Infrastructures, systèmes d'observation et services d'information environnementale et climatique

Compte-rendu du séminaire

Les données accessibles dans le domaine environnemental.

Quels freins ?

Mercredi 21 mai 2014 - 10h > 18 h

Muséum national d'histoire naturelle
amphithéâtre de la Grande Galerie de l'Evolution
36 rue Geoffroy St Hilaire, 75005 PARIS

Co-organisé par **Michel Guiraud** et **Cécile Callou** (MNHN)

Cadre législatif global commun

Intervenant : Marc Leobet, Mission de l'information géographique, Commissariat général au développement durable / Direction de la recherche et de l'innovation (MEDDE)

Introduction
Le triangle juridique de la donnée géographique environnementale

Le triangle juridique est constitué de la directive dite PSI (*Public Sector Information directive*, 2003/98/CE du 17 novembre 2003), sur la réutilisation libre des données publiques, de la convention d'Aarhus (signée en juin 1998), sur l'accès à l'information environnementale, au sens extrêmement large et de la directive INSPIRE (2007/2/CE du 14 mars 2007), sur le partage de l'information géographique publique entre autorités publiques. En tant que membres d'institutions publiques financées par le contribuable, qui fabriquent et diffusent de la donnée, et qui par ailleurs, travaillent dans le champ de l'environnement, les trois directives européennes s'appliquent.

PSI, réutilisation des données du secteur publique. En théorie, une personne peut demander la totalité des données dont l'on dispose. Il existe quelques exceptions pour un certain nombre de données, par exemple celles ayant trait à la sécurité publique ou la protection de l'environnement. Il n'existe aucun moyen de contrôler l'usage fait *a posteriori* des données, ce qui est une difficulté souvent soulevée par les producteurs. Il est possible pour une autorité publique de faire payer les données, sauf que le gouvernement a décidé d'encadrer strictement cette possibilité pour l'Etat et ses établissements publics. Une circulaire du Premier Ministre indique que les données qui étaient payantes avant le 1er juillet 2011 peuvent continuer à l'être ; en revanche, après le 1er juillet 2011, une donnée ne peut être payante que si elle est sur une liste établie par un décret du Premier Ministre. Or, il n'y en a à ce jour jamais eu. Il est donc théoriquement impossible de faire payer l'accès à de nouvelles données depuis le 1er juillet 2011. L'aspect le plus notable de la directive PSI est l'argumentaire juridique sur lequel les derniers gouvernements se sont appuyés pour établir la politique d'*open data*.

Convention d'Aarhus. Contrairement à PSI et INSPIRE, directives européennes, il s'agit d'abord d'une convention internationale. Dans ce cadre, toute décision environnementale doit être faite après recueil de l'avis des citoyens. Ainsi, le conseil constitutionnel a annulé une loi sur des schémas régionaux de l'énergie au motif que l'enquête publique s'est déroulée sur un délai trop court pour que ceux-ci puissent valablement participer. Pour que le citoyen prenne une décision ou donne un avis de façon pleinement responsable, il faut bien évidemment qu'il ait le même niveau d'information que les décideurs publics sur tout ce qui relève de

l'environnement au sens très large : état de l'environnement, son passé, son futur, sur tous les milieux, sur l'humain, etc... Depuis 2005, les acteurs publics (donc nous) doivent réaliser des catalogues en ligne des données qu'ils possèdent intéressant l'environnement. Dans les faits, personne ne l'a fait, et c'est sans doute une des raisons qui conduit à la priorité des métadonnées dans la directive INSPIRE. La convention d'Aarhus va très loin car elle dit que, dans un certain nombre de cas, pour tout ce qui traite des émissions dans les milieux, la propriété intellectuelle, le secret défense, ne peuvent pas être opposés. L'information concernant la quantité d'effluents radioactifs déversés dans la rade de Toulon par les sous-marins de la flotte nucléaire française a ainsi pu être obtenue par le Ministère de l'Environnement. La loi souligne que le responsable public doit peser l'intérêt de la diffusion publique par rapport aux différents intérêts particuliers protégés par la loi. Il n'existe pas encore de jurisprudence sur ce point.

Directive INSPIRE. Le périmètre sur laquelle elle s'exerce est plus restreint, mais son action est plus « invasive ». Si les deux premières sont peu exploitées car méconnues, le Parlement européen a choisi d'aller plus loin par la mise en ligne sur Internet. Dans le cas des autres textes, il faut faire une demande et donc savoir que les données existent, la directive INSPIRE impose la diffusion en ligne des données, avec des structures de données normalisées etc... Les restrictions d'accès public, de propriétés intellectuelles, de secrets statistiques sont les même que pour PSI. En revanche, entre autorités publiques, la liste des restrictions au partage entre autorités publiques est plus limitée, ce qui ne va pas sans poser quelques questions. INSPIRE respecte les modèles économiques en cours et permet de faire payer les données.

Les restrictions d'accès public. Il ne faut pas confondre « sécurité publique », c'est-à-dire Vigipirate (plan de protection des points d'intérêts vitaux pour la nation), qui entraine une vraie restriction et « sécurité publique » dans le sens adopté par certain comme tranquillité publique pour les décideurs ! Les droits de propriété intellectuelle sont évidemment protégés par la Convention de Berne, mais il s'agit du droit de propriété intellectuelle des tiers puisque les droits de propriété intellectuelle des autorités publiques ne peuvent pas s'opposer à une loi qui les oblige spécifiquement de les diffuser.

La confidentialité des informations commerciales ou industrielles apparaît dès qu'il y a des travaux en commun avec des entreprises privées ou des entreprises de type EDF qui sont sous statut privé ; le droit de propriété des tiers parties-prenantes dans une co-production est protégé, et il y a bien évidemment une restriction normale de diffusion des données qui appartiennent en commun aux chercheurs et à ces entreprises. Tout ce qui se rapporte à la loi Informatique et Libertés, le secret statistique, la protection de l'environnement est également protégé. Cependant, ces raisons ne s'opposent pas à la diffusion des catalogues de données sur Internet ; s'il existe d'authentiques restrictions d'accès public, il faudra préciser, sous couvert de l'article concerné, "pourquoi vous ne pouvez pas avoir accès à ces informations". Il reste que, dans notre domaine, les métadonnées en ligne sont obligatoires.

Des subtilités juridiques sources de blocage. Pour rappel, la propriété des agents de l'Etat n'existe pas en tant que telle dans le cadre de leur fonction, sauf pour les chercheurs, qui reste propriétaires de ce qu'ils produisent. Dans les faits, les données d'un laboratoire de recherche sont considérées comme des données de chercheurs, qu'elles soient produites par des chercheurs ou des techniciens/ingénieurs.

Il existe également le droit voisin sur les structures de données ; beaucoup de données sont aujourd'hui payantes parce qu'il y a un droit sur les producteurs de base de données. Cela soulève la question du devenir de ce droit avec l'usage de services de transformation automatique qui transformeront la structure de la donnée. Il n'existe pas encore de jurisprudence permettant de régler cette importante question.

Ce qu'il faut retenir, c'est qu'il existe un certain nombre de solutions et de garde-fous juridiques pour permettre la conservation de patrimoine d'informations, payées avec l'argent du contribuable, et éviter de refaire plusieurs fois la même chose.

Enfin, pour relativiser ces questions, un comptage sur les données du géo-catalogue réalisé en 2012 a montré que la donnée publique aujourd'hui est une donnée libre et gratuite, sauf dans quelques très rares cas.

La politique d'ouverture des données du gouvernement. Par décision des gouvernements successifs, c'est simple, toutes les données publiques doivent être gratuites. Le Président de la République a par ailleurs signé en décembre 2013 la Charte du G8 sur l'open data, et des décisions ont été prises dans les réunions du Comité interministériel pour la modernisation de l'action publique (Cimap).

La directive PSI a été révisée en 2013. Il se trouve que la France était allée beaucoup plus loin que ce que la 1ère version de la directive européenne imposait, si bien que la directive PSI actuelle ne changera pratiquement rien par rapport à nos habitudes françaises. Un projet de loi devrait cependant être discuté (annoncé par la secrétaire d'Etat au numérique, puis par Mme Lebranchu dans le cadre de la décentralisation), qui pourrait imposer l'ouverture des données, donc la gratuité et l'accès libre aux données à tout le monde, des établissements publics et Ministères mais aussi des collectivités de plus de trois mille habitants.

Les évolutions annoncées. Jusqu'à maintenant, il y avait obligation de partage avec la possibilité de recettes. La décision du Cimap va bien évidemment poser un certain nombre de problèmes, notamment aux trois établissements publics visés en particulier (IGN, Météo France et SHOM), parce qu'ils ont des recettes conséquentes liées à la diffusion de licences payantes. La gratuité imposée risque d'avoir des conséquences graves sur ces établissements. Si les données de la recherche ne sont pas explicitement citées dans le Cimap, les services du Premier Ministre suivent également cette question. La seule nuance entre les directives PSI 2003 et 2013 est l'extension des données publiques « aux documents détenus par les

établissements d'enseignement et de recherche, y compris les organisations créées pour le transfert des résultats de la recherche des écoles et des universités, à l'exception des bibliothèques universitaires ». Cela a soulevé des discussions au niveau européen et des Etats, dont la France, ont cherché à bloquer l'initiative facilitant l'accès aux données de la recherche. En 2003, la Commission européenne n'avait pas réussi à donner l'accès à ces données, mais certain nombre d'Etats soutenaient néanmoins cette demande. En 2013, le centre de gravité s'est déplacé. En 2023, pour la révision suivante, il faut probablement imaginer que les données de la recherche seront de plus en plus intégrées dans cette grande tendance à l'ouverture et au partage libre et gratuit des différentes données.

Conclusion. La position du Ministère est d'avoir les moyens de continuer à avoir des suivis temporels sur des très longues périodes pour être capable de remplir nos missions de préservation de l'environnement. Il y a donc un fort enjeu de préservations des données de référence.

Il y a également la question de la monétisation des données : dans certains cas, comme les permis de construire, la vente des données, même à faible prix, permet de financer des gens pour contrôler la qualité de la donnée. Cela entre dans la chaîne de production de nos services de l'observation statistique. La question de la valorisation économique fait débat : comment fait-on en sorte que les moyens de production de la donnée restent pérennes, à un bon niveau, pour permettre l'efficacité des services publics ? Autres problèmes qui touchent aussi les données de la recherche, l'utilisation croissante de la force de collecte ou d'identification des données par le citoyen. L'appel au *crowdsourcing* pose la question de la propriété intellectuelle et sur les pratiques qu'on pourra avoir vis-à-vis de la conservation des données obtenues de cette manière. Le côté positif est que cela oblige à réfléchir sur nos pratiques, et donc à les faire évoluer.

Puisque le délai de mise à niveau des modèles économiques des établissements est de cinq ans, la question est sans doute désormais de savoir ce qu'on pourra faire payer dans cinq ans. La réponse semble être le service, point. C'est-à-dire la valorisation que nous serons capables d'apporter à la donnée vis-à-vis des services publics ou de nos concitoyens.

Discussions

Non présentée : *Directrice d'une infrastructure de recherche, basée sur des données ouvertes, gratuites, en ligne, je n'ai pas de souci avec l'ouverture de données. Par contre, il existe une tendance lourde sur les indicateurs de mesures de nos infrastructures de recherche, qui sont basées sur le nombre de brevets, sur l'utilisation de savoirs et très précisément sur l'utilisation des données. Nous n'avons pas d'indicateurs corrects, acceptables. La réelle ouverture des données nous met en difficulté. Donc une tendance lourde devrait être suivie d'une autre*

tendance lourde qui serait de développer les indicateurs qui vont avec.

M. Leobet : *Absolument d'accord. Des collègues danois ont fait un gros travail d'ouverture de leurs données, les adresses, les numéros SIREN etc... et leur problème maintenant est de savoir "où sont mes clients/usagers ?". Leur problème est assez basique. Comme ils diffusent tout gratuitement, sans aucun frein, ils sont en difficulté pour défendre leur budget vis-à-vis de leur Ministre. Et les problèmes qu'on a dans nos infrastructures de diffusion de données ouvertes, gratuites etc... c'est en effet que si on n'est pas capable de dire à nos financeurs, à nos budgétaires, à quoi sert l'argent dans nos serveurs, dans nos ingénieurs informaticiens etc... on court le risque d'avoir une érosion des budgets et de mettre en danger, non pas seulement la diffusion de la donnée, mais l'infrastructure de gestion, y compris pour l'interne.*

Ceci dit, je pense que tout le monde détesterait qu'on ait une batterie d'indicateurs décidés par le Ministère de la Recherche, tout seul, ou par le Ministère de l'Environnement, tout seul. La façon dont je vois les choses, c'est qu'à un moment vous devez bâtir les indicateurs qui correspondent à votre métier et mettre en place par exemple les outils statistiques sur vos serveurs, la consommation de l'infrastructure, le nombre de données consommées, téléchargées., etc.. C'est à vous de voir ce qui est intéressant pour vous, mais je pense qu'on ne peut pas se permettre d'ouvrir tous nos placards sans à un moment être capable de dire à quoi et éventuellement à qui ça sert.

J.-P. Le Duc (MNHN) : *Je voulais intervenir sur deux points : 1°. Dans le cadre de la convention d'Aarhus, un certain nombre d'organismes, qui répondaient "cher monsieur, je ne peux pas vous communiquer cette donnée pour x ou x raison", répondent aujourd'hui "la donnée n'existe pas". Pour y avoir accès, il faut donc d'abord démontrer que la donnée existe, ce qui est quand même un obstacle et une tendance assez sidérante parce qu'elle vient souvent du secteur public de l'Etat ; 2°) Vous avez posé fort justement le problème des moyens pour faire vivre les banques de données au départ. Une très forte tendance actuelle - particulièrement au niveau européen mais aussi au niveau mondial – est de financer non plus les banques de données, mais des banques de banques de données, ou des banques de banques de banques de données. Il y a actuellement un certain nombre de gros projets, en particulier dans le domaine de la biodiversité, qui ne vivent qu'avec des données qui viennent de banques de données qu'on rend publiques, gérées par des organismes et bénéficiant de financement public. Il est aujourd'hui plus facile de se faire financer l'utilisation d'un système qui utilise des banques de données plutôt que de faire fonctionner les banques de données au départ pour continuer à récolter l'information et sa disponibilité.*

M. Leobet : *Deux réponses courtes. Ce qui m'intéresse dans l'utilisation d'INSPIRE faite par l'UMS BBEES (CNRS INEE-MNHN), c'est que ce n'est pas pour l'application simple d'une loi mais parce qu'un problème de gestion collective du patrimoine des données de la recherche a été identifié. Pas pour la diffusion ou l'open data, mais parce que quand un chercheur part à*

la retraite ou change de poste, les données restent bien souvent sur un disque dur et ne sont plus valorisables. J'ai le même discours vis-à-vis des chefs de service du Ministère : le catalogue des données est un outil de préservation du Patrimoine, comme il existe des listes d'ordinateurs qui sont dans vos laboratoires parce qu'il faut rendre compte ; mais les données ne sont pas gérées. Or, ce sont ces catalogues qui vont nous permettre de rendre minoritaire la pratique disant "non je n'ai pas".

