# Un système décisionnel pour l'analyse de la qualité des eaux de rivières

Sandro Bimonte, Kamal Boulil, Agnès Braud, Sandra Bringay, Flavie Cernesson, Xavier Dolques, Mickaël Fabrègue, Corinne Grac, Nathalie Lalande, Florence Le Ber, et al.

# Un système décisionnel pour l'analyse de la qualité des eaux de rivières

Sandro Bimonte [1], Kamal Boulil [2], Agnès Braud [3], Sandra Bringay [4], Flavie Cernesson [5], Xavier Dolques [3], Mickaël Fabrègue [3,5], Corinne Grac [6], Nathalie Lalande [5], Florence Le Ber [3], Maguelonne Teisseire [5]

1. UR TSCF, IRSTEA, Aubière, France
   sandro.bimonte@irstea.fr
2. LI, Université de Tours, France
   boulilkamel@yahoo.fr
3. ICube, Université de Strasbourg, ENGEES, CNRS, Strasbourg, France
   agnes.braud@unistra.fr,{xavier.dolques,florence.leber}@engees.unistra.fr
4. LIRMM, Université Montpellier 3, CNRS, Montpellier, France
   sandra.bringay@lirmm.fr
5. TETIS, IRSTEA, AgroParisTech, Montpellier, France
   prenom.nom@teledetection.fr
6. LIVE, Université de Strasbourg/ENGEES, CNRS, Strasbourg, France
   corinne.grac@engees.unistra.fr

Ingénierie des systèmes d'information – n° 3/2015, 143-167

## 1. Introduction

L'objectif de préserver ou restaurer le bon état des masses d'eau a été imposé par la Directive cadre européenne sur l'eau (DCE) (PE&CE, 2000). Par là même, cette directive a mis en exergue la nécessité de disposer d'outils opérationnels pour aider à l'interprétation des informations concernant les cours d'eau et leur fonctionnement, ainsi que pour évaluer l'efficacité des programmes d'actions engagés. Dans ce contexte, le projet Fresqueau [1] s'est donné pour tâche de prendre en charge deux questions particulières : (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. Répondre à ces questions nécessite l'exploitation de plusieurs sources de données relatives à la qualité de l'eau, l'hydrologie, les stations de mesures, etc., mais également de sources de données permettant de caractériser l'environnement des cours d'eau. La complexité et l'hétérogénéité de ces données, leur caractère à la fois temporel et spatial, soulèvent plusieurs problématiques propres à la gestion et à l'analyse de données environnementales, et qui s'inscrivent plus largement dans le nouveau champs de recherche des science des données (Le Ber et al., 2014).

Les systèmes d'information décisionnels (SD) rassemblent un ensemble d'outils et méthodes pour la modélisation et la restitution des données dans le but d'aider les utilisateurs dans un processus décisionnel. Ces systèmes ont été utilisés avec succès dans différents domaines d'application comme la santé, le commerce et l'environnement. Ils se fondent sur une architecture multi-niveaux (tiers) qui est typiquement composée d'un système d'information (SI) central et d'un ensemble d'outils d'analyse comme les outils OLAP (On-Line Analytical Processing) ou de fouille de données. Les entrepôts de données sont des bases de données spécifiques, historisées, dédiées à l'intégration et au stockage de gros volumes de données (Inmon, 2005). Les entrepôts stockent ces données au niveau le plus fin et les organisent de façon à faciliter leur analyse et leur agrégation. Les outils d'analyse OLAP permettent une exploration interactive des données de l'entrepôt à différents niveaux de détail, selon une approche multidimensionnelle (Abelló et al., 2006 ; Malinowski, Zimányi, 2008). Ces outils agrègent les données de l'entrepôt dans des structures multidimensionnelles appelées cubes de données. Les dimensions sont organisées en hiérarchies de niveaux d'agrégation, permettant de calculer des indicateurs d'analyse à différentes granularités, et ainsi une exploration rapide de ces cubes.

La fouille de données et, plus largement, l'extraction de connaissances à partir de bases de données (ECBD), recouvrent des techniques d'exploration de données qui permettent d'en extraire des éléments, répétitions, règles, corrélations ou autres régularités, interprétables par un expert du domaine. L'ECBD se décompose en cinq étapes, (1) la sélection de données en vue de leur analyse, (2) le prétraitement, (3) la transformation, puis (4) la fouille de ces données, avant (5) la restitution des résultats

