement :

Vos nom, prénom, adresse mél, (facultatif) :

ÉVOLUTION DU CONTENU DE VOTRE ACTIVITÉ PROFESSIONNELLE

- Etes-vous concerné/e par l'avènement des nouvelles techniques à haut débit de production des données : oui ☐ non ☐
  Si oui, dans quel domaine :
  – « Omiques » ☐
  – Imagerie ☐
  – Données satellitaires ☐
  – Données climatiques ☐
  – Phénotypage ☐
  – Autre ☐, précisez :
- Votre pratique professionnelle a-t-elle changé du fait de l'avènement du haut débit : oui ☐ non ☐
  Si oui, depuis quand avez-vous constaté ce changement :

  Qu'est-ce qui a changé :
  - Dans quel(s) domaine(s) :
  - Avez-vous été confronté/e à des nouvelles questions scientifiques : oui ☐ non ☐

    Si oui, pouvez-vous citer des exemples :
  - Avez-vous eu besoin de vous former : oui ☐ non ☐

    Si oui, dans quel domaine :

    Quelles ont été les durées des formations suivies ?

MODE DE FONCTIONNEMENT ET COLLABORATION

- Quels sont vos modes de collaboration et d'interaction :
  - Intégration d'ingénieurs dans vos projets de recherche oui ☐ non ☐
    Si oui, quelle aide vous apportent-ils ?

- Demandes faites à des prestataires externes oui ☐ non ☐

- Avez-vous constaté des changements :
  - Dans le nombre de projets traités : oui ☐ non ☐
    Si oui, donnez quelques exemples :
  - Dans les délais impartis pour l'obtention de résultats : oui ☐ non ☐
    Si oui, donnez quelques exemples :

ASPECTS CALCULATOIRES

Vous est-il nécessaire de faire appel à de gros moyens de calcul : oui ☐ non ☐
Si oui, précisez d'où viennent ces moyens (mésocentre local, grille, gros centre de calcul (CCRT, IN2P3, IDRIS)) :

Si oui, êtes-vous satisfait ? : oui ☐ non ☐

Si non, pourquoi :

Que comptez-vous faire pour améliorer la situation :

REGROUPEMENT DES MOYENS

L'afflux de données se traduit souvent par la mise en place de plates-formes, c'est-à-dire le regroupement d'équipement et de compétences dans une problématique bien ciblée (bioinformatique, microscopie électronique, imagerie, etc.) qui sont ouvertes vers les utilisateurs extérieurs. Ainsi le traitement des données peut passer d'une vision « cas par cas » à une vision « haut débit » (pôle de compétence, mutualisation des expériences, ressources matérielles).

- Avez-vous utilisé de telles plates-formes récemment : oui ☐ non ☐
  Si oui, qu'est-ce que le regroupement des moyens en plate-forme a induit comme changements dans la façon d'élaborer vos projets de recherche :

QUESTIONS DIVERSES

- Quels changements anticipez-vous à cinq ans :
- Comment comptez-vous y faire face :

COMMENTAIRES LIBRES

A.3. DIRECTIVE INSPIRE

La directive Inspire a été publiée le 14 mars 2007. Elle impose aux autorités publiques (l'État, les collectivités territoriales et leurs groupements, les établissements publics) de publier sur Internet leurs informations géographiques. Son objectif général est d'établir une infrastructure d'information géographique dans la Communauté européenne pour favoriser la protection de l'environnement. Inspire doit s'appuyer sur les infrastructures d'information géographique établies et exploitées par les États membres, et notamment pour la France sur le Géoportail mis en œuvre par l'IGN et le BRGM.

Le champ d'application

La directive ne s'applique qu'aux données existantes, détenues par une autorité publique ou un tiers qui a accès à cette infrastructure Inspire, et qui sont relatives à l'environnement et à l'un des 34 thèmes mentionnés dans les trois annexes de la directive. Ces annexes correspondent à un ordre de priorité, l'annexe I devant être traitée le plus rapidement.

Le partage des données

La directive impose aux autorités publiques :

- De partager leurs séries et services de données géographiques
- De remédier au cloisonnement qui peut exister entre les différents niveaux d'autorités publiques.
- Un niveau de qualité des données qui doit être indiqué de façon précise.

