« Bases de données et accès public aux résultats »

Compte-rendu de la table ronde
Colloque du programme de recherche « écosystèmes tropicaux »
des 7 et 8 novembre 2006 dédié à Daniel Lachaise

Tour de table

Patrice Mengin-Lecreulx, chef du département recherche de l'ONF

Il convient en premier lieu de préciser les objectifs recherchés à travers la constitution et mise à disposition de bases de données. En particulier :

- Quels sont les publics bénéficiaires visés (chercheurs, gestionnaires et professionnels, public plus large) ?
- Que souhaite-t-on mettre à disposition selon les cas : des données de recherche ou certaines données de gestion (inventaires forestiers par exemple), des données élémentaires ou des informations plus générales sur les données (métadonnées : quoi et où) ?
- Eviter de perdre des investissements lourds consentis pour l'acquisition de données jugées importantes, en documentant et sécurisant des bases ou fichiers de données épars.
- On devrait veiller à créer des synergies et de la valeur ajoutée en commençant par la mise à disposition de données au service de projets bien identifiés, pouvant relever de la recherche ou répondre à des questions posées par la gestion ou les tutelles (par exemple : bilan de carbone pour la Guyane, etc.). Partir, au moins pour commencer, de questions structurantes plutôt que du seul principe de mise à disposition de bases de données, sans projets de valorisation à la clé.

Une attention particulière doit être portée aux données d'inventaire et de monitoring qui peuvent être structurantes pour de nombreuses études, de recherche comme de gestion, quand elles portent sur de vastes surfaces ou des gammes de temps longues, et bénéficient parfois d'une assurance qualité.

Il serait par exemple très utile de mettre à disposition sous forme numérique les données de grands inventaires stratégiques, par exemple : l'inventaire papetier en Guyane, l'inventaire de reconnaissance du Sud Cameroun, et bien d'autres inventaires réalisés en Afrique dans les années 60 à 80 par le CTFT.

Les risques de pertes de données de ce type sont bien réels, avec des exemples à la clé.

Jean-Luc Peyron, Directeur d'ECOFOR

Importance de la gestion des données

L'information a toujours été essentielle en appui aux décideurs sans parler ici évidemment de la désinformation.

Sa maîtrise doit beaucoup aux nouvelles techniques de l'information et de la communication.

La composante « systèmes d'information » est aujourd'hui, en environnement comme ailleurs, incontournable dans la plupart des institutions et constitue, à ECOFOR, l'un des quatre grands axes de travail.

Un des facteurs expliquant la carence de la connaissance sur les milieux tropicaux est la complexité de leurs écosystèmes. Celle-ci incite tout d'abord à l'observation attentive et à la description. Cette tâche commence par l'énorme travail d'inventaire des espèces qui est en cours depuis le 18ème siècle et apparaît aujourd'hui sans fin. Mais elle concerne plus largement l'ensemble des milieux naturels et, de la même façon, le milieu socio-économique avec lequel il est en interaction. Envisager cette description là où elle est inexistante et nécessaire, l'organiser de manière intelligente, la réaliser rigoureusement puis l'analyser en profondeur est un acte de recherche à part entière. Ceci semble une évidence mais n'est pas toujours reconnu comme tel. De même, le besoin de suivi continu (monitoring) est important pour connaître le fonctionnement et l'évolution des écosystèmes. La comparabilité dans l'espace est évidemment indissociable du suivi temporel. La possibilité d'étendre les échelles s'avère également attrayante vis-à-vis des questions qui se posent souvent dorénavant au niveau régional voire planétaire.

Quelques perspectives sur les moyens de progresser en la matière

La carence d'information doit inciter à valoriser du mieux possible l'information disponible et, par voie de conséquence, à donner l'envie de recueillir de nouvelles données.

Pour caricaturer, on peut distinguer deux grands types de gestion des données

- une gestion centralisée telle que celle qui est pratiquée pour l'évaluation des ressources forestières mondiales, le suivi de la convention sur la diversité biologique, celui de la convention cadre sur la changement climatique, le protocole de Kyoto, le programme européen de suivi continu des forêts « Forest Focus »…,
- une gestion décentralisée, sur la base des projets de recherche, voire de réseaux plus ou moins volontaires.