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

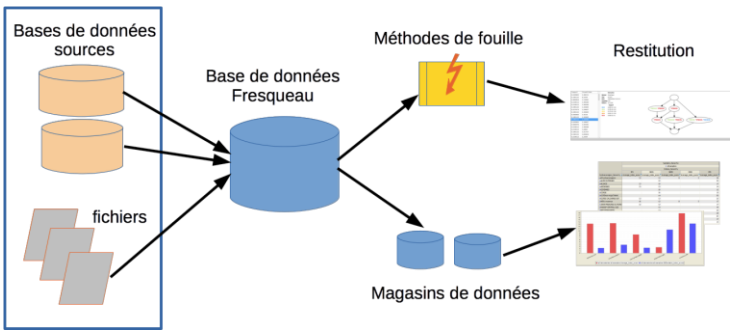


Figure 1. Architecture globale du système

à l'analyste (Fayyad et al., 1996). Les méthodes de fouille développées portent sur différents types de données (données quantitatives, qualitatives, données temporelles, spatiales, etc.). Elles regroupent des techniques plus ou moins anciennes, analyses statistiques, classification, arbres de décision, et des techniques plus récentes, telles que la recherche de règles d'association (Agrawal, Srikant, 1994) ou de motifs temporels (Agrawal, Srikant, 1995), qui permettent de traiter de gros volumes de données. Ces méthodes ont été utilisées dans de nombreux domaines d'application, en lien ou non avec un système décisionnel.

Dans le contexte des systèmes d'information décisionnels environnementaux, les fonctionnalités de chaque niveau sont enrichies et adaptées aux particularités des données environnementales (par exemple, composantes spatiale et temporelle), et aux analyses associées (par exemple, agrégations temporelles, interpolations spatiales). C'est un tel système que nous avons proposé de mettre en place dans le projet Fresqueau, pour traiter les deux questions générales posées dans le cadre de la DCE. Ce système contient une base de données intégrée complétée d'un entrepôt ainsi qu'un ensemble d'outils permettant l'exploration, la visualisation et l'analyse des données ainsi rendues disponibles (cf. figure 1). 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). L'expérience présentée ici est plus ambitieuse en termes de volume et de variété des données et informations considérées et d'analyses produites. Elle porte en effet sur deux grands bassins, correspondant aux districts Rhin-Meuse et Rhône-Méditerranée et Corse, pour la période 2002-2010.

De plus, cinq catégories de données sont considérées, au lieu des seules deux premières dans le précédent projet : (i) les données relatives à la qualité de l'eau, bio-indicateurs 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, apportant les informations issues des différents réseaux ; (iii) les données

décrivant le réseau hydrographique, afin de comparer ou compléter les informations sur les stations de mesures et leurs 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. Pour exploiter ces données, un système OLAP et des outils de fouille ont été mis en place. Le premier permet à l'utilisateur d'effectuer des analyses multidimensionnelles sur les données physico-chimiques et biologiques. Les seconds ont pour objectifs l'analyse des séquences temporelles de prélèvements physico-chimiques, l'analyse des données relationnelles portant à la fois sur la physico-chimie et les peuplements des stations de mesures, ou encore l'analyse des occupations du sol au voisinage des stations de mesures.

Le plan de l'article est le suivant. Dans la section 2, les questionnements des hydro-écologues sont présentés, puis nous détaillons les différentes catégories de données collectées et les problématiques posées par ces données. La section 3 présente les étapes de modélisation et d'intégration des données. La section 4 décrit l'entrepôt mis en place, qui comprend deux cubes décrivant les données selon plusieurs dimensions temporelles, spatiales et thématiques ; quelques exemples des résultats obtenus par exploration de ces cubes sont également présentés. La section 5 présente une des méthodes de fouille développées et le type de résultats obtenus. Dans la section 6 notre travail est situé et discuté par rapport aux travaux voisins. Finalement nous dressons quelques conclusions et perspectives.

## 2. Questionnements thématiques et sources de données

Le projet Fresqueau a pour objectif principal la mise en place d'un système décisionnel pour permettre l'analyse de données concernant les masses d'eau, afin de répondre à certaines questions des hydro-écologues. Ces questions relèvent de deux grands sujets, cités en introduction, (1) l'étude des différentes métriques permettant de caractériser la qualité des cours d'eau et (2) l'étude des relations entre sources de pressions sur le milieu et qualité des cours d'eau. De manière plus concrète, le système construit doit permettre de réaliser par exemple les tâches suivantes :

– appréhender les données dans leurs différentes dimensions, temporelles, spatiales et thématiques : par exemple, quelles sont les valeurs d'indices biologiques dans les différentes hydro-écorégions ? ou leurs évolutions sur certains cours d'eau ?

– relier les variations d'un paramètre physico-chimique avec la variation des populations de taxons (faune ou flore) présentant certaines caractéristiques : par exemple, le trait respiration est-il influencé par des variations des taux de matières organiques ?

– relier la composition chimique de l'eau et les notes d'indices biologiques relevées à l'aval : par exemple, l'état de la qualité en nutriments (nitrates, phosphates) influence-t-il les notes d'IBMR (indice biologique macrophytique en rivière (AFNOR, 2003b)) ou d'IBD (indice biologique diatomées (AFNOR, 2003a)) à l'aval ? sur quelle période de temps ces paramètres ont-ils une influence ?