Sur le deuxième élément, nous avons un travail vis-à-vis de la Commission Européenne, pour faire prendre en compte la gestion initiale des bases de données et même la création de bases de données, dans les appels de recherche des PCRD par exemple. Nous n'avons pas eu jusqu'à présent un grand succès, même si on a commencé à ouvrir une petite brèche, mais clairement, des lobbys font que les budgets européens sont plutôt sur des gros outils informatiques et pas sur le financement de la base de la base qui permet après d'alimenter les grandes macro-bases dont vous parlez.

G. Tallec (Irstea): *Quelle est la responsabilité légale des établissements publics vis-à-vis de ces données qu'ils doivent laisser en accès libre ?*

M. Leobet : *C'est une question récurrente, la réponse est parfaitement claire. Il n'y a aucune responsabilité, de la part des producteurs, sur une mauvaise utilisation de quoi que ce soit. Vous avez un navigateur GPS dans votre voiture, vous prenez un sens interdit parce qu'il vous a dit de tourner à gauche alors qu'il y a le panneau, c'est votre problème, ce n'est pas celui du producteur de la base de données du navigateur.*

Construction, maintenance, archivage et sauvegarde

Certaines grandes bases de données sont directement et abondamment exploitées par des acteurs privés ou public à des fins d'expertise et de besoins économiques. Quelles structures et contraintes s'appliquent pour ces bases de données ? Comment se partagent les coûts de construction et de maintenance, et dans quelle mesure la valorisation économique des données permet-elle de contribuer à ces coûts ?

Modérateur : Nicolas Arnaud, Institut national des sciences de l'Univers (CNRS-INSU)
Intervenants : Philippe Santoni, Météo France
Marie-Louise Zambon, Institut national de l'information géographique et forestière (IGN)
Nathalie Leidinger, Service hydrographique et océanographique de la marine (SHOM)
Présentation : table ronde 1.pdf

Le coût du cycle de vie de la donnée, par N. Arnaud

Derrière les questions de savoir comment on construit, on maintient, on archive, on sauvegarde se cache un mot-clef : le coût. Trois représentants institutionnels vont vous présenter leur vision à travers leur expérience d'organismes et d'institutions particulières sur cette question du cycle de vie de la donnée, des coûts et du partage actuel des coûts entre un certain nombre d'acteurs publics et puis de la redevance.
Collectivement, nous vous proposons trois axes essentiels de réflexion sur lesquels nous débattrons à la suite de leur présentation :
1°. Qu'est-ce qui peut être éventuellement mutualisé par l'intermédiaire des synergies, entre organismes, entre utilisateurs sur ce cycle de vie. De nouveaux acteurs non-institutionnels interviennent sur ce champ de la donnée, et reposent la question des coûts sous un éclairage différent. Lorsqu'on parle du cycle de vie, la question de la part de ce coût assumé par la puissance publique, mais aussi par l'apport de la redevance se pose. Ce point est largement évoqué dans le rapport Trojette, sur l'ouverture des données publiques, publié en juillet 2013. Au-delà même de la production, de la diffusion de la donnée, l'expertise des producteurs, qui est garante de l'exploitation à long terme, est rarement prise en compte dans le cycle de vie. On pourra s'interroger sur le coût, en séparant la notion de données et de services ;
2°. Que peut-on mutualiser, quelles sont les économies d'échelle qui existent déjà ?
3°. Enfin, la question des nouveaux acteurs. De nombreux producteurs non académiques et non institutionnels de données, la science participative au sens large, permettent un

accroissement des corpus de données beaucoup plus important sans doute que ce qu'individuellement les producteurs institutionnels pourraient produire. Mais se pose immédiatement derrière le coût, la notion de qualité ; garantir cette qualité a également un coût qu'il convient d'évaluer.

Météo France, par Ph. Santoni

Météo France est un établissement public de l'Etat, dont la mission principale est d'exercer les attributions de l'Etat en matière de sécurité météorologique des personnes et des biens. Parmi les autres missions, il y a l'observation du temps présent, la climatologie qui concerne le passé – avec des contrôles climatologiques qui sont effectués sur les données de l'observation- et les prévisions qui vont de quelques minutes à quelques mois, voire quelques trimestres, et enfin la recherche. De ce fait, il y a trois grands types d'utilisateurs : les clients institutionnels qui sont chargés de la sécurité des personnes et des biens, le secteur aéronautique (comme tous les Etats du monde, la France est engagée à fournir aux avions qui survolent le territoire des informations météorologiques) et enfin, le secteur économique dans son ensemble et le grand public.

Quel est le cycle de vie de l'information ? Comment arrive-t-elle jusqu'au public ? On la collecte par des moyens très différents, que sont les satellites, les bouées, les petites stations blanches qu'on voit au bord des routes, etc... Ensuite, il faut pouvoir à partir de ces données fabriquer d'autres données, notamment les données de prévisions, grâce à des modèles mathématiques. L'information contenue dans les bases de données est présentée via différents portails. Chacune de ces quatre phases représente un coût élevé (estimation coût annuelle : quarante millions d'euros), et la subvention de l'Etat ne couvre qu'une petite partie du budget (environ deux millions d'euros). Pour le secteur de la recherche, Météo France accorde la gratuité sous réserve des frais de mise à disposition, qui ne sont pas très élevés. Pour les autres acteurs, la position défendue par l'établissement aujourd'hui est celle de la participation raisonnable des usagers au coût de production et de diffusion, d'où la redevance demandée.

Selon Steward Brand, dont le propos est souvent tronqué ("l'information doit être gratuite"), le paradoxe est plus large : la phrase exacte est « d'un côté, l'information veut être payante parce qu'elle a une grande valeur, la bonne information au bon endroit change simplement votre vie, et de l'autre, l'information veut être gratuite parce que son coût d'obtention diminue constamment ». Face à ce paradoxe, plusieurs attitudes sont possibles. Ce qui a été choisi par Météo France, c'est cette participation limitée des usagers au coût de production.

SHOM, par N. Leidinger

La mission du SHOM est de connaître et de décrire l'environnement marin dans ses relations avec l'atmosphère, avec les fonds marins et les zones littorales, et d'en prévoir l'évolution. Il assure aussi la diffusion des informations correspondantes et met à disposition les données. Le

SHOM est un EPA depuis 2007, sous tutelle du Ministère de la Défense. Le budget est de 56,7 millions d'euros. Nous avons des recettes commerciales à hauteur de 4,5 millions d'euros, dont 1,7 millions d'euros de redevances sur les données, redevances liées à la vente de licence, soit à peu près 3% des ressources du SHOM.

Pour acquérir les données, de lourds moyens à la mer sont mis en œuvre : levés bathymétriques, levés LIDAR, observations de hauteur d'eau par installation de marégraphes, collecte des données externes. Les données sont ensuite qualifiées, validées et gérées au sein de bases de données. L'exploitation se fait en interne pour fabriquer des produits, qui permettent de répondre à nos trois missions : 1°. assurer la sécurité de la navigation, 2°. soutenir la Défense et 3°. soutien aux politiques publiques maritimes et littorales. Le SHOM a également pour mission de conserver ce patrimoine de données, donc de constituer des archives pérennes exploitables et de les diffuser.

Un portail a été mis en service en 2013 (http://data.shom.fr/), totalement conforme aux exigences de la directive INSPIRE et qui permet d'accéder aux données à travers des services, du téléchargement etc... Certaines données sont disponibles en *open data*, mais d'autres sont disponibles sous forme de licence et donc payantes.

Du côté du coût, l'ordre de grandeur annuel est le suivant : la flotte coûte annuellement 26,6 millions d'euros (coût imputable actuellement à la Marine, hors budget du SHOM), le processus d'acquisition dans le budget du SHOM coûte 14,2 millions d'euros (personnel, instrumentation scientifique et part des projets) ; la gestion, le traitement des bases de données coûte environ 3,2 millions d'euros ; la fabrication des produits et services, 25,5 millions d'euros ; il est difficile d'évaluer aujourd'hui le coût de gestion des archives, car peu numérisé au regard de l'existant (trois cents ans de données). Un gros chantier a été entrepris pour moderniser cette infrastructure, et permettre de pérenniser ce patrimoine de données. Enfin, le coût de l'infrastructure de la plateforme de diffusion est de 0,25 millions d'euros/an. Le pôle le plus important est donc le processus d'acquisition.

Quelques pistes de réflexion : la valeur d'une information s'évalue parce qu'on la trouve et qu'on peut l'exploiter. Le défi majeur à relever est que les technologies sont en constante évolution, alors que les budgets sont en constante diminution. On constate aussi que la valeur des réseaux d'acquisition est insuffisamment reconnue. Si on trouve assez facilement des financements pour les portails, ce flux de financement remonte très rarement jusqu'au système de collecte, qui est très coûteux dans le cas du SHOM. On constate que l'investissement est financé au moment de la construction d'un projet, mais pas le fonctionnement. On a beaucoup de difficulté pour passer de la phase R&D à une phase "opérationnelle" pour un certain nombre de système.

IGN, par Marie-Louise Zambon

L'IGN a connu une réorganisation en 2013, regroupant au sein d'une même direction à la fois les agents chargés de faire du commerce et ceux chargés de la mission de service public. L'IGN est un établissement public à caractère administratif, sous double tutelle du Ministère du Développement Durable et du Ministère chargé de la Forêt depuis sa fusion avec l'Inventaire Forestier National en 2012. Sa vocation est de décrire, d'un point de vue géométrique et physique, la surface du territoire national et l'occupation de son sol, d'élaborer et de mettre à jour l'inventaire permanent des ressources forestières. Les missions sont la description du territoire de façon neutre, la partie du territoire couverte par la forêt, puis de faire toutes les représentations appropriées, archiver et diffuser les informations correspondantes et administrer le portail de diffusion INSPIRE de l'Etat (Géoportail).

Le cycle d'acquisition comprend plusieurs sources : prises de vue aériennes, acquisitions LIDAR, images satellites, actions d'acquisitions complémentaires de terrain comme la photogrammétrie, pour avoir une description géométrique précise. On s'appuie sur des remontées collaboratives pour mettre à jour nos bases de données, avec une politique partenariale pour s'associer à d'autres acteurs publics qui détiennent des informations mutualisables. Les bases de données sont gérées en interne, pour en faire des produits qui sont diffusés sous forme de bases de données ou de cartes ; elles sont diffusées sous forme papier, au travers du Géoportail, ou bien de manière numérique au travers de téléchargement.

Quelques chiffres et étapes-clefs : 1738 agents pour un budget de l'ordre de 154 millions d'euros. Nous devons rechercher environ 35% de ressources, puisque la subvention ne couvre pas la totalité du budget (cf rapport d'activité sur le site de l'IGN). Le chiffre d'affaires grand public s'élève à 12 millions d'euros et le marché professionnel 16 millions d'euros. Dans le cadre du Géoportail, infrastructure lourde portée par l'IGN, nous avons 3 millions de visites par mois. On parlait d'indicateurs de suivi de ces infrastructures, on peut distinguer un usage grand public, c'est-à-dire ceux qui viennent consulter sur le site, de ceux qui utilisent l'infrastructure au travers d'une API, pour eux-mêmes développer des infrastructures ou des sites qui s'appuient sur nos données et sont ensuite mis à disposition. On constate une augmentation de la fréquentation du Géoportail au travers d'autres sites qui s'appuient sur cette infrastructure. En termes d'orientation stratégique, l'IGN mène depuis plusieurs années des actions d'ouverture. En 2009, toutes les données de l'IGN ont été mises à disposition gratuitement dans la sphère enseignement et recherche, hormis les données qui sont coéditées avec des tiers. En 2011, toutes les données du référentiel à grande échelle ont été mises à disposition au sein de la sphère publique pour un usage de mission de service public. Dans ce cadre, des négociations ont été menées avec les tutelles à la fois ministérielles et budgétaires pour abonder les pertes de recettes en contrepartie. Avec le passage à la gratuité, il a été

constaté que le volume de données téléchargées a été multiplié par vingt, sans que l'on sache précisément quel usage en est fait.

Par ailleurs, entre 2012 et 2013, des chantiers stratégiques ont été conduits pour redéfinir les orientations, parce que nous sommes dans un contexte en forte évolution avec le déploiement des mobiles, des smartphones, des données gratuites facilement accessibles. L'IGN se recentre donc sur sa mission de service public et a travaillé sur la définition d'un nouveau modèle économique, sur une simplification de la tarification, pour essayer de mettre en place des produits et des services qui sont à un premier niveau gratuits pour tous, à un deuxième niveau gratuits pour la sphère publique et des prestations, services ou données qui sont payantes, avec aussi la volonté de développer plus de services s'appuyant sur la plateforme de diffusion du Géoportail.