C'est à ce second aspect qu'il s'agit ici de s'intéresser dans la mesure où il concerne de près le fonctionnement des programmes de recherche et où il nécessite plus de réflexions et de stimulations que le premier.

Plus on a affaire à des données dispersées, moins on peut songer à les rassembler dans une même base. D'ailleurs, même l'INSEE ne le fait pas et garde une gestion partiellement décentralisée.

Il faut avoir conscience de la grande différence qui existe entre constituer une base de données pour un projet de recherche et la constituer pour une utilisation éventuelle ultérieure, voire par d'autres. Pour cette raison, il serait sans doute justifié d'inciter spécifiquement les équipes de recherche à constituer durant leurs travaux une base de données dans les règles et d'une manière qui la rende effectivement réutilisable. On pourrait par exemple faire en sorte que la constitution d'une telle base de données soit équivalente à une publication et qu'elle soit partie intégrante du projet de recherche et de son financement.

L'efficacité d'une telle démarche repose évidemment sur la volonté de partager les données. Une telle perspective est parfois vue par certains chercheurs comme une expropriation, un problème pour leur carrière. Sur ce point, il faut aller plus loin que de mettre face à face l'intérêt général de la communauté scientifique et l'intérêt particulier de tel ou tel chercheur. Il faut en effet reconnaître que cette stratégie est gagnante aussi bien au plan individuel qu'au plan collectif. Dans ce domaine, le protectionnisme est une illusion qui masque les occasions que le partage des données peut faire naître : occasion de collaborer, occasion de s'approprier une fraction de ce que d'autres font avec ses propres données ; sans compter la possibilité d'utiliser soit même l'information communautaire.

Libre accès aux données ne signifie pas forcément accès gratuit.

Dans cette démarche, l'assurance qualité joue un rôle important et incite à s'adapter à des protocoles.

Un problème important est celui du rassemblement de données éparses. Cependant, un tel rassemblement à l'intérieur d'une même base de données est souvent illusoire. La solution consiste à mettre en avant non pas les données mais les métadonnées, c'est-à-dire les informations normalisées sur les informations. De la même façon, l'information constituée sous forme de métadonnées conduit mécaniquement aux données.

Quelques expériences d'Ecofor

Ecofor travaille sur un projet de conférence internationale pour novembre/décembre 2007, sur le thème de la gestion de l'information dans les forêts tropicales (knowledge management for tropical forestry).

Par ailleurs, Ecofor travaille sur un catalogue des sources d'information sur la forêt (Ca-Sif), donc sur les métadonnées forestières. Ce projet est effectué en lien avec le système d'information sur la nature et les paysages (Sinp) du ministère de l'écologie et du développement durable, dont Ecofor est administrateur secondaire pour les aspects forestiers. Ce catalogue s'est pour l'instant surtout intéressé aux sources d'information sur les forêts tempérées hexagonales, mais pourrait s'étendre en direction des forêts tropicales, au moins celles des départements et collectivités d'outre-mer.

Egalement, les réflexions animées par Ecofor sur le thème des indicateurs de biodiversité dans le cadre du programme « Biodiversité et gestion forestière » des ministères de l'écologie et du développement durable, d'une part, de l'agriculture et de la pêche, d'autre part, ont mis en évidence, dans le domaine forestier, d'un manque d'indicateurs sur la richesse, l'abondance et l'équilibre spécifique et donc, parallèlement, sur le besoin de susciter de nouvelles données, y compris, par l'intermédiaire de réseaux décentralisés de naturalistes, à l'instar des pratiques dans le domaine du réseau Vigie-Nature du Muséum national d'histoire naturelle.

Jacques Trouvillez, responsable du service du patrimoine naturel qui a la responsabilité de l'inventaire du vivant au Museum

Quels sont les types de données concernés par la mise à disposition, s'agit-il de :

1) données élémentaires ou données de synthèse ?

