Une expérience de constitution d'un système d'information multi-sources pour l'étude de la qualité de l'eau

Agnès Braud, Sandra Bringay, Flavie Cernesson, Xavier Dolques, Mickaël Fabrègue, Corinne Grac, Nathalie Lalande, Florence Le Ber, Maguelonne Teisseire

Une expérience de constitution d'un système d'information multi-sources pour l'étude de la qualité de l'eau

Agnès Braud [1], **Sandra Bringay** [4], **Flavie Cernesson** [2], **Xavier Dolques** [1], **Mickaël Fabrègue** [1,2], **Corinne Grac** [3], **Nathalie Lalande** [2], **Florence Le Ber** [1], **Maguelonne Teisseire** [2]

1. *ICube, Université de Strasbourg, ENGEES, CNRS, Strasbourg, France*
   agnes.braud@unistra.fr,{xavier.dolques,florence.leber}@engees.unistra.fr
2. *TETIS, IRSTEA, AgroParisTech, Montpellier, France*
   prenom.nom@teledetection.fr
3. *LIVE, Université de Strasbourg/ENGEES, CNRS, Strasbourg, France*
   corinne.grac@engees.unistra.fr
4. *LIRMM, Université Montpellier 3, CNRS, Montpellier, France*
   sandra.bringay@lirmm.fr

RÉSUMÉ. *Pour mieux appréhender le fonctionnement des hydro-écosystèmes, sont disponibles des données nombreuses et diverses : données relatives à la qualité de l'eau ou aux stations de mesures, données décrivant le réseau hydrographique, etc. Ces données spatialement définies sont complexes à structurer et à relier de par leur volume et leur nature variables. Elles ont des origines, des valeurs, des structures spatiales et des répartitions temporelles diverses. Cet article fait l'état des problématiques rencontrées dans la collecte et la structuration de ces données, pour les deux zones étudiées, les districts Rhin-Meuse et Rhône-Méditerranée et Corse.*

ABSTRACT. *To better understand hydrosystem functioning, several and various data can be used: data on water quality, data characterizing sampling reaches, data describing the hydrographic network, etc. All these data are spatial and complex to structure and to interconnect because of their volume and their nature. They are characterized by a high heterogeneity due to their origins, their values, their spatial structures and their temporal variability. This article reports problems encountered for data gathering, modeling and integration. The inventory is carried out on two french districts: Rhine-Meuse and Rhône-Mediterranean and Corsica.*

MOTS-CLÉS : *Base de données intégrée, retour d'expérience, données multi-sources, hydrologie*

KEYWORDS: *Integrated Database, Feedbacks, Multi-source data, Hydrology*

DOI:10.3166/HSP.volume.1-15

1. Introduction

Il existe des méthodes génériques pour le développement de systèmes d'information (SI) pour l'environnement. On citera par exemple la méthode proposée par (Lemoisson *et al.*, 2010) pour construire des systèmes d'information dans le cadre d'observatoires des activités agricoles. Ces méthodes se déclinent en différentes étapes et s'appuient sur un ensemble d'outils logiciels. Toutefois leur mise en œuvre soulève la plupart du temps des difficultés propres à la réalité du terrain, existence et accessibilité des données, disponibilité et implication des acteurs fournisseurs de l'information, diversité des besoins des utilisateurs, etc. Nous allons ici traiter des problématiques liées à la collecte de données multi-sources.

Cet article rend compte d'une expérience de constitution d'un système d'information, menée dans le cadre du projet ANR Fresqueau [1], et qui s'applique à la problématique de l'évaluation de la qualité des hydroécosystèmes, fondée sur des expertises en hydrologie et hydro-écologie. Ce système contiendra une base de données intégrée, qui fait l'objet de cet article, ainsi qu'un ensemble d'outils permettant l'exploration, la visualisation et l'analyse des données ainsi rendues disponibles. Les utilisateurs sont des chercheurs en hydro-écologie et des ingénieurs-experts de bureaux d'études travaillant dans ce domaine. Une précédente expérience concernant les milieux aquatiques de la plaine d'Alsace a été décrite dans (Grac *et al.*, 2011 ; Braud *et al.*, 2011). L'expérience présentée ici est plus ambitieuse en termes de volume et de variété des données et informations considérées. Elle porte en effet sur deux grands bassins, correspondant aux districts Rhin-Meuse (33.000 km^2) et Rhône-Méditerranée et Corse (130.000km^2), pour la période 2002-2010. De plus cinq catégories de données sont considérées, au lieu de deux dans le précédent projet :

(i) les données relatives à la qualité de l'eau, bioindicateurs et paramètres physico-chimiques, permettant de qualifier de façon détaillée et complémentaire la qualité des cours d'eau,

(ii) les données relatives aux stations de mesures, traduisant la complémentarité des informations issues des différents réseaux,

(iii) les données décrivant le réseau hydrographique, afin de comparer ou compléter l'influence de la station de mesures et de ses caractéristiques,

(iv) les données relatives aux activités humaines pour estimer les pressions anthropiques ponctuelles et diffuses qui s'exercent sur les cours d'eau,

(v) les données relatives aux variables de forçage climatique ou de contexte afin de caractériser l'environnement des rivières et des points de prélèvements.