Discussions

N. Arnaud (CNRS-INSU) : *On a vu dans chacun de vos exposés le coût très important, assumé pour l'instant largement par la puissance publique, pour la production, le maintien, la diffusion des données. Est-ce que nous avons des éléments factuels qui nous montreraient qu'on est au bord de la rupture dans ce modèle économique-là ? Est-ce que l'ouverture et la diffusion gratuite des données ne va pas mettre en danger cette capacité à produire ou à pérenniser ?*

N. Leidinger (SHOM) : *C'est clairement une question qu'on se pose quasiment tous les jours, une journée à la mer coûtant à peu près 30000 euros. Le budget décroit et notre contrat d'objectif et de performances nous demande d'augmenter nos recettes. Chaque année se pose la question du nombre de jours à la mer, du nombre de capteurs à maintenir, etc.*

M.-L. Zambon (IGN) : *A l'IGN, il y a eu une réflexion stratégique par rapport à ce bouleversement qui nous impacte, puisque on trouve justement d'autres données gratuites sur le net. Les principes de tarification des données jusqu'à présent étaient ceux de la loi CADA : comment établir une tarification, combien coûte la donnée, combien j'estime pouvoir vendre de licences sur la durée de vie du produit ? En 2013, nous avons changé les principes de tarification, travaillé pour estimer quelle est la valeur d'usage de la donnée, qu'est-ce qui est acceptable pour l'utilisateur, et pour proposer une nouvelle grille de tarification beaucoup plus simple (74 pages de tarifs auparavant pour 4 aujourd'hui), on a baissé les tarifs pour certains produits (jusqu'à 5 fois). Nous n'avons qu'une année de recul, mais il y a une augmentation de l'utilisation, notamment par la sphère privée. On a essayé d'élargir l'offre pour faciliter l'usage des informations et rentabiliser cet investissement qui est fait par l'Etat dans la politique de description du territoire.*

Ph. Santoni (Météo France) : *Du côté de Météo France, on est en train de réviser aussi un certain nombre de choses. Aujourd'hui les coûts sont là et nous avons les moyens de les assumer parce qu'il y a cette partie des coûts pris en charge par l'usager, mais si demain l'environnement devait changer et si la pression à la gratuité devenait plus forte, on arriverait effectivement à une situation assez difficile pour la production de la donnée.*

Ph. Feldmann (Cirad) : *Est-ce qu'on a une réflexion sur la valorisation économique globale de l'utilisation de ces données ? Puisqu'il s'agit de données publiques, qui ont un impact économique pour la France dans ses différentes activités, est-ce qu'on a des éléments indiquant quelle est la valeur ajoutée qui peut être produite par l'économie par la mise à disposition plus facile et moins chère des données ?*

N. Arnaud (CNRS-INSU) : *Question pertinente puisqu'elle fait écho à l'augmentation du nombre de téléchargements de données comme montrait l'IGN, par exemple. Sachant qu'on ne sait pas qui télécharge et que c'est peut-être pour ne rien en faire derrière. Mais je pense que c'est compliqué à estimer.*

M.-L. Zambon (IGN) : *On n'a pas mené d'études sur la valorisation à l'IGN. Mais je voudrais citer un exemple : quand il y a eu la tempête Xynthia, une Commission du Sénat a fait une enquête pour voir le bilan de ces dégâts, estimé à 2,5 milliards. Un des axes préconisés suite à cette enquête, était de disposer d'éléments qui permettent de maîtriser l'aménagement du territoire, et le projet Litto3D®, qui est un projet qui vise à décrire le littoral, la partie mer par le SHOM et la partie terre par l'IGN a été cité explicitement. Concrètement, ce projet est terminé côté terre et a coûté au total de l'ordre de 6,6 millions d'euros, 50% a été pris en charge par l'IGN et 50% a été pris en charge par les collectivités territoriales, locales, qui ont accepté de contribuer à la description de cette partie du territoire. Si on compare quelques millions d'euros pour investir dans des bases de données qui vont décrire et supporter des politiques publiques, par rapport à deux milliards de coûts de dégâts, il faut faire des choix politiques pour investir dans ces bases de données.*

N. Leidinger (SHOM) : *Pour compléter, la partie mer n'a effectivement pas été financée, cette partie étant de l'ordre de 1800 euros du km², bien supérieure à l'acquisition terrestre puisqu'il faut multiplier les axes de vol. Pourtant, connaître ce qui se passe en mer, comment les vagues vont déferler, est aussi très important pour prévoir les dégâts sur la terre.*

Ph. Santoni (Météo France) : *A ma connaissance, il n'y a pas eu d'étude extrêmement abouties sur la valorisation en France. On pressent bien qu'effectivement la valorisation, qui est extrêmement importante, est à mettre en balance avec les investissements qui pourraient être faits par la puissance publique.*

N. Arnaud (CNRS-INSU) : *J'ai trouvé l'estimation du rapport Trojette, par ailleurs très complet, un peu flou sur la question du type "ce doit être extrêmement important pour*

l'économie".

M. Leobet (DRI/MEDDE) : *C'est souvent idéologique et pas très scientifique, les estimations économiques !*

N. Arnaud (CNRS-INSU) : *Je suis personnellement convaincu que la puissance publique doit rester un acteur majeur de financement. On pressent bien qu'il peut y avoir une valorisation économique forte par la sphère privée et que donc, indirectement par les impôts sur les sociétés, les plus-values etc., l'Etat peut récupérer sa mise de fond et continuer à financer. Est-ce que nous pensons collectivement que puisque l'Etat donne pour mission à Météo France, à l'IGN, au SHOM, au Cirad etc. de produire la donnée, il est normal qu'il paye pour ? Et, au-delà de ça, est-ce que nous pensons que la puissance publique doit demeurer l'acteur majeur de financement de l'obtention de la donnée de très bonne qualité ?*

M. Guiraud (MNHN) : *On demande de mettre des services payants pour valoriser les données. Mais, quelle est la part des recettes qui sert vraiment à l'acquisition et quelle est la part qui sert en fait simplement à mettre à disposition les données ? On a vu que les gens téléchargent à partir du moment où c'est gratuit, ça veut dire que les gens ne sont pas prêts à payer. Avez-vous une idée de la participation des recettes allant vraiment à l'acquisition et non pas à la mise à disposition ?*

M.-L. Zambon (IGN) : *J'ai l'exemple du programme des zones inondables pour la Direction Générale de la Prévention des Risques. Je ne sais plus combien de kilomètres de terrain sont concernés mais au total, la description fine des zones inondables a coûté 13 millions d'euros, hors littoral, donc les lits des rivières etc. Treize millions d'euros, dont un tiers a été financé par l'IGN, une partie par le Ministère et une partie par les collectivités territoriales, et en terme de recettes, l'année dernière, on a perçu cent cinquante mille euros.*

M. Guiraud (MNHN) : *Il n'y aura jamais un équilibre entre les recettes et les dépenses, c'est évident.*

N. Arnaud (CNRS-INSU) : *Il est important que la puissance publique demeure un investisseur majeur, notamment pour garantir la qualité de la donnée. Cela reboucle sur un des points qui est : est-ce qu'on a une évaluation du coût que ce contrôle qualité suppose dans le cas de cette masse de données produites par les producteurs non institutionnels, qu'un certain nombre de nos institutions, organismes, commencent à s'organiser pour récupérer ?*

M.-L. Zambon (IGN) : *Nos processus de collecte sont faits actuellement de plusieurs manières, soit on a des zones, des sources d'acquisition (par exemple, les images aériennes) et on fait de la restitution de sources fiables, soit on croise différentes sources, des informations qui viennent de différents partenaires. Mais, dans ce cas, on met en place la remontée d'informations au travers de partenaires de confiance (community sourcing), par opposition*

au crowdsourcing. Par exemple, pour mettre à jour les noms de rues dans nos bases de données, ce sont les pompiers (service de sécurité de confiance, acteurs de confiance) qui nous signalent ces informations ; si c'est une remontée d'un citoyen via le Géoportail qui nous dit "non, ce n'est pas du tout la bonne rue", on va croiser avec d'autres sources d'informations. On privilégie pour l'instant les partenaires de confiance, même si les données ne sont pas toujours intégrées directement. Le projet de « Base Adresse Nationale » va nous permettre de tester la contribution directe des communes, ce qui permettrait de déporter sur des partenaires de confiance la constitution des bases de données. Quand nous fournissons des bases de données au Ministère de l'Agriculture qui fait des contrôles sur les parcelles agricoles qui sont cultivées, nous devons assurer au Ministère de l'Agriculture un niveau de qualité. Pour l'instant, on n'a pas encore défini comment on peut intégrer ces informations qui viendraient de différents citoyens directement, donc on a besoin d'un filtre de contrôle.

N. Leidinger (SHOM) : *Le SHOM est référent national pour l'observation du niveau de la mer, alors qu'il n'est pas le seul établissement en France à collecter de la mesure de hauteur d'eau. Cette fonction de référent a en fait permis de coordonner l'action du réseau, d'éviter des doublons d'acquisition, de conseiller aussi en amont sur le choix et l'installation de nouveaux marégraphes, et d'indiquer comment s'y prendre pour avoir des mesures de bonne qualité. Elle a également permis de mutualiser en France les fonctions de gestion, d'archivage et de diffusion qui sont assumées par le SHOM. Sur cette expérience, je pense qu'on y a tous gagné en finançant un référent et en permettant à tous de collecter. Et les données d'observation sont disponibles en open data, gratuitement sur un portail, qui est le portail REFMAR.*

N. Pouvreau (SHOM) : *Je m'occupe de l'animation de ce rôle de référent, qui existe depuis 2010. Les données diffusées sont des données brutes, pas des données validées, parce qu'on n'a pas les ressources humaines suffisantes pour pouvoir réaliser ce travail. L'important dans ce réseau était de découvrir qu'il y avait beaucoup d'organismes de l'Etat qui faisaient des observations au niveau de la mer. On essaie par ce moyen de diffuser tous les standards et d'améliorer les méthodes d'acquisition, travail compliqué du fait de manque de ressources humaines.*

Ph. Santoni (Météo France) : *L'Etat a aussi donné à Météo France pour mission de coordonner des réseaux d'observations qui appartiennent à d'autres que Météo France. Et effectivement, plus on s'y prend en amont, notamment en diffusant des guides de bonnes pratiques, plus on arrive à une qualité de données qui est très bonne, et donc on n'a plus besoin d'assurer de lourds contrôles derrière.*

N. Arnaud (CNRS-INSU) : *Une réaction du* Muséum*, en particulier de la science participative, sur ce point ?*

R. Julliard (MNHN) : *Le souci de la qualité des données est évidemment au cœur du programme Vigie-Nature mais la qualité s'entend à plusieurs niveaux. C'est à la fois la*

précision de ce qui est collecté et qui effectivement dépend de la compétence de l'utilisateur, mais aussi du respect des consignes, des protocoles de collecte. Toutes les constantes montrent que le fait qu'on s'appuie sur du volontariat fait que les gens qui participent se sentent très concernés par ce respect, et du coup ça permet d'assurer les moyens de comparer ces données dans le temps, dans l'espace. Et puis il y a la quantité aussi de données, la distribution des observateurs, qui vont permettre aussi la qualité de la base de données cette fois-ci, et aussi permettre de faire des traitements statistiques a posteriori du style post-stratification, reconstituer les gradients etc... La qualité de la donnée n'est pas uniquement liée à la compétence, elle est liée à ces trois composantes, et le fait de s'appuyer sur des vastes réseaux d'observateurs va maximiser l'apport de cette standardisation de ce qui est collecté, pourvu qu'il y ait un bon partage entre la conception des protocoles et puis l'acceptation de ces protocoles par les observateurs, et puis le traitement statistique a posteriori sur ces données.

R. David (CNRS) : *J'ai l'impression qu'on est encore dans la gestion du risque plus que dans la recherche de valeur ajoutée. Quel est ce ratio au sein de votre service ? Une deuxième question est de savoir s'il peut y avoir d'autres motivations. On vient de le voir avec les sciences participatives, préserver un patrimoine, ce n'est pas forcément rechercher une valeur ajoutée, et ce n'est pas non plus gérer un risque, ou peut-être un risque tellement lointain qu'on ne peut pas mesurer la perte de valeur sans avoir à ne pas préserver ce patrimoine.*

Ph. Santoni (Météo France) : *Vous avez fait un bon résumé des questions qui se posent quand on réfléchit à la façon de mettre à disposition la donnée. L'aspect de la valorisation est ce qui nous guide actuellement. La donnée existe, elle est là et autant qu'elle serve et mieux vaut s'en servir de façon de plus en plus large. Mais on voit bien que la question est démultipliée aujourd'hui grâce aux systèmes de mise à disposition, aux portails, à Internet etc.*

M.-L. Zambon (IGN) : *Nous avons également un patrimoine de données, par exemple 3,5 millions de photos aériennes. Nous avons un programme de numérisation de ces photos pour qu'elles soient disponibles gratuitement et en ligne sur le Géoportail. Il s'agit d'un vrai investissement sur plusieurs années, mais ce n'est pas sur la mise à disposition gratuite de ces clichés scannés qu'on a une valorisation puisque cela fait partie de notre mission à l'IGN. Quand on parle de valorisation, on entend ressources complémentaires pour continuer à faire nos missions. Dans nos actions, nous devons répondre à des politiques publiques, comme par exemple la description fine de l'occupation des sols, alors que nous ne disposons pas de financements pour le faire. Nous devons donc chercher des ressources complémentaires. La valorisation peut se faire via la fourniture de services, mais aussi via la mutualisation, c'est-à-dire en se rapprochant, entre acteurs publics qui travaillons sur les mêmes secteurs, et en mettant en commun nos moyens pour remplir certaines missions.*

M. Leobet (DRI/MEDDE) : *M.-L. Zambon a parlé des travaux qui ont été faits pour*

l'altimétrie dans les zones inondables. Au départ, la Direction Générale des Risques s'apprêtait à faire un appel d'offres par région, donc à du privé pour faire un travail découpé par région. J'ai pu leur expliquer les problèmes de qualité de la donnée finale, avec des contrôles hétérogènes, alors que nous voulions avoir un modèle national de modélisation des inondations. Nous avons un opérateur public, dédié, dont c'est la mission, et il était sans doute plus intéressant de conventionner avec l'IGN, plutôt que faire vingt-deux appels d'offres en métropole (pour information, le budget total était estimé à 14,5 millions d'euros, sur quatre ans !). L'argument de la mutualisation a été parfaitement entendu et le Ministère a délégué à l'IGN l'ensemble de la passation, de la sous-traitance du contrôle qualité, de la montée en base nationale etc... Les aspects de mutualisation, de confiance dans l'opérateur de référence et de contrôle qualité ont été déterminant pour aboutir à une convention entre la DGPR et l'IGN. Lorsqu'il y a des enjeux bien identifiés - on est par exemple en train de monter avec l'IGN le Géoportail de l'urbanisme pour que tous les documents d'urbanisme soient en ligne et opposables, ce qui coûte encore quelques millions d'euros – on peut trouver les moyens et les filières pour les obtenir.

Accès aux données, traçabilité, propriété intellectuelle

La recherche se nourrit des bases de données, qui à leur tour alimentent la science. Quel contrôle d'accès aux données (droit des producteurs) ? Existe-t-il des modèles de contrôle efficaces ? Quel mode de citation de la base et du jeu de données d'origine (traçabilité, respects des producteurs). Quel équilibre entre charge de gestion et d'enrichissement du système ?