Données de synthèse : celles publiées dans les articles

Données élémentaires : utiles aux chercheurs, utiles à la constructions d'indicateurs (qu'ils soient à destination de la recherche, des décideurs, …).

2) données nécessaires aux chercheurs ou issues de la recherche ?,

3) données utiles aux gestionnaires, qui permettent par exemple d'établir des listes rouges d'espèces, des réglementations, …

Il convient également de savoir qui est propriétaire des données, à ce titre plusieurs natures de « propriétés » sont à envisager :

1) la propriété intellectuelle,
2) la « propriété du financeur » : voir Convention d' Aarhus signée par la France à ce sujet sur les données environnementales dont la récolte a été financée par les pouvoirs publics

Exemples de mise en commun en cours

- GBIF (global biodiversity information facility) géré en France par le Museum (GBIF : www.gbif.org, www.gbif.fr) assure l'interopérabilité des banques de données,

- SINP du MEDD (France métropolitaine et collectivités d'outre-mer en cours) permettra d'accéder aux métadonnées sur la flore, la faune et les paysages,

- Inventaire national du patrimoine naturel : données rassemblées dans un atlas.

Jean-Paul Rudant, Université de Marne-la-Vallée

La rédaction d'un cahier des charges pour la constitution d'une base de données rassemblant les données des chercheurs peut sembler une première tâche sur laquelle travailler. Un travail d'étudiant pourrait permettre de cerner les méthodes les plus adaptées pour mettre en œuvre le recueil et à la pérennité des données, et ce sur un ou des exemples répondant à des préoccupations précises, en liaison avec des chercheurs intéressés.

Concernant la mise en commun opérationnelle de rapports de recherche et études, il indique l'existence du site www.pefac.net, financé par le MAE dans le cadre du projet FORINFO au Gabon (Forêts d'Afrique centrale). Ce site réunit des documents au format pdf, adapté au faible debit des réseaux locaux, et permet d'extraire totalité ou partie d'un rapport en en prenant livraison sur sa boite mail.

Dans le domaine de la télédétection, dans le cas de données diffusées par une agence commerciale, seul l'acquéreur des données (et dans certains cas les équipes associées au projet) peut les utiliser, elles ne sont théoriquement pas partageables.

Apport au débat faite en conclusion du colloque par Eric Vindimian, chef du service recherche et prospection du MEDD à cette analyse

Accéder facilement à des données existantes, permettrait de valoriser le travail d'analyse et d'interprétation de celles-ci. Puisqu'en s'affranchissant de leur collecte, l'analyse deviendrait le travail central d'un projet.

Interventions et questions de la salle

Sylvie Gourlet-Fleury, CIRAD

La pérennité des données est un important facteur à considérer notamment quand l'organisme qui gère ou recueille ces données change de politique (exemples de données perdues : inventaires du CTFT en Guyane dans les années 70, seules les données agrégées existent

encore). Faut-il imaginer un système de financement des bases de données sur le long terme pour assurer cette pérennité ?

Réponse de Jacques Trouvillez, MNHN

Le stockage des données n'est pas pris en compte dans le coût d'un projet de recherche. Un partenariat pourrait être envisagé entre les producteurs de données et leurs utilisateurs. Dans la mission du Gbif, le Museum assure un peu le rôle de « bibliothèque de données ».

Réponse de Finn Kjellberg, CNRS

Le risque de stocker en un lieu unique l'ensemble des données c'est de perdre tout cet ensemble si ce lieu disparaît ou a des défaillances (parallèle fait avec la bibliothèque d'Alexandrie).

La gratuité d'accès aux données permet une diffusion de l'information favorisant sa pérennité.

En ce qui concerne la mise à disposition sur internet d'articles publiés, il paraît peu risqué de le faire sur le site personnel de l'auteur de l'article mais les sites institutionnels refusent de le faire en raison du risque encouru de voir les éditeurs des revues porter plainte.

Guy Landmann, ECOFOR

Le GIP Medias France gère les bases de données spatiales de Météo France, du CNES et du CNRS. Ceci est un exemple de gestion commune de base de données par un organisme extérieur à l'organisme producteur des données.