Le plan de l'article est le suivant. Le projet Fresqueau est brièvement présenté (section 2), puis nous détaillons les différentes catégories de données collectées et les problématiques posées par ces données (section 3). La section 4 présente les étapes de modélisation et d'intégration des données. Dans la section 5 notre travail est situé

1. http://engees-fresqueau.unistra.fr

et discuté par rapport aux travaux voisins. Finalement nous dressons quelques conclusions et perspectives au travail présenté ici.

2. Le projet Fresqueau

L'objectif de préserver ou restaurer le bon état des masses d'eau, imposé par la Directive Cadre Européenne sur l'eau (The European Parliament and the Council, 2000), met en exergue la nécessité de disposer d'outils opérationnels pour aider à l'interprétation des informations complexes concernant les cours d'eau et leur fonctionnement, ainsi que pour évaluer l'efficacité des programmes d'actions engagés. Dans cette perspective, le projet Fresqueau a pour but de développer de nouvelles méthodes pour étudier, comparer et exploiter l'ensemble des paramètres disponibles concernant l'état des cours d'eau et de leur environnement.

Plus précisément, le projet a pour objectif de traiter deux enjeux spécifiques : (1) mettre en évidence des liens entre différentes métriques permettant de caractériser la qualité des cours d'eau et (2) relier les sources de pressions sur le milieu à la qualité physico-chimique et biologique des cours d'eau. Pour cela, il a été nécessaire de constituer une base de données spécifique à partir d'un ensemble de données relatives à la qualité de l'eau, l'hydrologie, les stations de mesures, etc. mais également des données permettant de caractériser l'environnement des cours d'eau. Les données collectées sont caractérisées par une grande hétérogénéité en raison de leur origine (valeurs de mesures ou d'expertise), des objectifs qui ont conduit à leur acquisition (suivi à long terme, référentiel, rapportage européen, études ponctuelles, etc.). À cette hétérogénéité, se rajoutent la diversité de leurs valeurs (quantitative, semi-quantitative ou qualitative), leur variabilité temporelle (fréquence et durée de l'échantillonnage) et leur structure topologique (spatiale ou non). Nous avons également constaté des évolutions des protocoles et des formats sur la période d'étude 2002-2010. Toutes ces données sont localisées et sont associées à des objets spatiaux sous forme de points, lignes ou surfaces. Tous ces facteurs ont rendu délicates et complexes la structuration et l'interconnexion des données.

La première étape du projet s'attache au recensement puis à la structuration et à la mise en forme des données dans une base de données intégrée. Les étapes suivantes, non décrites ici, concernent le développement d'un entrepôt de données et la mise en œuvre de différentes approches de fouille pour explorer les données collectées, avant d'aboutir à un système d'aide à l'interprétation du fonctionnement des cours d'eau, rassemblant les différents éléments.

3. Sources de données

Les zones concernées par notre étude sont représentées sur la figure 1. Elles couvrent les districts de l'agence de l'eau Rhin-Meuse dans le nord-est de la France et de l'agence de l'eau Rhône-Méditerranée et Corse dans le sud-est. Pour ces deux zones, nous décrivons successivement les différentes catégories de données, leurs principales

caractéristiques, leurs producteurs, leurs protocoles d'acquisition et de bancarisation ainsi que leurs conditions d'accès. On remarquera la grande variabilité des souces, qui dépendent souvent de plusieurs producteurs, et dont les objectifs diffèrent dans le temps et dans l'espace.

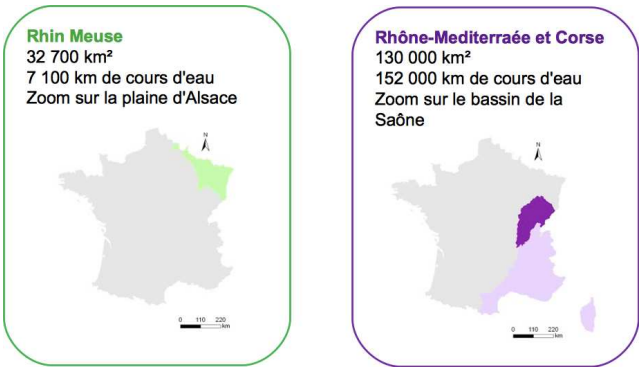

FIGURE *1. Localisation des deux bassins d'étude*

3.1. Paramètres de qualité de l'eau et des milieux aquatiques

L'état du cours d'eau, comprenant l'eau et les milieux aquatiques, se décline en trois parties :

1. L'état physico-chimique de l'eau et des sédiments ; il est subdivisé en un état physico-chimique soutenant la biologie, qui se traduit par les paramètres caractérisant les macropolluants (d'origines naturelles ou anthropiques), et en un état chimique, qui se traduit par des paramètres caractérisant les micropolluants (pesticides, ...) toujours d'origines anthropiques ;

2. L'état des peuplements biologiques floristiques (macrophytes et diatomées) et faunistiques (invertébrés, poissons) : cet état est mesuré par des échantillonnages des peuplements et synthétisé dans des indices biologiques, parmi lesquels l'indice biologique global normalisé (AFNOR, 2004) est le plus fréquemment utilisé ; les caractéristiques (traits de vie) des taxons sont également disponibles ;

3. L'état physique : il s'agit de l'hydromorphologie du cours d'eau, soit l'état des berges, du lit mineur, du lit majeur, des continuités longitudinales, latérales et verticales, des annexes, des conditions hydrologiques (débits) et hydrauliques (vitesse, géométrie du cours d'eau).