Modérateur : Michel Guiraud (MNHN)
Intervenants : François Robida, Bureau de recherches géologiques et minières (BRGM)
Anne-Sophie Archambeau, Système mondial d'information sur la biodiversité (GBIF)
Pierre Cotty, Institut français de recherche pour l'exploitation de la mer (Ifremer)
Emmanuelle Jannès-Ober, Institut national de recherche en sciences et technologies pour l'environnement et l'agriculture (Irstea)
Présentation : table ronde 2.pdf

Michel Guiraud : *Nous allons donc migrer des services ou des institutions, aux relations entre les gens qui fabriquent les données, les acteurs de la recherche notamment. Comment les institutions gèrent le droit de propriété intellectuel des chercheurs ? Comment arrive-t-on à la fois à fournir, à alimenter ces bases de données tout en respectant les fournisseurs ?*

BRGM, par François Robida

Il y a des freins, mais aussi des accélérateurs qui sont en jeu.
Le premier est de tirer les leçons de ce qui a marché ou qui marche. *Infoterre* est un site de diffusion du BRGM, se disant qu'il ne fallait plus seulement faire du travail de géologue pour des géologues, mais mettre nos informations à disposition pour qu'elles soient utilisables pour d'autres choses. Par ce biais, on accède au service qui fournit la carte géologique, qui alimente un certain nombre de cartes de risques ou aux forages etc... Des indicateurs permettent de mesurer l'augmentation régulière de ces services. Cent dix-sept pays participent à la construction d'une infrastructure mondiale, *One Geology*. L'idée était de dire, quelle que soit la qualité de la carte de votre pays, on va vous aider à la mettre en ligne ; on aura des choses très hétérogènes mais c'est mieux que de garder chacun de notre côté nos données. Le *Géocatalogue,* catalogue national d'INSPIRE, est réalisé pour le compte du Ministère de l'Ecologie. Il est en forte croissance sur les métadonnées, sur les services et sur l'utilisation des services. L'application CARMEN réunit 180 producteurs. Sa facilité d'utilisation le rend performant et le plus consulté (environ dix mille visites par jour, et plus de huit mille services

web OGC qui ont été construits de cette façon). A travers ces exemples, il apparaît qu'on ne fait pas de l'interopérabilité pour être technologiquement compatible, on fait de l'interopérabilité parce qu'on a besoin ou envie ou obligation de travailler ensemble.

Tout ça ne marche qu'avec des standards, qui demandent des investissements lourds si on veut vraiment s'y plonger et encore plus si on veut être influent. Les standards pour les données géospatiales existent, un standard sur les capteurs aussi, mais qui est très peu connu. La partie ontologique est plus compliquée, mais finalement toutes les infrastructures se bâtissent comme des legos à partir de pièces qui sont à peu près toutes les mêmes.

La mutualisation, est un des éléments importants. Il n'est pas question que chacun réinvente son outil, il faut éviter la multiplication et partager les savoir-faire. Sur le Géocatalogue, la puissance publique a financé l'outil qui s'appelle Géosource, téléchargeable et utilisable gratuitement. L'idée est qu'on met en commun des moyens, mais que chacun reste maître de ses données. C'est moi qui décide de les publier, ici ou là peu importe, je choisis un opérateur ou je le fais moi-même mais je garde le contrôle complet de ma donnée. La diffusion de la donnée ne se délègue pas, la responsabilité de la diffusion ne se délègue pas, elle reste dans les mains de son fabricant. Et puis il faut publier, publier, publier, et cataloguer, pour réutiliser. Plus on publiera, plus on aura de justifications à aller réclamer de l'argent. Dans le cadre d'AllEnvi, il faudrait créer le catalogue de l'observatoire français de l'environnement.

Je pense aussi, parce qu'il faut regarder ce qu'il se passe ailleurs, et mener plusieurs types d'actions : formations, mutualisation des efforts sur les technologies (exemple : dans le cas de GEOS sur les systèmes broker qui permettent de fédérer des systèmes qui ont des standards différents) etc. C'est un peu complexe à mettre en place, mais ça peut apporter énormément de services à une communauté aussi diverse que la nôtre.

GBIF, par Anne-Sophie Archambeau

Le GBIF est financé par la contribution des pays membres votants, qui participent à hauteur de leur PIB, et par celle, forfaitaire, des pays en voie de développement. Le GBIF participe également à des programmes européens ou internationaux, comme les espèces invasives pour GEOS. Concrètement, le GBIF permet l'accès aux données de biodiversité, le financement étant là pour développer l'infrastructure, la maintenir et développer des guides de bonne pratique pour la publication des données, pour la qualité des données, pour la numérisation. Toutes les données peuvent être téléchargées librement et gratuitement sans limite de taille. Les données peuvent ensuite être réutilisées par d'autres portails, il y a différents API sur les registres du GBIF, sur les cartes, sur la taxonomie. Les services développés sont gratuits et accessibles à tous en *open source*. L'outil de publication des données du GBIF (IPT), par exemple, facilite la connexion des données sur le GBIF, mais aussi sur l'INPN, ou tout autre portail adapté. Le GBIF n'utilise que des standards internationaux, tels que Darwin Core, ABCD et EML ; il contribue au TDWG qui est un congrès sur les standards de bio-

informatique, il participe à la mise en place de ces standards et bien sûr il contribue ensuite à essayer de les instaurer et de les déployer pour créer plus d'interopérabilité. Les informations affichées permettent la traçabilité (exemple d'une plante chilienne dans l'herbier), le responsable de la donnée restant son producteur de la donnée. Ainsi, en termes de droit et responsabilité, il n'y a pas de transfert vers le GBIF. Nous essayons cependant d'améliorer le processus, au niveau traçabilité, citation, en maintenant un lien vers l'URL source. Les métadonnées, que le fournisseur de données nous fournit, sont mises en ligne et téléchargeables, mais il y a 195 champs Dublin Core à peu près, souvent peu renseignés. Pourtant, plus une donnée est renseignée, plus les métadonnées sont décrites, et meilleur l'usage et la réutilisation qui en seront faits derrière. Il faut donc insister sur l'importance de remplir ces champs. Chaque donnée a évidemment un identifiant unique et une citation qui est téléchargée en même temps qu'elle. Le projet de mise en place d'un modèle de citation attribué à chaque jeu de données, basée sur les DOI, est en cours de développement. Cela améliorera la traçabilité des données et permettra de connaître précisément les données qui sont valorisées par le GBIF et utilisées dans les publications scientifiques. Pour inciter le chercheur à fournir ses données, le GBIF a mis en place les *data papers*, en collaboration avec les éditeurs *Pensoft*, permettant ainsi d'avoir une publication scientifique avec un DOI. Lorsqu'un utilisateur vient sur le portail du GBIF pour la première fois, il doit signer le *data use agreement* qui lui rappelle les droits sur la propriété intellectuelle sur le jeu et l'obligation de citer la source des données. La difficulté se pose dans le cas du chargement d'une grande quantité de données issue de plusieurs jeux de données. Autre projet : la mise en place d'un processus de lecture automatique de ces licences de données par les machines et qui va être mis en place d'ici la fin de l'année.

Ifremer, par Pierre Cotty

A l'Ifremer, on appelle observations tout ce qui est système d'acquisition, voire même de rassemblement de données, et l'on considère que les bases de données ou la gestion de données commence à partir du moment où l'on dispose d'informations informatisées et cette fonction-gestion de données s'arrête là où commence la recherche ou la R&D. C'est-à-dire qu'on est capable en gestion de données de produire des services, de produire des données brutes mais également des produits en automatique, et à partir de là commence la recherche. Notre phase d'observations est essentiellement occupée par de très grosses infrastructures productrices de données, au premier rang desquelles on va trouver la flotte océanographique française qui existe depuis une quarantaine d'années. Le centre de données qui rassemble ces informations, a également été renforcé par l'hébergement du centre mondial Argo (TGIR-Argo), et récupère les données obtenues dans le cas de projet inter-organisme comme Coriolis. D'autres infrastructures produisent des données comme les réseaux d'observations côtiers, les données satellites intéressant le milieu marin, ou également des données liées à l'observation de la pêche et de la biotique. Il n'y a pas véritablement de problématique de récupération de données

de laboratoire, mais plus un flux de données issu de grosses infrastructures. L'organisation des bases de données se fait plutôt par filière de données, permettant sur une même base thématique de retrouver plusieurs sources de données et, grâce à une bonne description par des métadonnées, de faciliter l'inter-comparaison, le regroupement de données et l'interopérabilité. Les métadonnées produites par les infrastructures de recherche sont évidemment très précises car, si on a des observations décennales, c'est plus souvent du service temps réel. Il s'agit de mettre à disposition les données dans un délai très court après les avoir reçues, parce qu'elles présentent un intérêt pour leur actualité ; il n'y a donc pas de logique opérationnelle au sens de Météo France. Concernant la traçabilité en amont, savoir comment la donnée a été acquise, par qui, est une nécessité scientifique pour interpréter la donnée; la traçabilité en aval est de savoir qui fait quoi avec des données qui sont pour la plupart à accès libre. Les collaborations sont très nombreuses et très anciennes, le milieu de l'océan étant partagé ; il y a donc forcément beaucoup de structurations de l'interopérabilité entre pays.

La caractéristique des données marines est une grande variabilité, d'où l'importance de la précision des métadonnées, et avoir des campagnes d'observation qui tiennent compte de l'ensemble de cette variabilité. Le géoréférencement est omniprésent, d'où l'application des normes ISO OGC. L'océan est un domaine partagé, il y a énormément d'appels d'offres qui font la promotion de ce partage de données au niveau mondial et au niveau européen notamment. De ce fait, l'*open data* était pratiqué dans le milieu de l'océanographie physique depuis de très nombreuses années. L'activité base de données à l'Ifremer représente environ 35 ETP et environ 2,5 millions d'euros de budget annuel. Le recouvrement des recettes se situe entre 10 et 20%, le financement public assurant le reste. Il y a des choses qu'on ne peut pas financer. La flotte océanographique (acquisition de données), infrastructure de recherche labellisée du Ministère de la Recherche, coûte plus de quarante millions annuels. On considère que la gestion des droits sur les données et de la traçabilité fait partie intégrante de la gestion de données.

Les interactions principales de l'Ifremer sont avec l'IODE (UNESCO, environ 40 ans), qui organise les échanges de données entre Etats sur la base du *National Oceanographic Data Center*, l'Ifremer étant le NODC français. Le protocole d'échange de données géographiques OBIS (UNESCO) est plutôt pour la biodiversité. Au niveau de l'Union Européenne, on est soit coordinateur soit important contributeur à des projets comme Seadatanet 1 et 2, qui permettent de sélectionner des jeux de données dont la provenance est diverse à partir d'un portail unique. Au niveau national, les pôles d'observation de la Terre sont en cours de construction : atmosphère, océan, terre solide et surface continentale. L'Ifremer contribue fortement à la construction du pôle Océan qui permettra encore de fédérer l'ensemble des données intéressant l'océan. Si on applique la directive INSPIRE, il existe également la DCSMM pour le milieu marin, la DCE pour le côtier, la DCF *Data Coalition Framework* pour l'halieutique, au niveau

de la politique commune des pêches européennes, et pour les données de Géosciences, géologie, le code minier s'applique également en mer. Les principes de traçabilité, de contrôle d'accès et de citation, sont souvent établis en coopération dans les projets, et en général on va vers le plus petit commun multiple entre tous les participants. Sur le projet Seadatanet 2, des DOI sur les jeux de données sont utilisés par exemple ; la question se pose est de savoir quel est le conditionnement des jeux de données : gros jeux de données avec l'espoir d'avoir un maximum de citation, mais pas de distinction à l'intérieur des jeux de données pour les navires et la flotte océanographique ? est-ce qu'on le fait pour la flotte en entier, par navire ou par campagne à la mer ? Le choix est plutôt par campagne à la mer, les jeux de données sont donc un peu hétérogènes. Les données sont certes gratuites, mais les utilisateurs sont identifiés, ce qui permet là aussi d'avoir la traçabilité en aval au moins dans un premier temps, parce qu'après cela peut être transmis de proche en proche.

Irstea par Emmanuelle Jannès-Ober

Irstea est un EPST, créé il y a trente ans. Il y a en effet déjà beaucoup à dire sur les freins avant de parler des opportunités, mais il faut avoir les deux éléments en tête.

La culture du chercheur à Irstea est généralement de se dire que puisque on est payé par les deniers publics, il est normal de diffuser gratuitement des données ; les questions de « modèle économique », les problématiques de valorisation ne sont pas du tout ancrées dans la culture de l'établissement. A partir de 2009, il y a eu une réorganisation de l'établissement et une volonté d'affirmer son rôle de valorisation et de créateur de richesses, dont les données et bases de données comme produit de la recherche. La culture de la sécurité et de la propriété intellectuelle était également quelque chose de peu intégrée par nos chercheurs jusqu'en 2009, mais qui se développe avec la création d'une direction de la valorisation et du transfert.

Des communautés, comme les hydrologues, ont eu traditionnellement l'habitude de capitaliser, numériser et de diffuser un certain nombre de données et poursuivent dans cette voie (base BDOH) ; de même des bases ont été créés en conformité avec la directive INSPIRE, par exemple sur des projets pour lesquels Irstea est en délégation de service public (bases de données sur les avalanches de l'EPA et de la CLPA). Mais derrière "données", on parle à la fois de la donnée brute, des jeux de données, des relevés de terrain (y compris papiers), et des données structurées informatisées comme les bases de données. Notre DSI n'a pas encore apporté de réponse standard pour gérer et valoriser les données, et répondre à l'ensemble des besoins des chercheurs dans ce domaine. Pour se mettre en conformité avec la directive INSPIRE et réaliser un inventaire des bases de données scientifiques produites à Irstea, un Géocatalogue interne a été mis en place pour cataloguer les bases de données scientifiques ; plus de 120 bases ont été recensées mais elles ne sont pas encore intégrées au catalogue, les chercheurs n'ayant pas encore pris l'habitude d'alimenter ce dernier. Parmi les freins majeurs, on notera le temps supplémentaire que prend, pour les chercheurs, la saisie et l'absence

d'accompagnement coordonné. Je serais incapable de vous dire combien de bases de données scientifiques il y a réellement au total dans l'institut ; d'une manière générale les métadonnées sont qu'incomplètement renseignées ; en particulier, les informations relatives à la propriété intellectuelle exacte du contenu ne figurent pas. C'est un problème majeur.