Dans le domaine des données écologiques, les organismes de recherche ne sont pas dans cette logique ni même dans celle qui consisterait à monter leur propre base de données.

Eric Loffeier, CIRAD

Les observatoires de recherche en environnement (ORE) devaient permettre une mise à disposition de données, qu'en est-il ?

Réponse de Jacques Trouvillez, MNHN

Ils n'ont pas permis de rendre publiques les données

Réponse de Guy Landmann, ECOFOR

Le site www.ore.fr permet de constater l'évolution des différents projets d'ORE et des bases de données associées. Les données seront bien publiques.

Sylvie Gourlet-Fleury, CIRAD

Exemple de base de métadonnées forestières qui n'ont pas fonctionné (pas de mise à jour, peu d'information) : GFIS, Tropis.

Que va apporter Ca-SIF, le projet d'ECOFOR, dans ce contexte ?

Philippe Birnbaum, CIRAD

Des données stockées peuvent perdre leur pertinence, car les données des chercheurs évoluent (particulièrement en génétique). La mise à disposition sur un site de données évolutives nécessite donc leur mise à jour. Cela est lourd à gérer pour le chercheur. Un

système qui viendrait se mettre à jour depuis la base personnelle du chercheur serait à encourager.

Finn Kjellberg, CNRS

Devant le manque d'argent des organismes de recherche et la réduction corrélative des effectifs des techniciens qui assuraient la collecte de données, on s'oriente vers une « collecte » organisée par des réseaux amateurs ou professionnels dans ce contexte. La « qualité » des données n'est plus assurée.

Jacques Trouvillez, MNHN

Pour assurer une certaine qualité il est nécessaire de rétribuer les amateurs.

Bernard Riéra, CNRS, MNHN, ECOFOR

Les bases de données qui fonctionnent (c'est-à-dire qui sont mises à jour) sont celles qui sont organisées autour d'un centre d'intérêt d'envergure relativement réduite et touchant une communauté motivée.

Christian Moretti, IRD

Les constats dressés par une étude récente réalisée par un étudiant en DEA sont les suivantes.

Il existe des bases de données accessibles mais elles manquent en général d'indications sur la qualité de l'information.

Par ailleurs, elles sont rarement viables car elles ne sont pas conçues pour les utilisateurs de données.

Dans la construction d'une base de données, il est indispensable de mettre l'utilisateur de données au centre du système (besoin d'informations sur les données elles-mêmes : qualité, protocole, …).

> Réponse de Guy Landmann, ECOFOR
>
> Il est difficile de savoir exactement quels vont être les utilisateurs des bases de données ; ainsi les concepteurs font des hypothèses à ce sujet mais la réalité peut ensuite être toute autre ce qui explique en partie le manque d'efficacité de certaines bases de données.

L'accessibilité aux données faciliterait le développement de modèles, gourmands en données.

Patrice Mengin-Lecreulx, ONF

La base de données ECOPLANT, rassemblant de nombreuses données de thèses permet de réaliser des études puissantes sur les niches écologiques des espèces, ou la prédiction des caractéristiques du sols à partir de la flore observée. On peut aussi concevoir des bases des données plus modestes.

Jean-Luc Peyron, ECOFOR

Il faut promouvoir la diversité des outils et des comportements pour faciliter la mise à disposition de données.

Guy Landmann, ECOFOR

L'animateur du projet Ca-SIF, une base de métadonnées, en cours de réalisation par ECOFOR prendra contact avec les organismes de recherche tropicale pour connaître leur position vis-à vis de ce projet.

Conclusion de Yves Gillon, IRD, Président du conseil scientifique du programme ET

Le sujet de la mise à disposition des données de la recherche est sensible et d'actualité.

La normalisation des données est un problème technique majeur mais le manque de rationalisation n'est pas non plus une raison pour ne pas conserver et mettre à disposition les données.