Pour répondre à ces questions nous avons collecté des données à partir de différentes sources sur une zone du territoire français qui couvre les districts de l'agence de l'eau Rhin-Meuse (33 000 km$^2$, 7 000 km de cours d'eau) dans le nord-est de la France et de l'agence de l'eau Rhône-Méditerranée et Corse (130 000 km$^2$, 152 000 km de cours d'eau) dans le sud-est (cf. figure 2). Ci-dessous, 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 sources, qui dépendent souvent de plusieurs producteurs, dont les objectifs diffèrent dans le temps et dans l'espace.

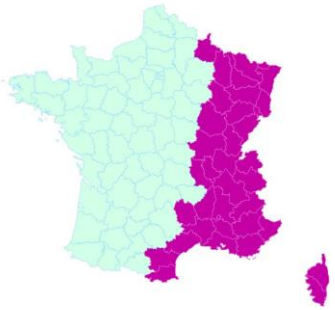

Figure 2. Localisation de la zone d'étude, couvrant l'est de la France (en foncé)

## 2.1. 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 naturelle et anthropique, et en un état chimique, qui se traduit par des paramètres caractérisant les micropolluants (pesticides...) toujours d'origines anthropiques ; ces paramètres peuvent être regroupés en macro-paramètres (matières organiques, matières azotées...) (MEDDE, 2012) ;

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 (IBGN (AFNOR, 2004a)) est le plus fréquemment utilisé ; les caractéristiques (traits de vie (Usseglio-Polatera et al., 2000)) 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 [3].

L'accès aux données est relativement aisé pour les notes des principaux 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 [4], ou sur le site de l'ONEMA. Les secondes, accessibles via les DREAL ou les agences de l'eau, sont de formats très hétérogènes et nécessitent presque toujours des prétraitements avant leur intégration. D'autres données dépendent de fournisseurs locaux (laboratoires de recherche, réseaux de surveillance spécifiques). Le tableau 1 décrit, pour certaines des 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.

### 2.2. Caractéristiques des stations de mesures

Les caractéristiques des stations de mesures comprennent les informations liées à leur dénomination, à leur localisation et à leurs objectifs. Ces informations peuvent être complétées par des informations synthétiques rendant compte du contexte hy-

2. DREAL : Direction régionale de l'environnement, de l'aménagement et du logement ; ONEMA : Office national de l'eau et des milieux aquatiques.

3. Service d'administration nationale des données et référentiels sur l'eau – voir http://www.sandre.eaufrance.fr

4. http://www.eaufrance.fr

Tableau 1. Exemples de modes d'accès et droits attachés aux données collectées

| | |
|---|---|
| **Données Agence de l'eau RM&C** | Données libres et gratuites sur demande<br>Conditions d'utilisations : pas d'utilisation commerciale ; http://www.rhone-mediterranee.eaufrance.fr/regles-droits.php#rmedregles<br>Mention : **Agence de l'Eau Rhône Méditerranée et Corse 2011** |
| **Stations RM** | Données libres et gratuites, téléchargeables via le portail Eau France pour les données physico-chimiques et les indices biologiques (de 1996 à 2007) ou sur demande pour les listes biologiques (à l'Agence RM pour les invertébrés, macrophytes, oligochètes, à la DREAL Lorraine pour les diatomées)<br>Conditions d'utilisations : pas d'utilisation commerciale, http://www.eau-rhin-meuse.fr/Mention : **Agence de l'Eau Rhin Meuse, 2012** |
| **BDD Image - ONEMA** | Droits d'usage : http://creativecommons.org/licenses/by-nc-sa/2.0/fr<br>Référence du document : "**IMAGE**"<br>Copyright : **"© ONEMA - 2012"**<br>Mention de copyright du coéditeur, le cas échéant<br>Dans l'hypothèse d'une utilisation qui aurait ou pourrait avoir une quelconque influence sur l'intégrité des données, l'utilisateur informe de manière visible en portant la mention «Origine des données : IMAGE - ONEMA – Données ayant fait l'objet de modifications par un tiers – La responsabilité de l'ONEMA et des producteurs de données ne peut être engagée». |
| **Données DREAL (ante-Naïades)** | Données libres et gratuites<br>Pas d'utilisation commerciale<br>Droits d'usage : http://creativecommons.org/licenses/by-nc-sa/2.0/fr<br>Mention **© SIE – ONEMA -2012** |

drologique 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, les échantillonnages se font sur des tronçons de rivières – dénommés points de 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.). Seul l'échantillonnage des diatomées, très localisé, peut être assimilé à un point. 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).

