VIII BASES DE DONNÉES

Coordinateurs : N. Barthes, M. Litaudon

L'atelier bases de données a été l'occasion de référencer les principales bases de données de nos disciplines et de permettre à leurs développeurs/gestionnaires de se rencontrer. En effet, les bases de données sont principalement conçues comme des outils de collaboration au sein d'une équipe et leurs utilisations sont généralement spécialisées dans un domaine de recherche. Ceci explique leur faible visibilité par la communauté scientifique au sens large. Cinq bases de données principales ont été identifiées dans le cadre des travaux en Écologie chimique :

L'extractothèque de l'ICSN de Gif-sur-Yvette

(http://www.icsn.cnrs-gif.fr/article.php3?id_article=144).

Cette base informatique, créée en 2005, recense une collection d'extraits naturels, préparés à partir de plantes supérieures provenant généralement des « points chauds de la biodiversité », répartis en plaques multi-puits. Pour chaque échantillon, sont répertoriées les données relatives (1) à l'organisme collecté : pays collaborant, taxonomie, date et lieu de prélèvement, photos, etc. ; (2) à la gestion des plaques multipuits : référencement, position des échantillons sur la plaque, déclaration de lots, déclaration des envois de plaques et retour des résultats ; (3) aux essais biologiques : cibles, domaine pharmacologique, « laboratoire propriétaire de la cible », résultats biologiques. A la mi-2011, l'extractothèque recensait approximativement 6500 plantes et 14000 extraits dont la plus grande partie était régulièrement transférée vers la Chimiothèque Nationale. L'objectif est d'entreprendre un certain nombre de criblages biologiques par la mise en œuvre d'essais robotisés et miniaturisés. Les extraits qui présentent une activité biologique significative sur une cible donnée sont étudiés pour en isoler les constituants responsables de cette activité. La base de données de l'extractothèque ICSN constitue aujourd'hui un outil de gestion et de recherche extrêmement puissant, mais limité toutefois aux seuls extraits des plantes supérieures. A moyen terme cette base pourra s'étendre à d'autres types d'extraits organiques comme des extraits issus de micro-organismes ou d'organismes marins. Il pourrait également être envisagé d'y intégrer les structures des molécules découvertes ainsi que l'ensemble des activités biologiques qui les concernent.

Cantharella de l'IRD à Nouméa

(http://cantharella.ird.nc).

Cette base de données opérationnelle depuis 2010 est particulièrement modulable. Elle a été pensée pour faciliter la gestion des différentes campagnes d'échantillonnage d'équipes travaillant sur des projets du milieu marin. L'outil a cependant été conçu de telle façon qu'il peut s'adapter aux produits naturels quelle qu'en soit l'origine. Bien que Cantharella référence le même type de données que l'extractothèque (taxonomie, gestion des échantillons – prélèvement, traitement, purification, analyses chimiques – et essais biologiques – cibles biochimiques, résultats d'analyses –), elle met fortement l'accent sur la gestion des droits d'accès aux données et permet d'ajuster très finement la diffusion des informations, que ce soit aux différents collaborateurs ou dans le cadre de restitutions vis-à-vis des pays, territoires, communautés d'où proviennent les orga-

nismes. Des développements de modules informatiques sont encore prévus (adjonction de documents, bibliothèque de molécule, SIG, etc.). Ceci étant, l'application Cantharella est pleinement fonctionnelle et actuellement utilisée pour partager et pérenniser l'information scientifique entre équipes d'un même projet. Elle permettra à terme la diffusion des données publiques. Il est à noter que l'application Cantharella sera prochainement diffusée sous licence libre (http://sourceforge.net/p/cantharella/home/Cantharella/), avec comme objectif d'élargir la communauté des utilisateurs et de rassembler des développeurs autour du projet.

La base de données collaborative du projet ANR Ecimar

Cette base de données a été initiée en 2007 pour réunir les données recueillies au cours des campagnes d'échantillonnage du projet ANR d'Écologie chimique marine ECIMAR. Pour chaque échantillon récolté, elle comporte des données de taxonomie, des illustrations des organismes *in situ* (éponges, cnidaires en majorité), et les signatures chimiques acquises selon plusieurs détecteurs. Elle regroupe aussi les informations concernant les lieux d'échantillonnage et la localisation précise des prélèvements. Les signatures chimiques regroupées dans cette base de données ont été acquises selon un protocole standardisé, et complétées par les caractérisations structurales de composés et des tests de bioactivité. Cette base a été alimentée par les différents partenaires du programme, et elle recense actuellement plus de mille échantillons prélevés sur environ 200 espèces marines. La structure de la base de données est évolutive, et elle va d'ailleurs servir à la création d'une base de données « fille » pour la gestion d'autres programmes de recherche (e.g. projet européen « Bammbo »).

L'EcoChimiothèque du GDR 2827 d'Écologie chimique

(http://ecochimiotheque.cefe.cnrs.fr).