Les données concernées par la directive Inspire

Annexe I :

1. Référentiels de coordonnées
2. Systèmes de maillage géographique
3. Dénominations géographiques
4. Unités administratives
5. Adresses
6. Parcelles cadastrales
7. Réseaux de transport
8. Hydrographie
9. Sites protégés

Annexe II :

1. Altitude
2. Occupation des terres
3. Ortho-imagerie
4. Géologie

Annexe III :

1. Unités statistiques
2. Bâtiments
3. Sols
4. Usage des sols
5. Santé et sécurité des personnes
6. Services d'utilité publique et services publics
7. Installations de suivi environnemental
8. Lieux de production et sites industriels
9. Installations agricoles et aquacoles
10. Répartition de la population, démographie
11. Zones de gestion, de restriction ou de réglementation et unités de déclaration
12. Zones à risque naturel
13. Conditions atmosphériques
14. Caractéristiques géographiques météorologiques
15. Caractéristiques géographiques océanographiques
16. Régions maritimes
17. Régions biogéographiques
18. Habitats et biotopes
19. Répartition des espèces
20. Sources d'énergie
21. Ressources minérales

A.4. L'AMÉLIORATION GÉNÉTIQUE DES BOVINS : COLLECTE ET GESTION DES DONNÉES À L'ÉCHELLE D'UNE INTERPROFESSION

En France, l'amélioration génétique des bovins s'effectue sur un modèle public reposant sur une collaboration directe des partenaires de l'interprofession. Il en est de même pour l'amélioration des ovins et caprins. En revanche celle des porcs relève du secteur privé.

Par la création de l'interprofession « France Génétique Elevage » (FGE), la Loi d'Orientation Agricole (LOA) de 2006 a confié la responsabilité du dispositif d'amélioration génétique des ruminants aux propres acteurs de la filière. L'organisation permet d'offrir aux éleveurs les services et moyens nécessaires à la sélection de leur troupeau. FGE rassemble les organismes représentatifs des éleveurs et des acteurs de la sélection : identification des animaux et élevages, performances, généalogiques, sélection et diffusion des reproducteurs par insémination ou monte naturelle, gestion des données et évaluation génétique.

La collecte, la gestion et le traitement des données s'effectuent de la manière suivante :

1. Au niveau départemental, rassemblement des données par les établissements départementaux d'élevage (EDE) et transmission à l'échelon régional.
2. Au niveau régional, 7 centres, les ARSOE (Association Régionale de Services aux Organismes d'élevage), regroupent et traitent les informations : déclaration de naissance fournie par les éleveurs, performances de production par les conseillers techniques et qualité du lait par les laboratoires d'analyse.
3. Au niveau national, le Centre Informatique du département de Génétique Animale (CTIG, INRA Jouy) rassemble les données provenant des ARSOE, gère le fichier des généalogies et produit les index de sélection qui sont retournés aux ARSOE.

A titre d'exemple les données gérées par l'ARSOE de Bretagne, la première région d'élevage de France correspondent à 22% de la production laitière et les cartes d'identité des animaux d'élevage sont établies pour plus de 2,5 millions de têtes présents chez plus de 45,000 éleveurs. Les tables gérées dans la base de données de l'ARSOE de Bretagne comptent quelques dizaines de millions d'enregistrements ; celles gérées par l'INRA quelques centaines de millions.

Le partage des données au sein d'un réseau à la fois très large mais aussi très structuré (en étoile) est une des premières raisons du succès de ce modèle. Il est à noter que la propriété des données fait encore l'objet de discussion.

A.5. TABLE DES SIGLES

ANR : Agence Nationale de la Recherche

CATI : Centre Automatisé de Traitement de l'Information

CEA : Commissariat à l'Energie Atomique et aux Energies Alternatives

CEMAGREF : Centre d'Etudes du Machinisme Agricole et des Eaux et Forêts

CERN : Organisation Européenne de Recherche Nucléaire

CIRAD: Centre de Coopération Internationale en Recherche Agronomique pour le Développement.

CNRS : Centre National de la Recherche Scientifique

DGRI: Direction Générale de la Recherche et de l'Innovation au MESR

ENS : Ecole Normale Supérieure

GIEC : Groupe d'experts intergouvernemental sur l'évolution du climat

IN2P3 : Institut National de Physique Nucléaire et de Physique des Particules

INED : Institut National des Etudes Démographiques

INIST : Institut de l'Information Scientifique et Technique du CNRS

INRA : Institut National de la Recherche Agronomique

INRIA : Institut National de Recherche en Informatique et Automatique

INSEE : Institut National de la Statistique et des Etudes Economiques

INSERM : Institut National de la Santé et de la Recherche Médicale

IGN : Institut national de l'information géographique et forestière

IRD : Institut de Recherche pour le Développement

IRSTEA : Institut national de recherche en sciences et technologies pour l'environnement et l'agriculture

MESR : Ministère de l'Enseignement Supérieur et de la Recherche

MNHN : Muséum National d'Histoire Naturelle

PEPI : Partage d'Expériences et de Pratiques en Informatique

Conseil Scientifique de l'Inra
147, rue de l'Université
75338 Paris cedex 07
Tél : 01 42 75 90 00
conseil.scientifique@paris.inra.fr