### 2.3. Réseau hydrographique

Concernant le réseau hydrographique, trois sources de données sont disponibles : la BD Topo®, la BD Carthage® et le réseau Syrah. La BD Topo® est la base de données vectorielles de référence produite par l'IGN (Institut national de l'information géographique et forestière). 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.

### 2.4. Activités humaines

Les activités humaines se traduisent par des prélèvements d'eau, des rejets dans le milieu naturel, et des modifications physiques des milieux, qui s'exercent comme des pressions positives ou négatives 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 issus 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 la photo-interprétation de données satellitaires ;
2. une sélection de la BD Topo® (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®.

### 2.5. 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.

## 3. Modélisation et intégration des données

Les données collectées sont caractérisées par une grande hétérogénéité en raison de leur origine (valeurs de mesures ou d'expertise) et 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. Nous présentons ci-dessous le modèle développé ainsi que quelques problématiques d'intégration rencontrées.

### 3.1. 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 l'entité concernant les stations de mesures [5] (cf. figure 3). 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 4, en utilisant la notation Merise du formalisme entité-association. Dans cette figure, nous utilisons les pictogrammes spatiaux de PictograF [6] pour représenter les

5. Outre son nom et ses coordonnées géographiques, la station est caractérisée par sa localisation sur le cours d'eau (point kilométrique) et un point caractéristique qui la symbolise sur la carte.

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

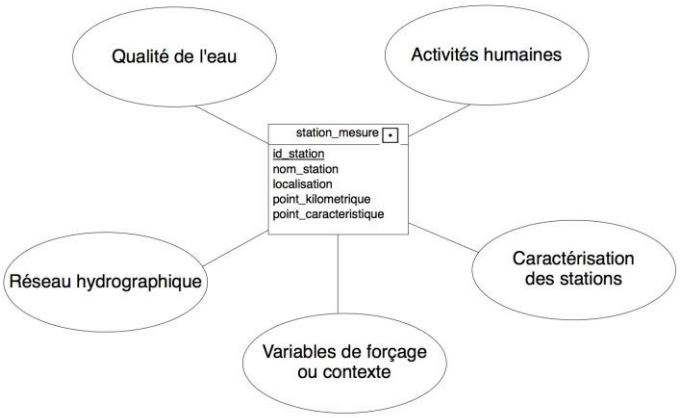


Figure 3. Modèle de données global

caractéristiques spatiales des objets ayant une géométrie. La station de mesures 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. L'association en tiretés ne fait pas partie du modèle initial, construit sur la base des données 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 des analyses. On remarque sur ce modèle plusieurs entités tronçon_* qui correspondent chacune à une des sources représentant cette information (BD Topo®, BD Carthage® et réseau Syrah).

Le modèle comporte également des entités 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).

### 3.2. Intégration

Les différentes sources de données requises ont dû être unifiées, structurées et intégrées dans la base de données Fresqueau. Pour cela, nous avons mis en place un