Majoritairement, les données de qualité d'eau sont issues de résultats d'analyse de prélèvements dont les protocoles sont normalisés. Toutefois ces protocoles, ainsi que les réseaux de mesures, dont les données sont issues pour la plupart, ont évolué

au cours du temps. La périodicité et la densité spatiale des mesures est variable selon les paramètres étudiés, les producteurs et les objectifs recherchés. Les producteurs de données sont principalement les agences de l'eau et services de l'Etat (DREAL, ONEMA) [2], des collectivités, des bureaux d'études mandatés, des laboratoires de recherche. Les objectifs relèvent de réseaux de surveillance ou d'études et recherche. Les données issues des réseaux de surveillance nationaux sont, jusqu'à présent, bancarisées au niveau de chaque district par l'agence de l'eau correspondante, suivant le format national du SANDRE (Service d'Administration Nationale des Données et Référentiels sur l'Eau).

L'accès aux données est relativement aisé pour les indices biologiques et paramètres physico-chimiques, plus complexe en ce qui concerne les listes floristiques et faunistiques. Les premiers sont accessibles sur demande ou téléchargeables *via* le portail Eau France [3], ou sur le site de l'ONEMA. Les secondes, accessibles *via* les DREAL, sont de formats très hétérogènes et difficilement intégrables à une base de données sans traitements préalables. D'autres données dépendent de fournisseurs locaux (laboratoires de recherche, réseaux de surveillance spécifiques). Le tableau 1 recense les différentes sources de données, les modes d'accès à ces données et leurs conditions d'utilisation.

Les paramètres physiques, ou hydromorphologiques (dimensions et forme du lit, caractéristiques du substrat, état des berges), déterminent l'état physique du cours d'eau. Les analyses de terrain peuvent être des observations d'experts, ou de non-experts, des mesures (vitesse du courant, profondeur maximale, etc.) ; ces observations ou mesures, suivant le cas, peuvent être agrégées sous forme d'indices numériques, de formules et d'intervalles variables. Par exemple, l'agence de l'eau Rhin-Meuse propose un indice QUALPHY, variant de 0 à 100, et qui repose sur trois variables caractérisant le lit majeur, les berges et le lit mineur. Les données des métriques et de la note QUALPHY sont téléchargeables sur le site de cette agence.

3.2. Caractéristiques des stations de mesures

Les caractéristiques des stations de mesures comprennent les informations liées à leur dénomination, à leur localisation et à leur objectif. Ces informations peuvent être complétées par des informations synthétiques rendant compte du contexte hydrologique ou environnemental. Les données ne sont pas toutes au format national SANDRE. De plus, certaines données sont codées dans le système de projection Lambert 2 étendu alors que le référentiel actuel est le système Lambert 93.

Une station de mesures est rattachée à un unique point géographique. En biologie, excepté pour les diatomées dont l'échantillonnage, très localisé, peut être assimilé à un point, les échantillonnages se font sur des tronçons de rivières – dénommés points de

2. DREAL: Direction Régionale de l'Environnement, de l'Aménagement et du Logement ; ONEMA : Office National de l'Eau et des Milieux Aquatiques.

3. http://www.eaufrance.fr

Tableau 1. Type, extension, objectif, origine et conditions d'utilisation (O pour oui, N pour non, nr pour non renseigné) des données collectées

Nom des bases	Catégorie de données					Couverture			Objectifs					Accès			Producteur(s)
	Qualité eau	Station mesure	Réseau hydrographique	Activités humaines	Forçage / Contexte	Nationale	District	Locale	Référentiel	Suivi long terme	Études ponctuelles	Recherche	Rapportage Europe	Gratuite	Usage commercial	Licence *Creative Commons*	
Image	×	×				×				×			×	O	N	O	ONEMA
Préfiguration Naïades	×	×				×				×	×		×	O	N	O	ONEMA et partenaires
Syrah			×	×	×	×			×					O	N	-	MEDDE, AE, ONEMA, IRSTEA
Qlté AE RMC	×	×					×			×	×		×	O	N	-	AE RMC
Qlté AE RM	×	×					×			×	×		×	O	N	-	AE RM
Qualphy	×						×			×				O	nr	nr	AE RM et collectivités
Rivières	×	×						×				×		O	N	O	LHYGES – ENGEES
RID 67	×	×						×		×	×			O	N	-	CG Bas-Rhin
Topo			×	×		×			×					O	N	N	IGN
Masse d'eau			×			×			×				×	O	N	-	AE et ONEMA
Carthage			×		×	×			×	×				O	N	-	IGN et AE
CLC				×		×			×					O	O	-	Union Européenne – SOeS
ROE				×		×			×					O	N	O	ONEMA et partenaires
RPG				×		×				×			×	N	O	-	ASP et Ministère de l'Agriculture
Climat					×	×						×		O	N	-	Théma – Univ. Besançon
Saône				×				×				×		O	N	-	TETIS – IRSTEA, AE RMC
HER					×	×			×				×	O	N	-	MEDDE et IRSTEA
Step RMC				×			×			×				O	N	-	AE RMC, MEDDE - DEB
Step RM				×			×			×				O	N	-	MEDDE - DEB
Geofla					×	×			×					O	O	-	IGN
Hydro		×			×	×				×				O	N	N	MEDDE - DEB
21sources	7	8	4	8	6	13	5	3	8	11	4	3	7	20	3	4	15 producteurs