En 2006, l'établissement a signé la déclaration de Berlin et s'est engagé de manière claire et volontaire dans le libre accès, en premier lieu pour la diffusion des publications. A partir de 2009, les choses s'organisent mais sur des axes qui pouvaient sembler *a priori* contradictoires (protection pour valorisation commerciale d'un côté et diffusion en libre accès de l'autre). Depuis cette année, une politique de gestion des données de la recherche est en cours de rédaction pour clarifier cela et est en cours de finalisation. Dans le même temps, un guide des bonnes pratiques pour l'acquisition, la gestion et la diffusion des données a été réalisé, définissant les processus d'organisation des données, de collecte, et de documentation pour assurer la traçabilité et garantir les usages futurs (réutilisation en particulier).On prévoit ici que la réutilisation des données se fera, selon les cas, sur un modèle soit totalement gratuit, soit payant lorsque l'accès aux données sera couplé à des services. Le chantier est énorme et l'on commence seulement à organiser les choses, en dehors de quelques exceptions remarquables. Il nous faut aussi clarifier des choses par rapport au nouveau cadrage de H2020, puisque les recommandations se révèlent en réalité des obligations pour les projets qui portent sur des données environnementales. Irstea tient absolument à garantir le droit de ses partenaires et établit des règles déjà pour les nouveaux projets ; le traitement rétrospectif de nos anciennes bases de données sera pour sa part beaucoup plus complexe. On structure beaucoup les choses autour des données, mais *quid* du lien entre publication et données ? Les éditeurs commerciaux poussent aujourd'hui de plus en plus les auteurs à céder leurs jeux de données ; la question de la propriété intellectuelle des jeux de données qui ne seront plus gérés, dans ce cas, par les établissements sera à traiter si les établissements font le constat qu'ils n'ont pas demain les moyens de gérer en interne leurs données et choisissent ce mode de sous-traitance. Concernant l'équilibre charges-gestion-ROI, on a peu parlé de coûts complets, et du poids humain qui pèse de plus en plus sur les établissements : les chercheurs pour documenter leurs données, les informaticiens, les professionnels de l'information scientifique et technique qui gèrent déjà les publications et qui vont documenter, gérer une partie des données. Il y a un travail de réflexion à conduire collectivement sur des axes de mutualisation inter-établissements : la question est de savoir quels moyens nous pouvons dégager maintenant pour faire avancer ce chantier de longue date, alors que nous sommes désormais au pied du mur ?

Discussions

M. Guiraud (MNHN) : *On a compris que les choses s'organisent par institution pour assurer cette traçabilité des données, ce lien entre etc. Mais est-ce que les outils sont là pour garantir*

que chaque partenaire retrouve ses données ? Dans le cas du GBIF ou de l'Ifremer : est-ce que ce sont vos données ou bien est-ce qu'on voit bien clairement que ce sont les données des laboratoires qui ont participés ? Pour reprendre l'idée de ce portail dont François Robida nous inspirait la création. Comment est-ce que vous percevez les choses ? Est-ce que vous pensez que dans les institutions, les chercheurs sont vraiment prêts à partager et à faire en sorte de se mettre en arrière par rapport à un portail national, européen ? Parce que derrière cela, il y a les indicateurs, et donc les ressources financières ou pas.

P. Cotty (Ifremer) : *Je viens répondre pour Ifremer : la gestion de données, c'est faire le lien entre la recherche et les coûts de production ou de gestion de données. Et quand je parlais de DOI, c'est dans le cadre du projet Seadatanet 2 qui va se mettre en place, les jeux de données provenant principalement des navires de recherche européens. On a réalisé des bibliométries pour la TGIR flotte océanographique française qui permet de dire voilà ce que telle campagne a généré. Le problème est que la campagne peut durer quinze jours et qu'il faut attendre jusqu'à 15 ans pour avoir des publications qui se réfèrent à cette campagne. On a aujourd'hui les outils pour faire ce genre de bibliométrie, très demandée pour les infrastructures parce cela coûte très cher, mais on va l'étendre aux jeux de données parce qu'on considère qu'ils peuvent moyennant un travail (interprétation, calcul) produire de nouveaux jeux de données, qui vont ensuite eux-mêmes générer des publications. Donc il faut créer les liens à la fois avec le producteur, grâce aux métadonnées, mais également sur les jeux de données pour avoir les deux accès. Autant qu'on a pu le voir, les chercheurs se réfèrent sans difficulté à leur publication quand ils publient en open access et remplissent facilement la case à cocher dans notre système, permettant de faire un repérage direct du lien entre la publication et la campagne.*

A.-S. Archambeau (GBIF) : *Le GBIF a pour rôle de valoriser les jeux de données. Les citations sont sur les jeux de données qui sont téléchargés. Après, savoir si les gens sont passés par le GBIF ou pas est très difficile, puisqu'ils ne citent que le nom du jeu de données. La difficulté est donc établir quelle a été l'utilisation du GBIF pour obtenir de l'information.*

F. Robida (BRGM) : *Je voudrais prendre l'exemple du Référentiel Géologique de la France (RGF), qui prend la suite de la carte géologique, et qui amène une révision assez complète de nos processus de production de données et de l'élaboration de produits dérivés. Dans l'historique de la carte géologique, une carte était produite par une équipe mixte composée de géologues du monde académique et de géologues du BRGM, et s'il y avait une publication, les auteurs étaient les auteurs de cette carte. Si la référence était claire, on perdait par contre toute la trace des observations de terrain et on enregistrait finalement que le produit final, qui est une interprétation. On a des acquisitions de données qui sont faites, soit par des gens du BRGM, soit par des gens d'université ou d'autres partenaires. On se doit donc de stocker toutes ses données et de les tracer pour ensuite être capable d'identifier quel est le créateur de la donnée. On essaye de stocker toutes les étapes qui amènent à la création de la carte donc*

les interprétations, le tracé, le produit, et maintenant non seulement une carte 2D, mais le modèle 3D. On est encore loin d'avoir résolu toutes les questions, celle des auteurs d'un modèle 3D, de la nature même d'une publication de ce type vs *une carte géologique. Les choses sont assez simples du point de vue des bases de données institutionnelles, type banque de données du sous-sol. Les choses deviennent de plus en plus compliquées au fur et à mesure que la granularité devient fine et que les données de projet ont été générées par un chercheur ou un ingénieur. Mais le changement de processus sur nos données, nous donne finalement l'occasion d'essayer de faire les choses à peu près proprement sur toute la chaîne.*

E. Jannès-Ober (IRTEA) : *Est-ce que les établissements seraient prêts à alimenter un portail commun ? Tout dépend des sujets et si les choses sont déjà organisées. Il n'y a pas de freins a priori à mutualiser sur des nouveaux projets, surtout s'il y a un partenariat, mais est-ce que l'on dispose chacun des moyens nécessaires ?*

Y. Biard (CIRAD) : *J'ai été très intéressé par la dernière présentation parce qu'on développe le même type de grand chantier stratégique complètement transversal au Cirad. Il y a donc là matière à échanger. Vous n'avez pas présenté les aspects éthiques, déontologiques qui tournent autour des jeux de données. Nous travaillons beaucoup avec des partenaires du sud, et cela nous amène à réfléchir sur les données des partenaires qu'on manipule et sur leur montée en compétence. Il y a aussi à réfléchir au niveau des RH et de la formation, puisque beaucoup d'activités vont se regrouper peut-être dans de nouveaux métiers, des métiers de détective, pour aller chercher les données qui sont à valoriser de façon prioritaire ; il y a aussi tout ce qui est animation et formation de ses propres collègues sur la propriété intellectuelle.*

R. David (CNRS) : *C'était très intéressant parce qu'il y avait des points de vue assez contrastés. J'ai l'impression souvent qu'on a tendance à présenter le bon côté des systèmes d'information et que dès qu'on réalise une véritable enquête sur les services, les usages, et les niveaux d'usage par rapport aux attendus, on est souvent déçu. Une question pour chacun d'entre vous : «* Quelle est la prochaine amélioration à prévoir pour augmenter son niveau et cette qualité d'usage ? *»*

A.-S. Archambeau (GBIF) : *Je crois que ça passe beaucoup par les métadonnées. Il faut vraiment que les chercheurs, que les fournisseurs de données se rendent compte que mieux ils vont détailler leurs données, et plus l'usage qui en sera fait après sera intéressant et adapté. Il y aura moins de risques que les données soient utilisées pour de mauvaises raisons. Il est difficile de faire la part des choses entre la qualité des données et l'adaptation à l'usage, parce une donnée qui ne sera pas assez qualifiée pour quelqu'un, le sera pour une autre étude. Alors remplir peut-être cinquante ou soixante champs, c'est un peu plus long au départ, mais à la fin au niveau de l'usage, ça changera beaucoup de choses.*

F. Robida (BRGM) : *Un des éléments fondamentaux est l'aspect RH, qu'est ce qui fait que les*

gens sont incités à aller dans ce type de démarche ? Sur le projet, on va mettre dans la mécanique même des systèmes d'information des périodes d'embargo sur les données, pendant que lesquelles les gens préparent leur publication, leur thèse. Les données en question sont accessibles en interne mais ne sont pas diffusées, par exemple. Ensuite, il y a : « je mets à disposition mes données pour les autres mais est-ce que moi j'ai intérêt à en récupérer aussi. Si je suis dans ma bulle et que je n'ai pas besoin de données, je n'ai aucun intérêt à aller en chercher. *» Il faut que le système s'initie pour que le système puisse commencer à fonctionner.*

P. Cotty (Ifremer) : *Je voudrais me défendre d'avoir fait une présentation trop favorable de la réalité d'Ifremer. En raison du coût des campagnes, les scientifiques sont particulièrement sensibilisés à la valorisation des instruments de l'observation et de la gestion de données qui a permis d'arriver à la publication. Donc je ne perçois pas de difficultés particulières à motiver les troupes. Certes, on ne fait jamais du 100 %. Il y a toujours du travail à faire pour mobiliser tout le monde pour rendre toutes les données et toutes les métadonnées, qui établissent le lien entre la publication et le producteur de données. Je pense que dans d'autres schémas où les données coûtent relativement peu cher à acquérir, il y a peut-être moins de motivation et plus cette appropriation personnelle au niveau des scientifiques.*

E. Jannès-Ober (Irstea) : *En dehors des questions RH évidemment, comment améliorer tout ça ? Une des pistes est de s'appuyer sur la démarche qualité. L'idée est de mettre autour de la table l'ensemble des acteurs, au démarrage de chaque projet de recherche, et même au moment de la réponse aux appels d'offres, pour se poser les bonnes questions et pour savoir justement comment, dans ce cadre précis, on va valoriser les résultats, mettre à disposition les jeux de données, dans quel cadre, sur quel support. On pourra alors décider si on gère ça en interne, ou si on le fait via un portail externe, ou autre. La difficulté aujourd'hui est que l'on intervient en général quand le projet est quasiment dans sa phase finale ; on se dit alors «* comment je diffuse les résultats ? *» Mais la pression est telle qu'on sait que les chercheurs répondent souvent aux appels d'offre dans l'urgence, et qu'ils n'ont pas le temps d'organiser la concertation préalable. Je pense que c'est un vrai défi.*

G. Reverdin (CNRS) : *Je vois deux problèmes qui ressortent des différentes présentations. Comment arriver à motiver suffisamment en amont les chercheurs ou les producteurs de données à s'impliquer jusqu'à la donnée non pas brute mais la donnée validée ? Parce que dans notre domaine [océanographie], la donnée est source de beaucoup d'erreurs qui sont difficiles à identifier, et il faut une implication forte du chercheur ou que des outils soient mis en place dans les bases de données pour en assurer la validation. Et, d'autre part, comment faire reconnaître cet effort qui sera fait par des acteurs extérieurs, et qui est essentiel pour l'utilisation de la donnée ?*

A.-S. Archambeau (GBIF) : *Le GBIF a travaillé sur ce sujet, parce qu'un des freins classiques à l'accès aux données, est : «* non, il faut que je publie, et puis de toute façon, ma

base de données, je ne vais avoir aucune reconnaissance dessus ; la reconnaissance que j'ai, c'est sur ma publication scientifique sur la recherche, pas sur ma base de données. *» C'est exactement pour cette raison qu'ont été créés les* data papers*, qui sont une publication scientifique sur la description d'une base de données, et pas sur le résultat de la recherche. C'est une publication de rang A, avec un DOI, qui peut donc être citée dans votre rapport d'activité.*

Valorisation, diffusion, diversité des données

Les équipes de recherche produisent des données dans le cadre de leurs programmes. Ces données peuvent être utiles pour le reste de la communauté et sont souvent soumises à une obligation de diffusion mais sont très diverses et très diversement structurées. Faut-il privilégier une organisation des données a priori ou a posteriori ? Peut-on respecter la diversité des données ? Quels sont les éléments de construction indispensables ? Quels périmètres (disciplinaires, thématiques, institutionnels) et quelles interactions possibles entre eux ?

Modérateur : Cécile Callou (MNHN)
Intervenants : Elisabeth Leclerc, Agence nationale pour la gestion des déchets radioactifs (Andra)
Francis Raoul remplace Gudrun Bornett, Institut nationale de l'écologie et de l'environnement (CNRS-INEE)
Jacqueline Boutin remplace Philippe Bertrand, Institut national des sciences de l'Univers (CNRS-INSU)
Patrick Farcy, Institut français de recherche pour l'exploitation de la mer (Ifremer) et Philippe Bertrand (CNRS-INSU)
Présentation : table ronde 3.pdf

C. Callou (MNHN) : Les données sont utiles évidemment à leur producteur, mais également à une communauté de plus en plus large avec cette obligation de diffusion, mais la difficulté est que ces données sont extrêmement diverses et surtout, diversement structurées. Faut-il privilégier l'organisation des données a priori ou a posteriori ? Est-ce que la standardisation des données est vraiment une solution ? Si oui, jusqu'à quel point faut-il tout standardiser ? Peut-on respecter cette diversité des données, puisque c'est une richesse ? On parlait de freins, mais c'est aussi le côté très positif de cette diversité. Cela prouve que la recherche est dynamique, active. Comment faire pour trouver à la fois des moyens de standardiser, mais également de préserver cette richesse et cette diversité même entre les institutions ? Quels sont les éléments de construction indispensables ? On a évoqué, même si le mot n'a pas été prononcé, de thesaurus, de référentiels, de vocabulaires contrôlés mais surtout des métadonnées. Ces dernières sont vraiment au cœur du problème. On a déjà insisté sur ce point, mais on constate qu'il est très difficile de valoriser des données si on n'a pas des métadonnées correctement renseignées. Autre question : faut-il rationaliser ces données par la mise en place de plateformes thématiques, disciplinaires, en dépassant les institutions, comme nous l'avons vu avec le Pôle Océan, ou comme cela est en train de se mettre en place pour l'écologie expérimentale (ANAEE) ? Et enfin, on voit bien qu'une fois que l'on a produit de la donnée brute, cela ne suffit pas, il y a donc tous ces aspects de valorisation et d'enrichissement de la donnée qui semblent assez intéressants à aborder.