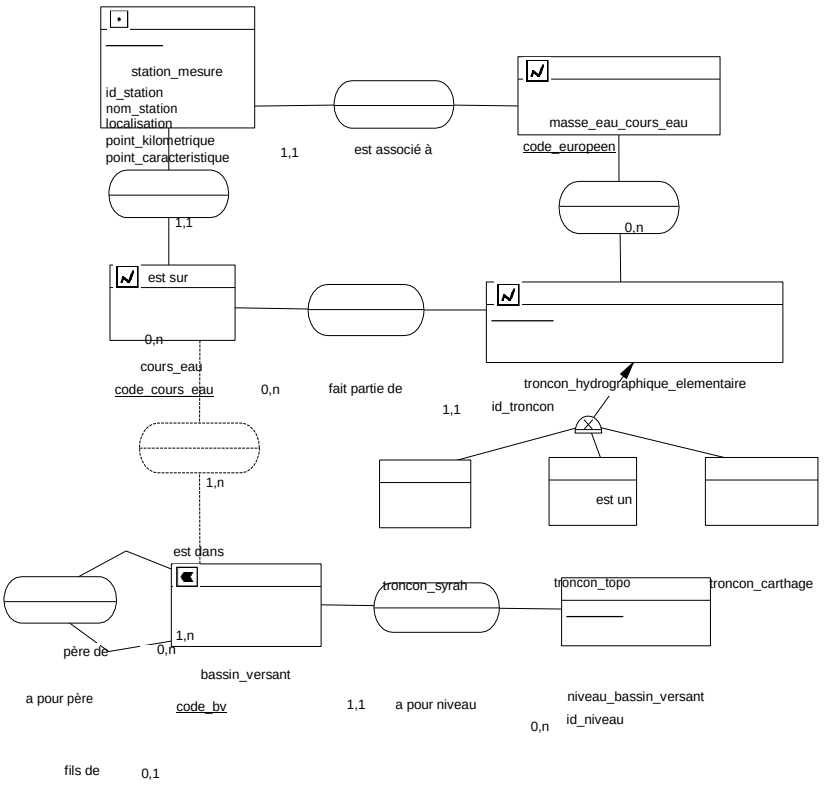


Figure 4. Modèle de données concernant les réseaux hydrographiques

processus ETL (Extract-Transform-Load) appuyé sur l'outil Spatial Data Integrator [7]. Ce processus a été rendu complexe du fait des caractères hétérogène et incomplet des données, mais aussi du fait de leur caractère spatio-temporel. En effet, comme indiqué précédemment, le concept de station de mesures est central dans notre modèle, et l'intégration des données correspondantes repose sur plusieurs réseaux de mesures, 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 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, environ 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

7. http://www.spatialdataintegrator.com

points kilométriques, les codes Cgenelin (code hydrographique du cours d'eau) et un voisinage de 10 à 50 mètres. Nous avons alors pu relier plus de 90 % d'entre elles sur les deux districts étudiés. 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 (Grac et al., 2011) 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, à son tour source potentielle d'erreurs.

### 3.3. Etat de la base

Le modèle a été implanté en utilisant PostgreSQL/PostGIS. Les données collectées sont stockées dans 81 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 qui permettent de faire le lien entre des objets spatiaux (par exemple, un cours d'eau et son bassin versant) ou apportent des informations agrégées utiles à l'analyse, comme des données sur les pressions présentes à proximité d'une station de mesures : par exemple, la table qui a pour schéma `station_300m_clc(id_station, code_clc_3, surface)` donne la répartition des occupations du sol (extraite de Corine Land Cover niveau 3) dans un cercle de 300 mètres autour d'une station. Vingt-cinq nouvelles tables, capturant l'information sur les pressions (bâti, réseau routier, bois, cultures, etc.) situées 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.

## 4. Un entrepôt de données

Dans cette section, nous présentons deux cubes de données, l'un concernant les données physico-chimiques, l'autre les données biologiques. Dans un premier temps, nous présentons leurs modèles conceptuels en utilisant le profil UML (Unified Modeling Language) défini dans (Boulil et al., 2015). Nous donnons des exemples des requêtes réalisables, puis nous décrivons brièvement l'implantation mise en œuvre.

### 4.1. Modèle conceptuel pour le cube des données physico-chimiques

Le cube présenté ici (cf. figure 5) permet de réaliser des analyses OLAP sur les données concernant les mesures physico-chimiques. Dans ce modèle, les mesures sont exprimées dans différentes unités de mesure (µg/L, cm/min, $cm^3$, $g/m^2$, etc.). Chaque mesure représente la valeur d'un paramètre physico-chimique qui peut être analysée selon sept dimensions, détaillées ci-dessous.

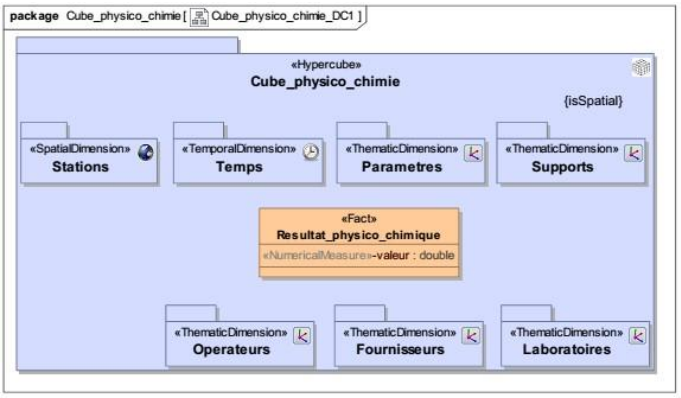


Figure 5. Modèle multidimensionel du cube « physico-chimie »

1. Dimension Paramètres : c'est une dimension thématique qui représente les informations sur les paramètres physico-chimiques de qualité de l'eau, organisée en hiérarchie : les paramètres physico-chimiques (par exemple, glyphosate) sont regroupés en sous-catégories (par exemple, pesticides), et les sous-catégories en catégories (par exemple, micropolluants) ; ceci permet d'agréger les mesures par paramètres, sous-catégories ou catégories.