prélèvement – délimités par des coordonnées amont et aval, et pouvant ne pas inclure les coordonnées de la station à laquelle ils sont rattachés, pour des raisons techniques (accès, faciès d'écoulement, etc.). Les stations de pêche sont caractérisées par un code station spécifique et stockées dans la base de données IMAGE de l'ONEMA.

Il existe de nombreux (plusieurs centaines) réseaux de mesures de la qualité de l'eau en France. Ils peuvent être permanents ou temporaires ; étendus sur tout le territoire national ou locaux, portés par des établissements publics ou privés. Enfin ces réseaux peuvent faire l'objet d'un suivi plus ou moins régulier. Les données concernant les stations des réseaux nationaux gérés par les agences de l'eau sont normalement aisément accessibles comme décrit ci-dessus (cf. tableau 1).

3.3. Réseau hydrographique

Concernant le réseau hydrographique, trois sources de données sont disponibles : la BD TOPO®, déjà citée, la BD Carthage® et le réseau Syrah. La BD Carthage® recense une information complète du réseau hydrographique réalisée à partir de la couche hydrographie de la BD CARTO® de l'IGN, enrichie par les agences de l'eau. De plus, elle offre un découpage en aires hydrographiques, non relié aux réseaux de mesures, et une représentation des masses d'eau (partie distincte et significative des eaux de surface). Le réseau Syrah, quant à lui, est composé de tronçons de cours d'eau géomorphologiquement homogènes ; il est produit par l'ONEMA et IRSTEA en collaboration avec les agences de l'eau. Ces deux dernières bases sont accessibles gratuitement sur demande.

3.4. Activités humaines

Les activités humaines se traduisent par des pratiques de prélèvement de l'eau, de rejet dans le milieu naturel, et de modification physique des milieux, sous la forme de pressions positives ou négatives exercées sur le milieu. Ces pressions peuvent être intermittentes (rejet d'une industrie), ou permanentes (un seuil barrant le cours d'eau). Les pressions peuvent aussi être diffuses, si elles sont liées à des processus de diffusion issues de sources surfaciques (épandages agricoles, eaux de ruissellement).

L'environnement des stations de mesures et les pressions diffuses liées à l'occupation du sol sont acessibles *via* trois bases de données complémentaires :

1. CORINE Land Cover, produite par le ministère en charge de l'écologie, qui met à disposition un inventaire de l'occupation du sol, issu de photo-interprétation de données satellitaires ;

2. une sélection de la BD TOPO®, base de données vectorielles de référence produite par l'IGN (réseaux routier et ferroviaire, bâtiments, végétation arborée, etc.) ;

3. le registre parcellaire graphique permettant de préciser les espaces agricoles. Il s'agit de données collectées dans le cadre des déclarations par les agriculteurs des

surfaces relevant de la politique agricole commune. Les données anonymisées sont accessibles *via* une licence payante.

Ces informations peuvent être complétées par les fichiers de rejets disponibles auprès des agences de l'eau. Enfin, l'information concernant les obstacles aux écoulements (barrages, seuils, etc.), collectée par les acteurs de l'eau et de l'aménagement du territoire, est recensée dans la base de données ROE (Référentiel des Obstacles à l'Écoulement), sous licence ouverte. Les localisations s'appuient sur le réseau hydrographique de la BD TOPO®.

3.5. Variables de forçage ou de contexte

Les variables de forçage ou de contexte mobilisées sont de différentes natures : données hydrologiques, données climatiques, hydro-écorégions (HER, régions homogènes pour les processus physiques dominants) mais aussi données administratives. Différentes bases de données les recensent : les premières relèvent de la base de données nationale HYDRO, administrée par le SCHAPI (Service Central d'Hydrométéorologie et d'Appui à la Prévision des Inondations) pour le compte du ministère en charge de l'écologie ; elles sont accessibles sous licence. Les données climatiques sont des synthèses sous la forme d'une typologie et d'un zonage des climats pour la France métropolitaine réalisées par des équipes de recherche. Ces données sont téléchargeables gratuitement. Les données concernant les hydro-écorégions, produites par IRSTEA sont également disponibles gratuitement, sur demande. Enfin les données administratives sont bancarisées dans la base GEOFLA®, produite par l'IGN, et accessibles gratuitement, sous licence ouverte.