Andra (Agence nationale pour la gestion des déchets radioactifs), par Elisabeth Leclerc :

En 2006-2007, l'*Observatoire Pérenne de l'Environnement* (OPE) a été mis en place en même temps que le démarrage du projet de stockage géologique de déchets radioactifs français, pour établir un état de référence de l'environnement sur plusieurs années. L'OPE est situé dans le Nord Est de la France, à la limite des départements Meuse et Haute-Marne, future localisation d'un potentiel stockage. On est sur l'observation d'une zone de 900 km^2 qui englobe plusieurs sous-bassins versant, dans un milieu karstique, avec petits cours d'eau en amont du bassin Seine-Normandie. Une des particularités de cet observatoire est que l'Andra garantit la pérennité de la surveillance et du suivi complet de tous les écosystèmes de ce territoire pendant 100 ans. L'OPE est donc un nœud de réseau, couvrant toutes les données de l'environnement, que ce soit les sols, l'homme, la faune, la flore, l'eau et l'air. Les questions scientifiques associées sont très axées sur les cycles biochimiques, donc les distributions : cycles et flux (sur le long terme), la dynamique, la sensibilité de la biodiversité, les indicateurs et capteurs environnementaux.

La base de données informatique a été développée à partir d'une base de données qui concernait les expérimentations et des observations géologiques au niveau du laboratoire souterrain construit sur le site depuis 1999, peu adaptée aux observations environnementales de surface (les oiseaux par exemple). Au-delà des données sur prélèvements d'échantillons, analyses, on a aussi des observations et des stations instrumentées (mesures en continu). On est donc en train de développer une nouvelle base de données. L'idée étant d'avoir une approche intégrée de l'environnement, les métadonnées sont indispensables (mesure de 400 à 500 paramètres sur chaque échantillon de sol, d'eau etc., méthode utilisée quand elle est normée, etc.). On s'appuie essentiellement sur les programmes nationaux et internationaux, avec des protocoles validés, l'entrée des données étant gérée et validée par les partenaires (exemple, Vigie-Nature).

Toutes nos données sont mises sur les réseaux accessibles, nationaux thématiques : réseau eau-Seine-Normandie pour l'eau, Atmolor pour l'air, plateforme ICOS (CO^2) etc. L'OPE est un SOERE et appartient à plusieurs réseaux nationaux ou internationaux (ex : de surveillance radiologique du territoire OPERA). Il y a donc une obligation de transparence et d'accès aux données. L'Andra ne possède pas les terrains d'échantillonnage et doit obtenir les accords des exploitants et des propriétaires des terrains pour avoir accès à leurs territoires et à leurs produits (agricoles notamment). L'OPE travaille avec des laboratoires de recherche, des bureaux d'études, des associations avec des prestations de services et des partenariats, la question de la copropriété des données et de la protection de certaines informations parfois sensibles est donc essentielle. La publication des données-chercheurs avant une diffusion en open access est aussi prise en compte. Les protocoles sont établis par les réseaux thématiques internationaux et nationaux, qui valident et diffusent les données. La plupart de ces données sont déjà accessibles via les réseaux. L'interface du site web de l'OPE permet d'accéder soit

directement aux liens des sites partenaires, soit à un formulaire d'accès aux données. Presque 2500 points sont suivis sur cette zone, ce qui génère beaucoup d'échantillons. Ces échantillons (sol, matrices biologiques etc.) sont conservés et doivent pouvoir être accessibles à la communauté scientifique, avec les mêmes problématiques d'accès, de droit, de traçabilité. La valorisation se fait aussi au niveau local, important pour la valorisation du territoire. L'Andra est certifiée ISO 9001, il y a donc déjà des normes qui permettent aussi d'avoir un certain processus de validation et d'assurance qualité, même dans les processus d'acquisition de données. Cela permet aussi de justifier ce que l'on fait et comment on le fait.

Expérience de la base dans une unité mixte de recherche CRNS-INEE, par Francis Raoul

Le laboratoire Chrono-environnement est une unité pluridisciplinaire dans laquelle on est appelé à manipuler des données d'ordre épidémiologique, écologique, hydrologique, géologique etc. d'environnements anciens et actuels – en essayant de faire le lien entre ces corpus de données. On s'inscrit nécessairement et très souvent dans du suivi long terme et spatialisé des systèmes étudiés, soit en prospectif avec la notion d'observatoire, soit en rétrospectif avec la notion de retro-observatoire. Nous sommes également amenés à gérer un certain nombre de banques d'échantillons de nature très variée, à disposition de la communauté internationale : os, crânes, sérums humain, pollens, tissus, ADN, etc. Jusqu'à maintenant, l'ensemble était majoritairement géré sur des formats Excel et donc pas forcément exploitable de façon rigoureuse. Sur la base de ce constat, depuis 2012, nous avons mis en place un axe transversal « base de données » au sein de l'UMR, en impliquant d'autres institutions partenaires. Cet axe est piloté par un tandem « spécialiste des bases de données » et « spécialiste en écologie ». Nous avons eu la chance de pouvoir mobiliser des crédits notamment de la Région Franche-Comté pour salarier des ingénieurs qui développent et modélisent les bases de données. Nous avons des informaticiens de l'UMR qui s'occupent de l'architecture serveur. L'effort à consentir en termes de temps de travail est de 1,2 ETP maître de conférences (Chrono-environnement) et environ 1,2 ETP ingénieur informatique, pris sur les ressources internes des URM (UMS OSU THETA dans le cas présent). C'est donc du ressort de la volonté politique des laboratoires ou des instituts de dégager du temps ou des moyens de recrutement pour travailler sur les bases de données. Dans notre cas, le Conseil Régional de Franche-Comté a vite compris l'enjeu qu'il y a derrière la gestion des bases de données et a accepté de financer sur trois années un salaire d'ingénieur de bases de données pour faire le travail.

Faut-il gérer la diversité des données ? Pour moi la réponse est oui, il faut faire ce qu'il faut pour pouvoir gérer cette diversité de données. Pour cela, les mêmes personnes s'occupent de tous les projets de bases de données de l'UMR, quelle que soit la thématique. L'idée est d'avoir une approche globale de toutes ces bases. On va également vers l'utilisation de standard communs en travaillant avec des partenaires comme l'UMS BBEES. Ça nous permet

de mettre le curseur à la bonne place entre une base de données unique, dans laquelle on met tout mais qui fait perdre finalement la diversité, et puis des bases de données trop fragmentaires et pas connectables entre elles. La réflexion est en cours pour la définition des métadonnées, dans le cadre d'un projet de l'OSU THETA. En termes de diffusion, les premières bases de données seront mises en ligne à l'automne 2014, et on réfléchit à une politique globale de diffusion de ces informations. Tout n'est pas encore en place mais l'idée est de témoigner ici de cette situation : une unité de recherche qui se lance dans la gestion organisée de Bases de Données. Je pense que nous ne sommes pas les seuls à être concernés. Puisque c'est là que la donnée de base est produite, c'est la première marche. Mais cela pose un certains nombres de questions. La pérennité du soutien financier pour le salaire des ingénieurs informaticiens qui font le travail. C'est un point clé. Et puis aussi, un problème que je ne soupçonnais pas forcement avant de travailler sur ces aspects : tout le monde n'a pas la culture « base de données ». C'est le cas du chercheur « moyen » dans le domaine qui me concerne. Il faut arriver, d'une façon ou d'une autre, à créer une culture sur les bases de données. Cela peut passer par exemple par la formation doctorale. Les écoles doctorales devraient être impliquées. Ce sont les futurs chercheurs qui vont être confrontés à ce genre de problème par la suite.

Le Pôle des Observations de l'Océan, par Patrick Farcy

Au sujet de l'intérêt que portent les chercheurs pour les bases de données, il est vrai qu'à l'Ifremer ils n'ont pas vraiment le choix : « vous voulez avoir des données à partir d'un navire. On finance le navire, mais les données doivent rentrer dans les bases de données » ! Grâce à ce phénomène d'incitation obligatoire, très peu de données restent dans les PC des ingénieurs et chercheurs.

Je vais vous présenter le Pôle des Observations de l'Océan, ou Pôle de données de l'Océan, qui est une démarche menée assez récemment à plusieurs organismes (Ifremer, Cnes, CNRS, IGN, IRD, Météo France, SHOM, les universités marines, et peut-être le BRGM), autour de ce besoin d'échanger des données, d'accéder à plusieurs types de données, de travailler ensemble sur des jeux de données que l'on n'a pas forcement tous disponibles. L'idée de la création de ces Pôles d'observations, ou Pôle de données, émane d'un groupe de travail sur des données environnementales, constitué d'un certain nombre d'organismes et piloté par le Cnes et le CNRS. Le Pôle Océan bénéficie peut être plus de guides pour nous structurer : le programme Mer d'AllEnvi au niveau National par exemple, qui définit un certain nombre de besoins concernant la connaissance du système Mer, ou encore la directive cadre sur la stratégie du milieu marin, qui est une directive européenne qui va s'appliquer à partir de 2016. Au niveau Européen, la tendance est de vouloir structurer les données par filière thématique : la physique, la biologie, la chimie, la bathymétrie, etc… (7 types de filière). On a un très fort besoin de structurer la recherche au niveau de l'Océan, ne serait-ce que pour répondre à des appels d'offres et aller chercher des financements. On a un nombre de systèmes d'observations

colossales, qui va des satellites jusqu'aux mammifères marins, voire bientôt des capteurs sur les plongeurs sous-marin, et de paramètres fondamentaux à mesurer. Cette complexité de retrouve dans la structure des bases de données à l'Ifremer, dans laquelle il y a six grandes thématiques : physique, géoscience, surveillance littorale, pêche-aquaculture, biodiversité, océanographie spatiale. Ces thématiques sont faites de manière à ce que chacune ait son propre portail. On peut accéder à la physique de l'Océan par le portail Coriolis qui permet de récupérer un certain nombre de données et de travailler dessus. Mais si dans ma base de données, j'ai une information qui me donne une évaluation de la pêche à tel endroit et que je veux la relier au paramètre physique de l'océan, je dois chercher dans les différents serveurs les paramètres qui vont m'intéresser. Les portails ne sont pas forcément identiques, ce n'est pas la même logique pour y accéder, vous n'avez pas forcement les outils qui vous permettre de localiser vos données. Le Pôle de données devrait faciliter l'accès unique pour vous permettre de poser de telle question et de savoir sur quel type de données vous pouvez compter pour y répondre.

Les objectifs : offrir un portail unique, faciliter l'accès aux données, fournir des services (documentation, navigation, colocalisation de données), des extractions, faciliter l'accès à des informations de faible niveau, ou d'événements extrêmes. Donc l'utilisation de technique de fouille de données, par exemple. Mais aussi optimiser les ressources de matériels et humaines.

Les 13 commandements du Pôle : Définition d'une stratégie Nationale-Européenne parce qu'on ne peut pas faire en France quelque chose qui soit complétement distinct de ce sera l'environnement européen ; Mise en place en France de 4 pôles de données, mais peut-être qu'il en restera un seul dans dix, quinze ou vingt ans ; prévoir rapidement une structuration inter-Pôle assez forte ; Une gouvernance commune entre l'ensemble des participants ; l'expertise scientifique au cœur des pôles ; s'appuyer sur des moyens mutualisés ; établir des liens clairs entre recherche et opérationnel ; rôle défini des structures privées, de plus en plus présentes ; importance de la formation et de la communication ; améliorer l'accès aux données ; un archivage géré au niveau national ; mise en place de laboratoires experts et enfin, prise en compte de l'arrivée des nouvelles technologies, comme la la fouille de données, par exemple. La phase de réflexion a débuté en 2014, pour un projet qui devrait durer cinq ans. Fin 2018, un colloque est envisagé pour évaluer cette capacité à consolider l'ensemble (ou plutôt une grande partie) des bases de données autour de l'océan.

CNRS-INSU, par Jacqueline Boutin

Un très gros travail de réflexion a donc piloté par le Cnes et le CNRS-INSU pour réfléchir à l'organisation des données et des bases de données, notamment vers un Pôle Océan et un Pôle Atmosphère en se plaçant au-delà des organismes. Des projets comme CYBER *Cycles biogéochimiques, environnement et ressources*, propres au CNRS-INSU, ou encore les activités autour de RESOMAR pour le côtier montrent que les données acquises sont

d'origines très diverses. Mais avant tout, si ce groupe de travail a eu lieu c'est parce que les bases de données sont conservées dans des endroits divers et pas forcément très visibles pour qui n'est pas de la communauté directe qui a acquis les données. Il y a donc une volonté de fournir aux utilisateurs de données une meilleure vision de ces données et une meilleure vision aussi de la qualité de ces données. Pour avoir l'information sur la qualité des données distribuées en open access, les métadonnées sont indispensables. Pour cela, il faut fournir aux producteurs/fournisseurs les moyens de le faire de façon simple. Clairement, le chercheur ne souhaite pas s'adapter à des standards compliqués. L'objectif des Pôles est de proposer une structuration basée sur ce qui existe, en évitant de recréer de nouvelles bases de données, sauf en cas de manque évident. L'idée de la proposition ANR MODIF, est d'arriver à construire autour de quelques projets pilotes des exemples de bases de données qui peuvent dialoguer entre elles.

Cécile Callou (MNHN) : Ce qui me frappe en vous écoutant, c'est l'existence de deux niveaux d'échelles : le niveau chercheur, celui que nous rencontrons très fréquemment, qui cherche à mettre ces données en base, qui sait qu'il faut passer à la vitesse supérieure, qui ne sait pas comment et qui a des problèmes de moyens (techniques, RH) et le niveau très organisé comme à l'Ifremer et à l'Andra, où le système est très contraint (bateaux, territoires) et donc toutes des données sont en bases. La première étape pour tous semble bien être cette volonté de porter à la connaissance l'existence de ces bases à l'aide de portails.