2. Dimension Stations : cette dimension spatiale contient les données caractérisant les stations de mesures (cf. section 2). Chaque station est représentée par un point dans l'espace. Cette dimension permet de calculer la distribution spatiale des mesures de qualité de l'eau. Elle est organisée selon plusieurs hiérarchies spatiales, qui permettent de réaliser différentes agrégations des mesures physico-chimiques à différentes échelles et unités spatiales :

- la hiérarchie administrative « Station < Municipalité < Département » regroupe les stations selon les découpages administratifs français ;

- la hiérarchie hydrographique « Station < Masse d'eau < Hydro-écorégion_1 » regroupe les stations selon les masses d'eau puis les hydro-écorégions de niveau 1 ;

- la hiérarchie « Station < Masse d'eau < Fr_Type » regroupe les stations selon les masses d'eau puis les types français de masse d'eau ;

- la hiérarchie « Station < Masse d'eau < Modification_Type » regroupe les stations selon les masses d'eau puis selon les types de modification (état naturel, artificiel, très modifié, etc.). Cette hiérarchie modélise ainsi des informations concernant les pressions physiques sur les cours d'eau (barrage, canalisation, etc.) ;

- la hiérarchie « Station < Cours d'eau < Rang Cours d'eau » regroupe les stations selon les cours d'eau et les rangs de cours d'eau (correspondant à l'inverse du rang de Strahler (Strahler, 1957)) ;

- la hiérarchie « Station < Cours d'eau < Bassin versant_3 < Bassin versant_2 < Bassin versant_1 » regroupe les stations selon les cours d'eau, puis les bassins versants de niveau 3 (par exemple, le bassin versant du Sânon, petit cours d'eau), les bassins versants de niveau 2 (par exemple, le bassin versant de la Meurthe, grand cours d'eau), et enfin selon les bassins versants de niveau 1 (par exemple, le bassin versant du Rhin).

3. Dimension Temps : cette dimension contient les dates des prélèvements effectués sur les stations. Elle definit deux hiérarchies: « Jour < Mois < Bimestre < Semestre < Année » et « Jour < Mois < Trimestre < Semestre < Année ». Elle permet de réaliser des agrégations temporelles sur les données physico-chimiques.

4. Dimension Supports : c'est une dimension thématique qui décrit les milieux dans lesquels les paramètres physico-chimiques ont été mesurés (eau, sédiments, etc.). Cette information est organisée dans une hiérarchie à deux niveaux : les fractions analysées (par exemple, eau brute) regroupées en supports (par exemple, eau). Des résultats obtenus dans différents supports ne peuvent pas être agrégés.

5. Dimension Opérateurs : cette dimension thématique décrit les personnes qui ont réalisé les prélèvements sur les stations.

6. Dimension Laboratoires : cette dimension thématique décrit les laboratoires qui ont réalisé les analyses. Il est intéressant de distinguer les laboratoires car ils utilisent différentes méthodes qui peuvent rendre leurs résultats spécifiques.

7. Dimension Fournisseurs : cette dimension thématique décrit les organismes (Agence de l'eau, par exemple) qui ont fourni les données.

Comme énoncé ci-avant, les cubes permettent de calculer des indicateurs servant à l'analyse en agrégeant les mesures le long des hiérarchies, selon différentes fonctions d'agrégation. Pour ce cube, nous avons défini plusieurs indicateurs en appliquant différentes fonctions d'agrégation à la mesure val_paramètre. Par exemple, l'indicateur Moy_val_paramètre est calculé en appliquant la fonction moyenne (Moy) sur toutes les dimensions. L'indicateur Compt_val_paramètre quant à lui renvoie le nombre de valeurs de paramètres pour une combinaison de dimensions (par exemple, une période de temps, un sous-ensemble de paramètres physico-chimiques et un sous-ensemble de zones géographiques). Nous avons également introduit de nouvelles fonctions d'agrégation pour calculer des indicateurs plus complexes, tels que les percentiles, qui sont couramment utilisés en hydro-écologie (Boulil et al., 2014).

Les résultats affichés en figure 6 concernent une requête portant sur le nombre de mesures par type de paramètres et par département (quatre départements sont figurés ici). On voit également sur cette figure les différentes hiérarchies définies pour le cube de données physico-chimiques.

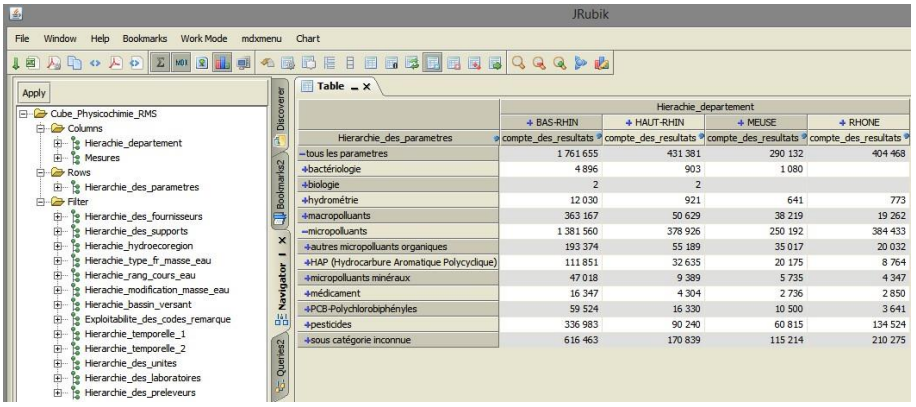


Figure 6. Visualisation du décompte des mesures physico-chimiques par type de paramètre et par département