4. Modèle et intégration des données

4.1. Modèle

Le modèle développé s'est appuyé en grande partie sur les modèles des bases sources. Nous donnons ci-dessous une vision globale de la base, centrée autour de la table concernant les stations de mesures (figure 2). On y retrouve les différents thèmes recensés, (i) qualité de l'eau, (ii) caractéristiques des stations, (iii) réseau hydrographique, (iv) activités humaines, (v) variables de forçage ou de contexte.

Une vision partielle du modèle concernant l'hydrographie est présentée sur la figure 3, en utilisant la notation Merise du formalisme entité-association. Pour ces figures, nous utilisons les pictogrammes spatiaux de PictograF[4] pour représenter les caractéristiques spatiales des objets ayant une géométrie. La station de mesure est ainsi associée à une forme ponctuelle, le cours d'eau à une forme linéaire et le bassin versant à une forme surfacique, toutes dans un univers de dimension 2. La relation en pointillés ne fait pas partie du modèle initial, construit sur la base des données

4. http://pictograf.scg.ulaval.ca/

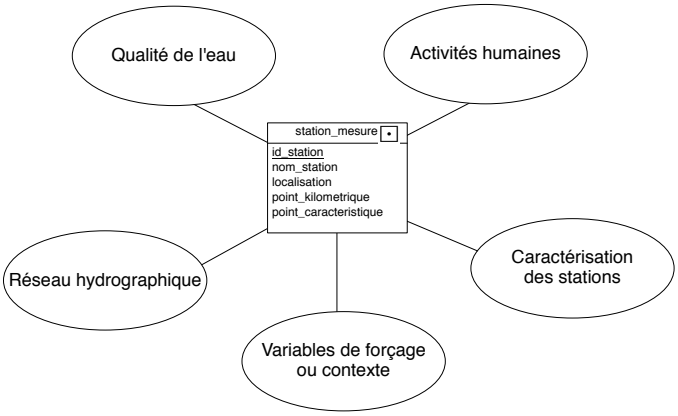

FIGURE 2. *Modèle de données global*

collectées, mais est calculée à partir des géométries des cours d'eau et des bassins versants afin de faire le lien entre les deux types d'objets lors de nos analyses. On remarque sur ce modèle plusieurs tables `tronçon_*` qui correspondent chacune à une des sources représentant cette information (BD TOPO®, BD Carthage® et réseau Syrah, voir section 3).

Le modèle comporte également des tables traitant de la qualité des données, appuyées sur la norme ISO 19115, qui définit les métadonnées de l'information géographique. Différentes méthodes ont été établies pour évaluer les différentes dimensions de la qualité des données à intégrer : *(i)* le comptage des valeurs manquantes pour la complétude des données ; *(ii)* la prise en compte des contraintes du domaine (contrôles de vraisemblance) ; *(iii)* la satisfaction de contraintes logiques pour la précision logique (contrôles de cohérence) ; *(iv)* la satisfaction de contraintes pour les précisions temporelles et spatiales. Ces méthodes ont été testées sur une partie de la base. De plus, des traitements pour compléter les données manquantes ont été expérimentés sur une partie des données concernant les paramètres physico-chimiques (Serrano Balderas *et al.*, 2014).

4.2. Intégration

Les différentes sources de données impliquées dans le projet ont dû être unifiées, structurées et intégrées. Les problématiques d'intégration sont bien sûr liées aux caractères hétérogènes et aux incomplétudes existant dans les sources de données. Le caractère spatio-temporel a ajouté de la complexité dans le processus d'alimentation de la base de données Fresqueau. En effet, comme indiqué précédemment, le concept