Il y a également cet aspect des données brutes qui ouvrent sur d'autres données. Vous avez évoqué le cas des échantillons, qui amènent un coût de conservation et de gestion. Car ces échantillons sont là pour être réutilisés et recréés de la donnée. On a l'impression que tout est fini quand les jeux de données sont dans une base de données. On a fait l'essentiel, mais ce qui reste à faire derrière est aussi important. Avez-vous une idée en termes de budget, ce que cela représente.

Discussions

P. Farcy (Ifremer) : Je pense que c'est un peu moins important que la phase de construction, mais que c'est quand même primordial. Tout dépend de ce que l'on veut faire après, qui va pouvoir utiliser les données qui sont dans un premier temps utilisées pour des besoins bien précis. En océanographie opérationnelle, les données viennent en temps quasi réel et sont utilisées pour alimenter des modèles. Mais, tous les ans, on rejoue des données pour améliorer des jeux de données précédents, qui eux-mêmes amélioreront les modèles prévisions qui vont être utilisés dans un an ou deux. Depuis 20 ans, grâce à l'apport de nouveaux systèmes, de nouveaux capteurs plus précis mais aussi parce que l'on a rejoué les données d'il y a cinq - dix ans, on s'est aperçu que l'on était capable d'améliorer certains paramètre, et donc de diminuer l'erreur. Ces jeux de données retraités, améliorés vont servir aux modélisateurs et également à

d'autres communautés qui vont faire plutôt des statistiques ou de la climatologie.

Cécile Callou (MNHN) : Même si cela entre finalement dans le cadre de la recherche, ça améliore donc aussi la qualité de la donnée. Dans ce cas, la réutilisation est une vraie plus-value.

F. Raoul (CNRS-INEE) : C'est clair, la plus-value est nette. Nous sommes en début de processus, ce n'est donc pas une priorité pour nous actuellement. C'est quelque chose qui est envisagé dans un deuxième temps. La diversité des échantillons, c'est aussi une partie du « capital de guerre » du scientifique qui doit être valorisé au même titre que la donnée de départ.

M. Lebouvier (CNRS) : Dans les interventions que nous venons d'entendre, est revenu à plusieurs reprises la question de la mobilisation des chercheurs. Ce qui m'évoque un séminaire qui s'était tenu dans les années 2000 où l'intervenant avait estimé qu'un chercheur devait passer 20 à 25 % de son temps à préparer et à structurer ses données pour la diffusion. C'est sans doute juste, mais de le dire comme ça c'est totalement contre-productif. Je pense que la mobilisation des personnes, elle peut se faire beaucoup mieux en démontrant l'utilité en interne. Dans l'unité, nous avons des thèses, des masters, des papiers en préparation avec des jeux de données. Et très rapidement, personne n'a plus aucune vue sur cette production. Plus que par la contrainte, c'est par ce biais qu'on peut intéresser les collègues. Nous avons à Rennes une école doctorale qui, pour une soutenance de thèse, exige un papier soumis ou accepté et introduit la production de métadonnées.

Mais ce n'est pas toujours évident en raison de la diversité des données. En lien avec le traité sur l'antarctique, chaque responsable de programme de recherche soutenu par l'institut Polaire Français est tenu de fournir chaque année des bases de métadonnées dans un format assez détaillé et basé sur le « global change master directory ». Il est difficile de correspondre à un schéma général, même si on souhaite un thesaurus, une structure commune. Donc les métadonnées en interne bien sûr et tant mieux si cela sert à la communauté. Une anecdote enfin : il y a quelques années, j'ai rencontré un collègue des Etats Unis qui travaille dans les LTER *Long Term Ecological research*. Il m'a expliqué que si dans sa recherche il tombe sur une métadonnée avec mon adresse pour demander les données, il passe à autre chose. L'intérêt est pour ce qui est immédiatement disponible.

C. Callou (MNHN) : Il est en effet impératif de montrer l'utilité des métadonnées aux chercheurs. Je pense que vous êtes tous sensibles à cette question. Le gros problème, auquel on se trouve confronté est l'évaluation du chercheur sur sa production scientifique, ses publications. L'embargo sur les données mis en place un peu partout maintenant (entre 3 et 5 ans selon les disciplines) fait que cela passe un petit mieux. Outre les *data paper*s, il existe un autre élément incitatif pour la diffusion des données : en octobre 2012, l'AERES a enfin reconnue « la constitution et la mise à disposition de bases de données, de logiciels, de corpus

et d'outils de recherche » comme des « productions scientifiques de rang A ».

J. Boutin (CNRS) : Un argument aussi en faveur de ces bases de données, c'est l'archivage pérenne. Il faut assurer au scientifique que les données acquises aujourd'hui seront accessibles dans dix ans. Or, dans les labos, on a assez peu cette capacité.

R. David (CNRS) : Quand on parle de la visibilité de la donnée, on imagine souvent une base de données plus ou moins accessible et plus ou moins utilisée. Cela rappelle les débats d'il y a 15-20 ans sur le web invisible. On a un ensemble de données visibles sur des pages, référencées par des moteurs. Et on a toutes les données derrière un formulaire, qui ne sont pas accessibles à ces moteurs de recherche. Aujourd'hui sont considérées publiques des données personnelles que quelqu'un met volontairement en ligne : des photos de vacances, des films, des schémas, des inventions, des brevets, etc… et donc répliquables à foison. On entre alors dans une autre logique de Web sémantique, qui va permettre de structurer l'accès aux données et d'augmenter son impact. Le chercheur aura un outil formidable qui lui permettra d'avoir tout de suite, sans effort, des informations transdisciplinaires.

Ph. Feldmann (Cirad) : Dans le cadre de l'évaluation des projets de l'ANR (mais aussi au niveau européen), dans un contexte contraint au niveau budgétaire, on fait de plus en plus attention au bon usage des fonds publics. Dans les critères d'évaluation, même s'il n'y a pas encore suffisamment de visibilité de publication, il y a une analyse qui est faite de la manière dont sont produits les résultats et la manière dont ils sont ensuite mis à disposition et maintenus accessibles. Comme c'est déjà le cas de collection de type ressources biologiques, avec dépôt des échantillons dans les collections de ressources. Autant les chercheurs ont raison de dire qu'ils sont évalués essentiellement sur des critères de publications, autant pour obtenir actuellement des financements, le fait de bien utiliser les résultats, donc de bien archiver, stocker les données et les rendre disponibles figurent parmi les critères d'évaluation.

Nous avons évoqué la « science participative » avec Vigie-Nature, on peut également parler du réseau Visionature qui, à l'échelle de région Ile-de-France en l'espace de trois ans, a collecté des millions de données environnementales. Alors que les années précédentes on en avait au mieux quelques dizaine de milliers. Il y a là des systèmes de production en masse de données qui ne sont pas forcément structurées de la façon dont un chercheur souhaiterait qu'elles le soient pour répondre à un projet précis de recherche. Il va donc falloir se poser la question de savoir comment analyser, interpréter et exploiter des données qui n'ont pas été préparées pour un objectif précis.

Y. Biard (Cirad) : La publication associée aux données c'est bien un système de gestion de la preuve. Sinon, au niveau de la motivation du chercheur, il y a différents niveaux d'actions : laboratoire, institut, nationale, et même parfois supranationale ou inter-institut. Mais à partir du moment où on descend jusqu'au niveau du chercheur, de l'individu, je pense que l'on est complètement dans le domaine du management. Si on veut éviter les échanges entre

chercheurs via Dropbox ou d'autres services tiers qui partent aux Etats-Unis, il y a une démarche pédagogique urgente à mener.

F. Raoul (CNRS) : Il y a clairement une révolution culturelle à faire dans certains domaines. Maintenant, je n'ai pas la solution mais les pistes que vous évoquez me semblent intéressantes. C'est peut être moi qui ai parlé du chercheur moyen, je ne sais plus…Le chercheur, enseignant qui plus est, avec des responsabilités administratives. Si en plus on rajoute l'aspect base de données alors que ce n'est pas dans mes habitudes. Je vais dire : t'es gentil, mais je n'ai pas le temps ! Je le dis, car c'est ce que les gens me disent. Je pense qu'il y a un vrai travail sur le long terme à faire pour faire évoluer cette mentalité et je pense que via les écoles doctorales il y a probablement une carte à jouer pour sensibiliser les chercheurs dès le début.

E. Leclerc (Andra) : Je voudrais signaler la volonté de certains d'avancer par thématique, comme ANAEE pour l'analyse et l'expérimentation des écosystèmes. Il y a des spécialistes dans chaque domaine donc laissons à l'atmosphère la compétence, à la mer la compétence etc… Mais le métier de créer une base de données est également une compétence.

A.-S. Archambeau (GBIF) : En ce qui concerne la science participative, le problème n'est pas que technique, il y a aussi une notion de mentalité. Prenons l'exemple de Diveboard, association de plongeurs sous-marins. Ils ont réalisés un site web international (vos carnets de plongée en ligne), avec de nombreuses descriptions et notamment des données d'observation sur les poissons. Nous avons pu ajouter avec eux 2/3 champs pour faciliter le flux de données entre nos systèmes et grâce à ce partenariat, enregistré l'apparition de poissons lion dans certaines zones des caraïbes. Ce n'est pas toujours la technique qui bloque, mais plus les gens.

F. Robida (BRGM) : Web sémantique et harmonisation des données sont pour moi un oxymore. Ce que l'on peut harmoniser, c'est la façon d'échanger les données. Je crois qu'il faut même militer pour la diversité des données. Il ne s'agit pas de tout faire rentrer. Une des questions qui se posent, c'est finalement l'approche open data *versus* l'approche « INSPIRE ». Approche INSPIRE : on type tout, on fait rentrer les données dans un modèle, cela permet d'avoir des échanges parfaitement standardisés et peut se révéler être très utile. Mais il y a d'autres choses qui relèvent plutôt de l'open data : j'ai un bout de fichier quelque part, structuré à ma façon que je le publie. Et je ne m'embête pas de savoir si c'est suivant les normes qui ont été définies par toute une collectivité ! Il y a un intervalle considérable pour lequel les outillages du web sémantique devraient me permettre de beaucoup mieux exploiter ces données complètements hétérogènes. Je crois qu'il faut faire attention, chacun dans nos métiers, dans nos domaines, à savoir exactement où on se situe et jusqu'où on veut aller dans normalisation, la standardisation de l'échange et la structuration de la donnée.

Discussion générale

- Quelle dynamique collective à mettre en place dans chaque communauté ?
- Quelle dynamique collective à mettre en place entre communautés ?
- Open Data
- Définition d'un plan d'action
-

M. Guiraud : *Je vous rappelle que l'idée de ce séminaire était d'avoir une discussion ensemble pour essayer d'établir, sinon une feuille de route, du moins quelques lignes directrices sur ce qu'il faut faire et comment il faut s'y prendre.*

Au niveau de chaque communauté, quelle dynamique faut-il créer ? Nous l'avons vu, certaines communautés sont bien organisées. L'accès aux données, la traçabilité, la mise à disposition des données, l'acquisition des données, tout cela fonctionne bien. C'est moins le cas dans d'autres communautés. Mais quand on parle de communautés, comment les identifier ? De quoi parle-t-on ? Nous avons entendu parler de pôles de données. Faut-il dépasser les institutions ? Faut-il s'organiser au sein de chaque institution ou plutôt d'une façon interinstitutionnelle et, dans ce dernier cas, dans quel paysage ? Comment s'articuler au sein de chaque communauté et entre communautés ?

De même, les choix sont un point qu'il faudra discuter. Nous avons les contraintes, les obligations d'accessibilité. Quoi ? Comment ? Quel est l'apport ou quel est le retour que l'on peut en avoir ? Rien n'est gratuit. Comment tout cela est-il pris en charge ?

Il faut enfin essayer d'échafauder un plan d'action. Vous reconnaissez-vous dans cette notion de pôles de données telle qu'elle a été suggérée ? En évitant l'écueil d'avoir autant de pôles que de bases de données. Cette notion de pôles de données induit aussi les questions d'appropriation. On voit beaucoup de top-down. A un moment, les institutions ou les organismes décident d'organiser les nombreuses bases de données, mais, en bas, les gens ont déjà organisés les choses à leur manière. Si on raisonne en termes RH, on comprend bien la fragilité de la construction des bases de données (1 à 2 ETP). Il y a la construction de la base, mais aussi le temps passé à rédiger les projets pour aller récupérer des financements etc. Cette base de données est leur bébé, donc à partir de quel moment ce bébé peut-il être transmis à des structures plus impersonnelles, et quel outil mettre en place ?

Un des problèmes réside là : comment organiser cette structuration autour de pôles en fonction de toutes les bases de données et des bonnes volontés qui existent ? Il semblerait que seules les méthodes coercitives de l'Ifremer fonctionnent. Vous voulez prendre le bateau, les données sont dedans. Ce modèle fonctionne avec les bateaux … mais pas forcément dans les bureaux !

J. Boutin (CNRS) : *Dans le cadre d'un projet européen, nous sommes obligés de mettre les données acquises dans une base de données, en espérant que celle-ci soit coordonnée avec une base de données nationale. Cette volonté européenne se retrouve également dans les discussions sur la DCSMM (directive cadre « stratégie pour le milieu marin »). Du point de vue du chercheur, il est parfois très difficile de répondre seul à cet engagement.*

C. Callou (MNHN) : *Dans le cadre des demandes auprès de l'ANR, des projets de bases de données, de diffusion de l'information sont de plus en plus systématiquement inscrits. Mais la réalité est que la somme allouée à un projet, s'il est accepté, est bien moindre que celle-demandée. Et qu'est-ce qui disparaît ? La partie base de données. Il faudrait que l'ANR fasse en sorte que cette partie soit obligatoirement financée. Et pas seulement avec un CDD de six mois ! On a vu que qu'en amont des Pôles, il existait des plateformes. Il faut aider ces éléments à atteindre un niveau technologique suffisant, ce qui a un vrai coût en termes de ressources humaines.*

D. Couvet (CNRS/MNHN) : *On peut se demander dans quelle mesure il ne faudrait pas aller plus loin et renforcer la dynamique collective. Ne faudrait-il pas créer une plateforme permanente, une sorte d'enviroscope, fixant quelques objectifs ? Nous avons besoin d'une plateforme d'échange des informations entre les observatoires. On voit bien que les initiatives sont multiples et qu'il faut échanger sur ces initiatives. L'autre problème est celui de la reconnaissance des chercheurs. Nous avons parlé des* data papers. *Si cette plateforme venait à susciter les collaborations, donc les publications, il s'agirait évidemment d'un mécanisme vertueux. Ensuite se pose le problème des moyens. Le contexte est peut-être favorable, notamment avec l'organisation des systèmes d'observation de l'environnement à l'échelle mondiale (GEOSS) qui fait actuellement un gros travail de normalisation. De son côté, GEO BON (*Group on Earth Observations Biodiversity Observation Network*) a créé le concept des variables essentielles de la diversité (EBV), qui permet de faire de la cartographie de la visibilité des différents observatoires biodiversité.*

Les sciences participatives sont en train de se développer et il y a là aussi des outils qui montrent qu'il peut y avoir synergie entre les données protocolées des chercheurs et les données non protocolées de la science participative, avec un renforcement de la qualité. Cet enviroscope pourrait également peser dans les arbitrages du ministère de la Recherche quant à l'allocation des moyens dédiée à l'observation. Il y a un décalage des moyens alloués entre les données d'observation des étoiles et celles de l'observation de l'environnement, alors que les questions environnementales sont devenues très présentes.