Cette base de connaissances, initiée en 2010, a pour objectif de fédérer, classer et permettre d'analyser les informations biologiques, chimiques et physico-chimiques disponibles dans le domaine des médiateurs chimiques. Elle permet aux chimistes et écologues de partager leurs informations liées à la taxonomie, aux lieux et techniques de prélèvement ou de traitement de l'échantillon ainsi que les données physico-chimiques, analyses chimiques effectuées sur ces échantillons, disponibilité et accessibilité par synthèse chimique. Développée sur la base de l'application de « chimiothèque départementale » de la Chimiothèque Nationale (pour lui assurer une pleine compatibilité), elle est actuellement en phase de tests et devrait être alimentée dès 2012.

La Chimiothèque Nationale

(http://chimiotheque-nationale.enscm.fr).

Née au début des années 2000 de plusieurs initiatives individuelles, la Chimiothèque Nationale est aujourd'hui un GIS, auquel adhèrent 34 établissements d'enseignement supérieur ou organismes de recherche (Directeur : Marcel Hibert). Elle est représentée dans cet atelier par Philippe Jauffret, directeur de l'UGCN (UPS 3035 du CNRS), unité support du GIS CN. La Chimiothèque Nationale a pour mission principale de regrouper les collections de produits de synthèse, de composés naturels et d'extraits naturels existant dans les laboratoires publics français et d'en promouvoir la valorisation scientifique et industrielle. Les informations (structure chimique, propriétés physico-chimiques, données pharmacologiques, etc.) concernant les molécules et extraits disponibles sont regroupées dans deux bases de données interrogeables via le WEB. En juin 2011, la CN référençait 46000 substances et 14500 extraits, tous libres de droits et disponibles. Ces produits sont conditionnés en vrac ou en microplaques de 96 puits (ou les deux), afin de permettre leur évaluation biologique, ciblée ou systématique. L'ICSN et le MNHN sont les principaux contributeurs des collections de composés et d'extraits naturels de la Chimiothèque Nationale.

Thématiques émergentes

En Écologie chimique comme dans beaucoup de domaines, les bases de données sont devenues un outil central et essentiel dans la gestion des données scientifiques. Les technologies actuelles permettent de stocker des informations de nature très variable, de les rendre accessible à distance et ainsi de les partager immédiatement avec des collaborateurs. Ces progrès technologiques ont permis à diverses équipes d'organiser plus efficacement leur programme de recherche en permettant un meilleur suivi de leur état d'avancement ou parfois en offrant un outil d'aide à la rédaction de rapports. En tâche de fond, la centralisation des données au sein de bases de données permet une pérennisation de ces données scientifiques par la sauvegarde massive et automatisée rendue possible par l'accroissement des capacités de stockage des disques durs. Il faudra tout de même veiller à mettre à jour les formats de données afin de permettre leur lecture dans les années futures et ne pas se laisser dépasser par l'évolution des normes de fichiers.

Une contrepartie importante à ces déploiements facilités de bases de données est que chacun a pu, personnellement ou dans le cadre de son équipe de travail, développer sa propre base conduisant a un morcellement de l'information au sein de multiples bases de petite taille. De plus, si les Systèmes de Gestion de Bases de Données (SGBD) sont généralement homogènes et principalement articulés autour de la norme SQL, les interfaces de connexion font appel à toute une multitude de technologies et de langages informatiques.

Un second écueil des progrès technologiques apportés aux méthodes de travail en recherche scientifique (dans tous leurs aspects) est que la quantité de données à traiter est devenue conséquente. Pour faire face à ces flots d'informations et arriver à les interpréter de manière pertinente, des outils statistiques sont en cours de développement ou d'utilisation pour les plus avancés (MSEASY ou XCMS par exemple pour le traitement assisté de données d'analyses chimiques). Ces outils doivent permettre de gérer au mieux les différentes sources de données et se confrontent souvent à un obstacle majeur : il n'y a pas de norme associée au format des fichiers de sorties générés par les machines, en particulier, pour ce qui nous concerne, dans le domaine de l'analyse chimique ou du séquençage.

Par ailleurs, le rapprochement de ces bases en Écologie chimique avec l'initiative d'inventaire et de gestion des bases de données réalisée par INEE est en cours, en particulier avec la création depuis septembre 2011 d'une Unité Mixte de Service (UMS 3448, CNRS-MNHN) Bases de données Biodiversité, Ecologie, Environnement et Sociétés (BBEES).

Barthes N., Jauffret P., Litaudon M., Petek Sylvain, Thomas O. (2012)

Bases de données

In : Hossaert M. (ed.), Barthes N. (ed.), Abbadie L. (collab.), Al-Mourabit A. (collab.), Bagnères A.G. (collab.), Caissard J.C. (collab.), Grison C. (collab.), Le Bris N. (collab.), Leblanc C. (collab.), Litaudon M. (collab.), Lucas C. (collab.), Pérez T. (collab.), Potin P. (collab.), Rebuffat S. (collab.), Schatz B. (collab.), Smadja C. (collab.) Prospective écologie chimique. Paris : CNRS, 36-38

(Les Cahiers Prospectives - CNRS). Ateliers de Prospectives sur l'Ecologie Chimique : Atelier 8. Bases de Données, Rennes (FRA), 2009.