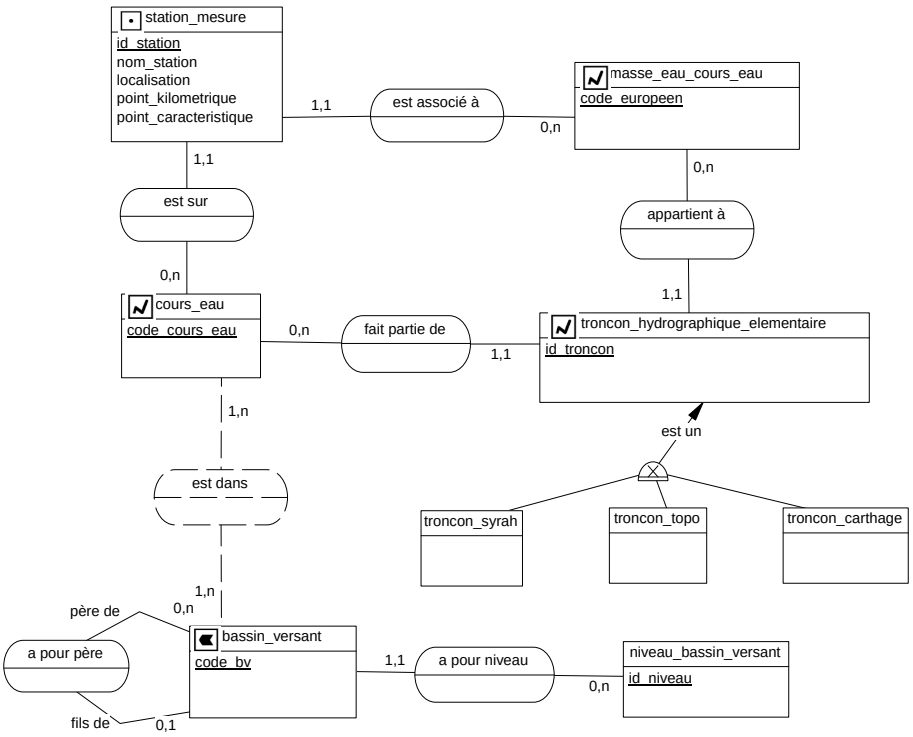

FIGURE *3. Modèle de données concernant les réseaux hydrographiques*

de station de mesure est central dans notre modèle, et l'intégration des données correspondantes repose sur plusieurs réseaux de mesure, afin de bénéficier de l'ensemble de ces informations. Cependant le rattachement des stations de qualité d'eau aux différentes données de réseaux hydrographiques n'est pas direct. Par définition, les stations de qualité d'eau sont rattachées aux masses d'eau qui servent de support pour le rapportage européen. Par ailleurs, les tronçons de la BD Carthage®, qui représentent un référentiel pour les agences de l'eau, sont beaucoup plus nombreux que ceux des masses d'eau. La mise en relation des stations de qualité d'eau et des tronçons de BD Carthage® a dû être réalisée et a nécessité le développement de requêtes spatiales et attributaires. Les correspondances spatiales exactes ne couvraient qu'un très faible pourcentage des stations : 1%. Pour seulement 85% des 7975 stations étudiées, nous disposions de suffisamment d'informations pour mettre en correspondance les stations et les tronçons Carthage via des requêtes intégrant les points kilométriques, les codes Cgenelin (code hydrographique du cours d'eau) et un *buffer* d'environ 10 mètres. Nous avons alors pu relier plus de 90% d'entre elles sur les deux districts étudiés (98% des stations reliées l'ont été avec un *buffer* de 50 mètres). Néanmoins 15%

des stations de la base initiale n'ont pas pu être mises en lien avec les tronçons du réseau hydrographique.

D'autres difficultés étaient liées à l'intégration des données sur l'état des peuplements biologiques. Tout d'abord, le référentiel pour les taxons est mis à jour régulièrement. Un taxon donné peut ainsi faire l'objet d'une modification et se voir attribuer un nouveau code. Dans le cadre de notre projet, nous avons utilisé la dernière mise à jour disponible au moment de l'intégration. L'intégration des données sur les traits de vie des taxons, issues de la base Rivières et fondées sur une mise à jour antérieure du référentiel, a donc nécessité un travail de mise en correspondance des codes taxons de la base Rivières avec ceux de la base Fresqueau. Cette mise en correspondance était partiellement automatique, mais a également nécessité un travail manuel fastidieux. Notons que ceci pose un problème pour la mise à jour de notre base qui, pour ce qui concerne les données liées aux taxons, ne pourra pas être un processus complètement automatisé. Par ailleurs, les listes des taxons (invertébrés, poissons, macrophytes, diatomées), identifiés et dénombrés lors des prélèvements effectués sur les stations, sont disponibles sous la forme de fichiers autonomes, dont le nombre est très important. Ceci peut engendrer des erreurs ou des oublis, au moment de la saisie par les opérateurs, et conduit à une intégration complexe dans la base.

4.3. Etat de la base

Le modèle a été implanté en utilisant `PostgreSQL/PostGIS`. Les données collectées sont stockées dans 80 tables. Ces tables se répartissent dans les grands thèmes évoqués précédemment comme indiqué dans le tableau 2.

Tableau 2. Nombre de tables par catégorie

Catégorie	Nombre de tables	Nombre de sources
Qualité de l'eau	31	8
Stations de mesure	7	4
Activités humaines	25	8
Réseau hydrographique	8	4
Variables de forçage ou de contexte	10	13

Afin de donner un aperçu du volume de données, nous livrons une estimation du nombre de lignes de certaines tables. Pour les deux districts considérés, on trouve notamment plus de cinq cent milliers de lignes correspondant à des mesures climatiques, plus de quatorze millions de mesures pour la physico-chimie, plus de neuf millions d'exploitations dans le registre parcellaire graphique, plus de huit millions de bâtiments et plus d'un million de tronçons hydrographiques. De plus, vingt-deux des tables créées possèdent au moins un attribut représentant une géométrie.

La base est complétée par des tables calculées à partir de ces données intégrées. Les données contenues dans ces nouvelles tables permettent de faire le lien entre des objets spatiaux (comme un cours d'eau et son bassin versant) ou ont été jugées nécessaires à l'analyse par les experts du domaine, comme des données sur les pressions

présentes dans un périmètre donné autour d'une station de mesure. Ainsi vingt-cinq nouvelles tables capturant l'information sur les pressions (bâti, réseau routier, végétation, cultures agricoles, ...) se situant dans un rayon de 300 à 2000 mètres autour des stations ont été créées. Cette information est quantifiée par la surface concernée par la pression dans le voisinage de chaque station.