N. Arnaud (CNRS-INSU) : *Il faut éviter de déshabiller Paul pour habiller Jacques. Battons-nous pour que l'ensemble de la recherche fondamentale soit valorisé, sans chercher à ne poursuivre que des défis sociétaux. Historiquement et culturellement, les communautés « astronomie » et « astrophysique » par exemple ont toujours placé l'information et le partage au cœur de leur démarche scientifique. Elles ont construit leurs infrastructures autour de cela*

et elles ont su les défendre, les valoriser scientifiquement. Fondamentalement, le point ultime de blocage est la manière dont on incite le chercheur à partager sa donnée. Il s'agit souvent d'un individualiste forcené qui a déployé beaucoup d'efforts pour la générer. Il faut donc lui simplifier la mise en ligne etc. mais, surtout, il faut que ce soit valorisant pour lui parce que, dans le cas contraire, il ne le fera pas.

Au-delà de la demande de moyens, AllEnvi, qui rassemble nombre d'organismes et d'établissements, peut essayer de faire passer le message dans tous les lieux d'évaluation – et jusqu'à l'évaluation individuelle – de la richesse que représente l'effort de mise à disposition de la donnée pour la personne qui a généré cette donnée. Les data papers sont importants à condition que les commissions qui évaluent les chercheurs reconnaissent leur valeur. Nous connaissons tous les travers du système.

Non présentée : *En termes de dynamique collective, il faudrait faire en sorte que la formation soit uniformisée par le biais des écoles doctorales qui traitent des sciences de l'environnement ou via les différents organismes d'AllEnvi. Je voudrais également revenir sur la valorisation des données réutilisées. Nous sommes entre chercheurs publics, mais il s'agit d'un point à préparer pour l'avenir. Mettre en pôles toutes ces données ou bases de données pour une utilisation de la recherche et pour le progrès des connaissances est une bonne chose, mais il pourrait également exister d'autres usages et d'autres usagers.*

C. Callou (MNHN) : *Des personnes qui n'appartiennent au monde académique sont présents aujourd'hui. Il serait intéressant d'avoir leur point de vue sur cette question.*

H. Pedersen (CNRS/UJF) : *J'ai deux propositions. La première serait d'œuvrer d'une manière très active pour constituer un groupe de travail autour des indicateurs associés aux infrastructures avec open data. J'ignore où en est le projet discuté un temps à la DGRI. La deuxième serait de jouer un rôle moteur pour les références de données de type DOI. Je sais que nous travaillons actuellement à un standard international en sismologie afin que chaque réseau ait un DOI associé. Nous travaillons aussi de manière active vis-à-vis des publications d'articles de recherche, pour faire figurer la référence formelle du data set qui a été. Cela veut dire que le chercheur va pouvoir compter les DOI qui ont été utilisés. De grands journaux du type Nature ou Science vont de plus en plus dans ce sens, c'est-à-dire que la référence formelle de la donnée doit figurer dans l'article publié.*

R. David (CNRS) : *Nous réfléchissons aux outils qu'il manque mais pas au calendrier qui permettra de les diffuser efficacement. Un peu comme si nous sortions une nouvelle voiture qu'aucun garage en France ne serait capable d'entretenir. Il faut éviter la fracture numérique au niveau des chercheurs et réfléchir activement à la meilleure manière de diffuser cette sensibilisation et cette connaissance.*

M. Guiraud (MNHN) : *Nous avons bien vu perçu les enjeux de l'open data. Quelle est votre perception des obligations de publication des données ? Dans quelle mesure peut-on assurer*

un certain droit de propriété par rapport aux services liés aux données ? La tendance est la gratuité. Y a-t-il une distinction subtile à faire entre les données, collectées dans le cadre d'un projet, et les services qui y sont liés, comme les outils de validation ?

C. Callou (MNHN) : *Si les données doivent être gratuites, il n'est pas injustifié que les services, qui constituent une plus-value, soient payants.*

Y. Biard (Cirad) : *Vous avez parlé d'embargo sur les données, de décalage temporel entre la publication scientifique et l'ouverture des données. Est-ce formalisé ? Est-ce la politique de tel ou tel établissement ? Et comment cela peut-il s'articuler par rapport aux contraintes règlementaires sur les données publiques ?*

C. Callou (MNHN) : *Du côté des sciences humaines, la pratique voudrait que l'on instaure une protection de cinq ans parce que les données sont plus longues à acquérir, donc plus longues à traiter. En revanche, du côté des sciences dites « dures », le délai est en général de trois ans. La tendance actuelle oscille donc entre trois ans et cinq ans. Cela ne veut pas dire que la donnée n'est pas en base mais plutôt qu'elle n'est pas encore rendue accessible. Il n'existe pas de texte formalisé, même s'il pourrait être intéressant, y compris dans le cadre d'AllEnvi, d'arrêter clairement une période d'embargo afin de laisser aux chercheurs le temps nécessaire de valoriser leurs données.*

Non présenté : *Dans certaines bases de données publiques, l'embargo se révèle parfois très compliqué à lever et la donnée reste indisponible tant que le chercheur n'a pas donné son accord. Quant à la notion de service, elle peut très compliquée parce que le service peut provenir de maints partenaires de la base. Le chercheur valide et améliore sa donnée, il en fait un produit qu'il fournit à la base. Les données de la base servent pour des cartographies ou divers produits secondaires. Ensuite les projets européens combinent des données de différentes bases de données pour en faire des produits. A cela s'ajoutent les services de type Copernicus (*European Earth Observation Programme*) qui demandent que ces données, ces produits soient dans le domaine public. Il est donc très difficile de séparer ce qui va être public de ce qui peut ne pas l'être.*

Non présenté : *La notion de contrainte et de droit (moral, d'usage) des agents publics ne fait pas complètement consensus entre les juristes et les services de valorisation de nos institutions. Un effort d'harmonisation, ou d'information, reste à faire.*

M. Guiraud(MNHN) : *Le droit de propriété du chercheur est reconnu et Marc Leobet a indiqué qu'il pouvait être étendu aux équipes de recherche, ainsi qu'aux ingénieurs. Parce que, statutairement parlant, les ingénieurs n'ont aucun droit de propriété intellectuelle étant donné qu'ils abandonnent tout à leur organisme.*

C. Callou (MNHN) : *Vous trouverez de nombreux documents à ce sujet sur le site du Réseau Bases de Données, réseau créé par le CNRS mais regroupant de nombreuses personnes appartenant à différentes institutions.*

N. Arnaud (CNRS-INSU) : *Je peux entendre que la donnée générée sous fonds publics soit totalement ouverte, mise à disposition tout de suite. Il y a la question de l'embargo avant publication etc., mais de quelle donnée parlons-nous ? De la donnée brute captée au sens le plus large du terme, y compris quand le capteur est humain ? A quel moment passe-t-on de la donnée au service ?*

F. Robida (BRGM) : *La classification des données est l'éternelle question. J'utilise une donnée. Pour moi, c'est une donnée brute. Pour celui qui me la fournit, c'est une donnée élaborée. La chaîne est souvent compliquée à déterminer. Des services d'affichage de données ou des services de téléchargement des données sont des services. Le « téléchargement des données » est parfois improprement utilisé. On doit pouvoir utiliser la donnée sans avoir besoin de faire une copie chez soi de la base de données du voisin. Il y a là encore un travail d'éducation à faire auprès des utilisateurs.*

J'ai aussi une suggestion dans la ligne de l'observatoire français de l'environnement dont j'ai évoqué la création possible. Quelqu'un a-t-il une idée du nombre de catalogues de métadonnées dans AllEnvi ? La première chose à faire serait peut-être de répertorier les catalogues de métadonnées afin que nous puissions avoir une première image de ce dont nous disposons.

C. Pichot (Inra) : *En termes d'informations sur les métadonnées, même si cela concerne parfois une information de premier niveau (le « porter à connaissance »), le Géocatalogue existe. Mais si, par curiosité, on entre le mot « AllEnvi » sur le Géocatalogue, il n'existe qu'une seule fiche de métadonnées. C'est peu.*

Les ressources qui nous intéressent dans le cadre de l'environnement ne se limitent pas aux jeux de données. Il s'agit de ressources plus globales, d'informations sur les équipes, les projets, les infrastructures etc. Elles ont aussi vocation à être décrites au travers de métadonnées, et pas simplement les jeux de données en tant que tels. Je milite pour qu'il y ait une remontée des fiches que nous produisons, soit localement, soit directement au niveau du portail national du Géocatalogue.

R. David (CNRS) : *De la même manière que la question s'était posée pour les sciences participatives, il serait intéressant de se demander quelle est la durée de vie moyenne d'un catalogue ? Là, je pense que c'est la science participative qui gagne.*

N. Arnaud (CNRS-INSU) : *Pour les bases de données dans les champs des pôles thématiques de données qui commencent à se structurer (Océan, Atmosphère, Surfaces et interfaces continentales ou Terre solide), ce travail de référencement ou de rassemblement de ce qui existe en termes de catalogues de métadonnées avance. C'est un travail difficile. Evidemment, ces pôles de données sont loin de couvrir l'ensemble des thématiques, en particulier dans le domaine du vivant et de l'environnement. Ils ne seront donc pas exhaustifs.*

M. Guiraud (MNHN) : *Ne faudrait-il pas commencer par identifier le nombre de pôles de*

données pertinents, s'assurer que les métadonnées sont cataloguées et qu'elles rejoignent le catalogue général, même si on peut se poser la question de la durée de vie ? Cela pourrait être une des actions concrètes à initier au niveau d'AllEnvi.

C. Callou (MNHN) : *L'INIST serait un bon partenaire pour réaliser ce travail. Il faudrait qu'ils viennent vers les chercheurs.*

D. Couvet (CNRS/MNHN) : *L'Ecoscope est l'organisation des observatoires sur la biodiversité sur le long terme. Avoir une plateforme d'échange est vraiment très apprécié. Développer les métadonnées pour dresser la cartographie des observatoires est nécessaire. Peut-être plus nécessaire dans la biodiversité que dans les autres domaines environnementaux en raison de la complexité des systèmes écologiques. Cela étant, la pérennité de l'Ecoscope dépendra des moyens significatifs qu'il permettra de drainer, en termes de ressources humaines, d'appareillage etc.*

N. Arnaud (CNRS-INSU) : *Les quatre pôles thématiques de données qui se créent, n'ont pas forcément valeur de modèles, mais ont l'avantage d'avoir permis une réflexion : comment structurer les choses, quels sont les acteurs, comment le mettre en œuvre etc. ? Au sein d'AllEnvi, nous pourrions essayer de voir comment se structure chacun de ces pôles, comment a été fait le travail de catalogage. Et on peut imaginer que l'Ecoscope joue le rôle d'un pôle thématique de données pour un certain nombre de champs qui ne seraient pas couverts.*

Le deuxième point serait de voir concrètement comment faire avancer le référencement des bases de données de manière à ce que le chercheur puisse très, très facilement faire référencer sa base de données.

Enfin, troisième point, il faudrait qu'AllEnvi puisse transmettre à l'ensemble de ses membres l'importance de la préservation des données issues de la recherche. Dès la formation, dès l'école doctorale, dès la soutenance de la thèse et ensuite dans la vie du chercheur, cette notion devrait être valorisée autant que la publication scientifique.

P. Cotty (Ifremer) : *Parmi les suggestions pour AllEnvi, je verrais bien une forme de labellisation qui ne porte pas nécessairement le nom de SOERE. Au niveau de la labellisation des Systèmes d'Observations traditionnels, des SOERE d'AllEnvi ou des TGIR au niveau du ministère, on a trop considéré que seule l'observation coûtait et que tout le reste pouvait en découler sans problèmes. Les intervenants qui se sont exprimés, chacun au nom de son organisme, ont tous affirmé que la gestion de données n'était pas couverte par des financements récurrents, voire par des projets.*

N. Arnaud (CNRS-INSU) : *Théoriquement, les SOERE n'étaient pas censés permettre le fonctionnement des observatoires. Au départ, ils étaient censés participer aux actions collectives de la deuxième étape, à savoir notamment la constitution des bases de données. Par la suite, nous avons assisté à une dérive collective assez forte des moyens des SOERE qui étaient donc utilisés par les observatoires individuels – les nœuds de terrain – parce que*

l'observation aussi coûte cher et que, au sein des différents organismes, les moyens dédiés à l'observation se raréfient. Une réflexion est menée actuellement au sein du groupe infrastructures d'AllEnvi afin de voir comment repositionner une partie des moyens des SOERE sur cette deuxième étape : la sécurisation de la donnée et le coût que cela implique.

M. Guiraud (MNHN) : *Nous allons clore cette journée et nous pouvons nous féliciter du nombre important d'inscrits, qui prouve qu'il s'agit d'un sujet aujourd'hui majeur. Les interventions ont d'ailleurs montré comment cela va impacter la vie quotidienne dans nos laboratoires.*

Le rôle d'AllEnvi est de relayer les questions que se pose la communauté et d'agir auprès des ministères pour essayer de remédier au mieux à la situation.

C. Callou (MNHN) : *Nous présentons toutes nos excuses aux organismes que nous n'avons pas sollicités. Il serait intéressant de refaire le même exercice d'ici deux ou trois ans afin de voir ce qui a évolué, puisque les choses bougent très, très vite.*

- - -

Fin de la journée