La normalisation des données devrait être discutée lors de la création de bases de données pour permettre les comparaisons fiables. En effet, l' évolution d'une donnée n'est pas forcément significative si par ailleurs les facteurs qui influent sur sa collecte ont changé (ex : statistiques de pêche : les engins de pêche étant de plus en plus performants, des pêches de plus en plus fructueuses ne signifient pas des stocks de plus en plus prospères. Dans ses statistiques, la FAO tient compte de ces biais).

Des sources de données existent en d'autres endroits que dans les organismes de recherche et notamment dans les cabinets d'études qui opèrent de nombreuses études d'impacts sur l'environnement.

Enfin, il apparaît être du rôle de programmes comme « écosystèmes tropicaux » d'alerter les ministères des questions soulevées par cet objet.

Résumés des discussions

Ce sujet est d'actualité notamment en raison de la Convention Aarhus signé par la France.

1. **Au niveau conceptuel, les chercheurs sont à la fois inquiets (peur de diffuser leurs propres données avant leur usage) et enthousiastes (possibilité d'obtenir des données nombreuses rapidement) vis-à-vis de cette idée. On peut résumer les points de vue par les idées suivantes.**

Intérêts de la mise à disposition

- accès à tous
- mutualiser les coûts de collectes d'information
- Le partage de données limite les risques de perte de données
- permet de travailler sur des modèles puissants (basé sur un grand nombre de jeux de données), de comparer des séries de données dans l'espace et dans le temps
- …

Elle pose la question de la propriété des données ?

- Le chercheur
- Le financeur
- L'éditeur (d'un article)
- Le pays d'origine des données acquises

2. **Au-delà du concept, il existe de nombreuses questions techniques et d'objectifs auxquelles il faut répondre avant de se lancer dans un projet de base de données. Parmi celles-ci les suivantes ont été identifiées lors de la table ronde.**

Quels types de données (chacun ayant des avantages et des défauts) ?

- Données élémentaires
- Données de synthèse
- Métadonnées

Des données pour qui ?

- Pour les gestionnaires
- Pour les chercheurs
- Pour les décideurs
- Pour le grand public

Des données pour des projets bien identifiés (sans que ce soit limitatif) ?

Des données produites par qui ?

- Des chercheurs
- Des bureaux d'études
- Des naturalistes
- Des organismes ou réseaux spécialisés dans la collecte de données (ex : IFN, Renecofor, …)

Les opérateurs

- Gestion centralisée : confier cette mission à un organisme privé ,d'Etat ou international, qui gérerait des ensembles de données écologiques
- Gestion décentralisée : laisser à chaque possesseur de données la responsabilité de la création / de la participation à une base de données
- …

Conséquences de la mise à disposition de données

- Elle a un coût qu'il faut pouvoir assumer sur le long terme, y compris dans la nécessaire transcription à chaque mutation technologique.
- …

Quelques clés de réussite (ou d'évitement d'échec)

- Adapter les bases de données au public visé
- Etre capable de donner des informations concernant la qualité des données (protocole de relevés, …)
- Faciliter pour le chercheur la mise à disposition de ses données dans la base (tout en assurant une mise à jour des données de façon automatisée ?)
- Favoriser la diversité des initiatives

3. Quelques exemples

Les opérations du passés

- Inventaires du CTFT en Guyane ou en Afrique : données parfois perdues, seules les métadonnées restent parfois.
- Inventaires papetiers : données perdues
- GFIS : données rares et pas à jour
- Tropis : données rares et pas à jour

Les opérations en cours

- Données spatiales de Météo France, du CNES et du CNRS gérées par le GIP Médias France
- SINP : système d'information sur la nature et les paysages (métadonnées) piloté par le MEDD
- GBIF : base de données internationale sur la biodiversité, point de contact français au MNHN
- PEFAC : système de partage de rapports de recherche au Gabon
- Ca-SIF : projet de système d'information sur les sources d'information forestières (métadonnées)
- ORE : observatoires de recherche en environnement coordonné par le Ministère de la recherche
- ECOPLANT : Base de données écologique sur les forêts gérée par le LERFOB, UMR INRA-ENGREF…