5. Travaux connexes

Les agences de l'eau ont développé des bases de données où sont recensées les informations sur les nombreuses stations qu'elles surveillent. Nous avons intégré celles de l'agence Rhin-Meuse et de l'agence Rhône-Méditerranée et Corse. Toutefois, même si elles recouvrent des zones géographiques étendues, les informations disponibles dans ces bases sont très limitées, en particulier, et comme nous l'avons fait remarquer ci-dessus, elles ne contiennent généralement pas les relevés taxonomiques établis sur les stations ; les données concernant les réseaux hydrographiques, les activités humaines et les variables de forçage, que nous avons regroupées dans la base Fresqueau, ne sont pas non plus présentes, ou de manière très partielle.

L'ONEMA pilote un double projet national : la constitution d'une banque de données « Naïades » – qui rassemblera, notamment, toutes les informations des bases de données des différentes agences de l'eau – et d'un outil d'interrogation de cette banque : le SEEE (Système d'évaluation de l'état écologique) capable de fournir des évaluations des masses d'eau consultées. Les autres données ne sont pas non plus prises en compte.

On citera également la base de données constituée dans le cadre du PIREN Seine (Programme Interdisciplinaire de Recherche sur l'Environnement de la Seine) [5]. C'est dans ce cadre qu'a été développé l'outil SENEQUE (Ruelland, 2004), une interface reliant un modèle générique de fonctionnement des cours d'eau à un système d'information géographique. Cette interface permet de sélectionner les informations nécessaires à la mise en œuvre du modèle. Elle est reliée à une base de données qui décrit la structure de ces informations : réseau hydrographique, contraintes de forçage, points de rejet, pollutions diffuses calculées à partir des informations d'occupation du sol. L'exemple traité dans l'article (Ruelland *et al.*, 2007) concerne la rivière Oise, au nord du Bassin Parisien, qui couvre un bassin versant de 17.000 km^2. Si la base est générale, les informations collectées sont limitées au cours d'eau étudié, et aux exigences du modèle utilisé, le modèle Riverstrahler, qui simule le fonctionnement biogéochimique des cours d'eau.

Hors domaine hydro-écologique *stricto sensu*, il existe différents travaux traitant de systèmes d'information et de données environnementales plus ou moins complexes. Par exemple, (Le Gal *et al.*, 2002) traite de la mise en place d'un SI dédié à la maintenance des réseaux hydrauliques au Niger. Dans (Mimouni *et al.*, 2007) est décrit

5. http://www.sisyphe.upmc.fr/piren_drupal6/?q=seneque

un système d'information géographique permettant de collecter et visualiser des données spatiales concernant la géologie et l'environnement (réseau routier et chemins, hydrographie, végétation et zones agricoles, etc.) au Maroc. Plus récemment, les travaux décrits dans (Vernier *et al.*, 2013) portent sur un SI, incluant un entrepôt de données, permettant de caractériser les activités agricoles dans des bassins versants, en relation avec des indicateurs de présence de pesticides. Ces travaux ne posent pas explicitement le problème de la collecte des données, probablement parce qu'il s'agit de travaux portant sur un territoire restreint et exploitant des données principalement issues de travaux de recherche.

A l'inverse, nous travaillons sur un grand ensemble de données, provenant de différentes sources et de ce fait présentant un certain nombre de difficultés, comme nous l'avons déjà souligné : incomplétude, imprécision, incohérence ... Notre objectif est en effet non pas d'alimenter et de tester un modèle, mais d'utiliser ces données dans un processus d'extraction de connaissances, en mettant en œuvre des méthodes de fouille de données (Dolques *et al.*, 2013 ; Fabrègue *et al.*, 2013) pour répondre à un ensemble de questions posées par les hydrologues et hydro-écologues.

6. Conclusion

Le travail décrit ici est une première étape du développement d'un système d'information pour l'étude et l'évaluation de l'état des écosystèmes aquatiques. Il a nécessité beaucoup de temps, temps consacré à collecter et comprendre les données, alors même que ces données sont réputées facilement accessibles. Il s'agit aussi d'un travail d'équipe, où les collaborateurs doivent disposer de et partager des métadonnées, des dictionnaires des données et un journal de bord. Finalement ce travail ne doit pas se faire sans explicitation des besoins des utilisateurs finaux. En effet, le choix des données conditionne les hypothèses et le domaine de l'analyse à mener sur ces données. Dans notre cas, un gros travail a été parallèlement mené pour expliciter les questions des hydrologues et les hydro-écologues. Par exemple, une question posée est de relier la valeur des bioindicateurs (IBGN ou autre) aux valeurs des paramètres physico-chimiques, sur une certaine période de temps précédant les prélèvements biologiques, et en prenant en compte certaines variables de contexte. A cette question est associé un sous-jeu de données, extrait de la base intégrée. Une autre question concerne les liens entre certains types de pressions (liées à l'occupation du sol, par exemple) et les paramètres physico-chimiques mesurés dans les stations situées à l'aval de ces pressions.

La diversité des données collectées, de leurs formats, la complexité des accès, nous ont conduits a développer une base intégrée plutôt que d'accéder "à la demande" aux bases existantes, comme l'autorise par exemple un outil de catalogage tel que MDWEB (Desconnets *et al.*, 2007). Ce choix se justifie aussi par le fait que la base intègre également des informations issues de fichiers disparates et que la mise en cohérence des données oblige à définir des tables supplémentaires. Enfin les méthodes d'analyse

que nous voulons mettre en œuvre nécessitent de disposer des données ensemble pour effectuer les prétraitements et sélectionner des sous-jeux adaptés aux questions posées.