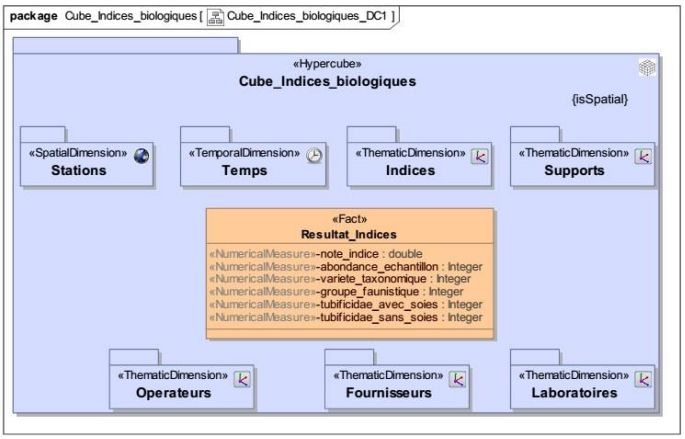


Figure 7. Modèle multidimensionel du cube « hydrobiologie »

## 4.2. Modèle conceptuel pour le cube des données hydrobiologiques

Ce cube a été défini pour l'analyse des prélèvements hydrobiologiques (cf. figure 7). Il est identique au cube physico-chimique pour six dimensions. Il possède en plus la dimension thématique **Indices** qui regroupe les indices biologiques (IBG, IBGA, IBGN, etc.) selon leur thème taxonomique (dans l'exemple, indices invertébrés). Outre la mesure principale, la note d'indice biologique, il y a ainsi cinq autres mesures biologiques de qualité d'eau, qui dépendent du thème taxonomique :

– l'abondance, c'est-à-dire le nombre total d'individus présents dans l'échantillon (par exemple, le nombre de poissons, échantillonnés toutes expèces confondues, pour l'indice poissons en rivières) ;

– la variété taxonomique, c'est-à-dire le nombre d'espèces différentes de taxons trouvés dans un échantillon ;

– le groupe faunistique pour les invertébrés, une valeur entre 1 et 9, qui correspond au groupe le plus polluo-sensible trouvé dans un échantillon de macroinvertébrés (par exemple, 9 si les familles les plus polluo-sensibles, Chloroperlidae ou Perlidae, ou Perlodidae ou Taeniopterygidae, ont été trouvées) ;

– les Tubificidae avec soies pour les indices oligochètes, c'est-à-dire le nombre d'individus Tubificidae avec soies trouvés dans un échantillon d'oligochètes [8] ;

– les Tubificidae sans soies pour les indices oligochètes, c'est-à-dire le nombre d'individus Tubificidae sans soies trouvés dans un échantillon d'oligochètes.

8. Les Tubificidae sont la famille la plus polluo-résistante parmi les oligochètes.

| | Indices_par_theme_taxons | | | | | |
|---|---|---|---|---|---|---|
| | tous les indices | diatomées | invertébrés | macrophytes | oligochètes | poissons |
| Hierachie_hydroecoregion | Compte_des_notes | Compte_des_notes | Compte_des_notes | Compte_des_notes | Compte_des_notes | Compte_des_notes |
| toutes les hydroecoregions | 34 241 | 8 493 | 16 273 | 1 095 | 48 | 8 332 |
| ALPES INTERNES | 2 205 | 642 | 1 412 | | | 151 |
| ALSACE | 1 927 | 1 065 | 708 | | | 154 |
| ARDENNES | 326 | 169 | 97 | | | 60 |
| CEVENNES | 1 249 | 316 | 834 | | | 99 |
| CORSE | 886 | 269 | 571 | | | 46 |
| COTEAUX AQUITAINS | 294 | 73 | 206 | | | 15 |
| COTES CALCAIRES EST | 3 531 | 1 345 | 1 640 | | | 546 |
| HER 1 inconnue | 7 553 | 157 | 552 | 1 095 | 48 | 5 701 |
| JURA-PREALPES DU NORD | 5 535 | 1 335 | 3 736 | | | 464 |
| MASSIF CENTRAL NORD | | | | | | |
| MASSIF CENTRAL SUD | 861 | 194 | 572 | | | 95 |
| MEDITERRANEEN | 4 975 | 1 430 | 3 157 | | | 388 |
| PLAINE SAONE | 1 950 | 541 | 1 235 | | | 174 |
| PREALPES DU SUD | 1 284 | 362 | 787 | | | 135 |
| PYRENEES | 387 | 92 | 212 | | | 83 |
| VOSGES | 1 278 | 503 | 554 | | | 221 |

Figure 8. Visualisation des décomptes des notes d'indices biologiques par thème taxonomique et par hydro-écorégion

Pour permettre l'analyse OLAP de ces mesures biologiques, nous avons défini de nombreux indicateurs en utilisant les différentes fonctions d'agrégation existantes (Moy, Min, Max, etc.). Nous donnons ici quelques exemples d'indicateurs : l'indicateur Min_note_indice est calculé en appliquant la fonction minimum (Min) à la mesure note_indice selon une dimension à préciser ; l'indicateur Compt_note_indice donne le nombre de notes d'indices pour une combinaison de dimensions (par exemple, une période de temps et un sous-ensemble d'indices).