Cette première étape du projet Fresqueau a été suivie du développement d'un entrepôt de données, permettant de parcourir et de visualiser les données selon différentes dimensions et différents niveaux d'agrégation (Boulil *et al.*, 2014). Des analyses faisant appel à différentes techniques de fouille de données sont en cours. Au-delà du travail mené sur les deux districts Rhin-Meuse d'une part et Rhône-Méditerranée et Corse d'autre part, une extension à l'échelle nationale est prévue prochainement. Ce sera l'occasion de valider la structure de la base et la procédure d'intégration mise en place.

Remerciements

Ce travail s'inscrit dans le cadre du projet ANR 11 MONU 14 Fresqueau. Merci aux différentes personnes impliquées dans le projet et aux organismes fournisseurs des données.

Bibliographie

AFNOR. (2004). *Qualité de l'eau : détermination de l'Indice Biologique Global Normalisé (IBGN).* XP T90-350.

Boulil K., Le Ber F., Bimonte S., Grac C., Cernesson F., Niel J. (2014). Multidimensional modelling and analysis of large and complex water quality data: an OLAP-based solution. *Ecological Informatics*. (En soumission)

Braud A., Nica C., Grac C., Le Ber F. (2011). A lattice-based query system for assessing the quality of hydro-ecosystems. In A. Napoli, V. Vychodil (Eds.), *Proceedings of the 8th International Conference on Concept Lattices and Their Applications (CLA 2001), Nancy*, p. 265–277. CEUR Workshop Proceedings.

Desconnets J.-C., Libourel Rouge T., Clerc S. (2007). Cataloguer pour diffuser les ressources environnementales. In *Inforsid 2007, Actes du XXVème congrès, Perros-Guirec*, p. 253–267.

Dolques X., Le Ber F., Huchard M., Nebut C. (2013). Analyse Relationnelle de Concepts pour l'exploration de données relationnelles. In F. S. Christel Vrain André Péninou (Ed.), *EGC'2013: 13e Conférence Francophone sur l'Extraction et la Gestion des Connaissances, Toulouse, France*, p. 121-132. Hermann-Éditions.

Fabrègue M., Braud A., Bringay S., Le Ber F., Teisseire M. (2013). OrderSpan: Mining Closed Partially Ordered Patterns. In *Advances in Intelligent Data Analysis XII, The Twelfth International Symposium on Intelligent Data Analysis (IDA 2013), London*, vol. LNCS 8207, p. 186–197. Springer.

Grac C., Braud A., Le Ber F., Trémolières M. (2011). Un système d'information pour le suivi et l'évaluation de la qualité des cours d'eau – Application à l'hydro-écorégion de la plaine d'Alsace. *RSTI - Ingénierie des Systèmes d'Information*, vol. 16, p. 9-30.

Le Gal P.-Y., Passouant M., Famanta M., Bélières J.-F. (2002). Conception et mise en place d'un système d'information dédié à la maintenance des réseaux hydrauliques à l'Office du

Niger (Mali). In P. Garin, P.-Y. Le Gal, T. Ruf (Eds.), *La gestion des périmètres irrigués collectifs à l'aube du XXIe siècle, enjeux, problèmes, démarches : actes de l'atelier du Pcsi, 22-23 janvier 2001*, p. 211–224. Montpellier, France, Cirad – Cemagref – IRD.

Lemoisson P., Passouant M., Foucher J.-F., Martinand P., Maurel P., Vinatier J.-M. *et al.* (2010). *Co-Obs : Méthode de conception collaborative d'observatoires*. Rapport technique. Cirad, Cemagref, Chambres d'Agriculture, RMT OAAT, APEM.

Mimouni N., Bouaziz S., Rebai N. (2007). Intégration des données géologiques et environnementales de la région de Monastir dans un SIG. In *SIG 2007, Conférence francophone ESRI*. ESRI France.

Ruelland D. (2004). SENEQUE, logiciel SIG de modélisation prospective de la qualité de l'eau. *Revue Int. de Géomatique*, vol. 14, p. 97–117.

Ruelland D., Billen G., Brunstein D., Garnier J. (2007). SENEQUE: A multi-scaling GIS interface to the Riverstrahler model of the biogeochemical functioning of river stystems. *Science of the Total Environment*, vol. 375, n° 2007, p. 257–273.

Serrano Balderas E. C., Berti-Equille L., Grac C. (2014). Data processing for controlling data quality on surface water quality assessment. In *Atelier "Systèmes d'Information pour l'environnement", Inforsid 2014, Lyon.*

The European Parliament and the Council. (2000). *Framework for Community action in the field of water policy.* Directive 2000/60/EC.

Vernier F., Miralles A., Pinet F., Carluer N., Gouy V., Molla G. *et al.* (2013). EIS Pesticides: An environmental information system to characterize agricultural activities and calculate agro-environmental indicators at embedded watershed scales. *Agricultural Systems*, vol. 122, p. 11–21.