Une requête OLAP sur ce cube peut porter par exemple sur le décompte des valeurs note_indice par thème taxonomique (regroupant les indices biologiques d'un même thème) et par hydro-écorégion de niveau 1 (cf. figure 8). On remarque sur la figure que les indices des thèmes diatomées et invertébrés sont plus représentés que les autres indices biologiques, c'est en effet les indices les plus souvent mesurés.

## 4.3. Implantation

Nous avons implanté les cubes dans une architecture OLAP relationnelle (ROLAP) utilisant uniquement des outils logiciels libres, et en trois parties distinctes.

La partie ETL est en charge du peuplement des cubes de données, à partir de la base intégrée Fresqueau et des fichiers annexes. Cette partie est constituée d'un ensemble de programmes JAVA pour l'extraction et la transformation de données, mis en œuvre au moyen des outils Spatial Data Integrator (SDI) et Talend Open Studio (TOS) [9]. L'outil SDI est utilisé pour l'intégration de données spatiales. Les entrepôts

9. http://www.talend.com

de données ont été implantés en utilisant PostGIS. Le serveur OLAP utilisé est Mondrian [10] et le client OLAP est JRubik [11].

## 5. Analyse par des méthodes de fouille

Nous avons exploité les données en utilisant différentes méthodes de fouille de données : données relationnelles à l'aide de l'analyse formelle de concepts (Dolques et al., 2013), données temporelles à l'aide d'une méthode de fouille de séquences qualitatives (Fabrègue et al., 2013), données spatiales à l'aide de méthodes issues des statistiques spatiales (Lalande, 2013). Nous décrivons ici pour exemple des résultats obtenus avec la méthode décrite dans (Fabrègue et al., 2013), qui recherche des motifs partiellement ordonnés fermés (CPO-motifs) dans une base de séquences temporelles. L'idée générale est d'extraire à partir d'observations temporelles (ici des mesures effectuées sur les stations de rivières) des répétitions fréquentes de valeurs qui sont ensuite synthétisées sous la forme de CPO-motifs. Un exemple illustratif est présenté sur la figure 9 : ce motif signifie que fréquemment, dans une base de séquences, il est observé qu'un bas taux d'oxygène et qu'un niveau élevé de pesticides (ces deux observations n'étant pas ordonnées entre elles) sont temporellement suivis par une dégradation du compartiment biologique.

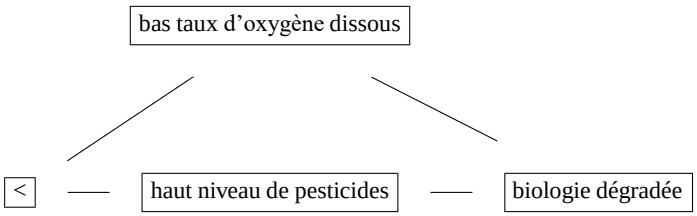


Figure 9. Exemple de CPO-motif (le signe < indique le début du motif)

Les CPO-motifs sont bien adaptés au caractère temporel des données et leur représentation sous la forme de graphes acycliques est facilement lisible par des experts. De plus, le fait que les motifs soient fermés permet de résumer au mieux les informations contenues dans une base de séquences. Malgré cela, la méthode produit trop de résultats et, pour les réduire, nous utilisons des mesures d'intérêt (Geng, Hamilton, 2006) qui permettent alors de mettre en évidence quelques motifs fréquents, discriminants et non redondants, parmi les milliers qui peuvent être extraits de la base Fresqueau.

Nous présentons ci-dessous un exemple des données traitées pour mettre en évidence le lien entre la physico-chimie et les indices biologiques. Ce traitement consiste

10. http://community.pentaho.com/projects/mondrian/
11. http://rubik.sourceforge.net

à utiliser les données physico-chimiques qui précèdent une mesure de la biologie dans un certain intervalle de temps. Dans l'exemple présenté ci-dessous, l'intervalle est de six mois. Toutes les données (biologiques et physico-chimiques) ont été discrétisées en s'appuyant sur les connaissances du domaine : les valeurs numériques des paramètres ont été transformées en cinq classes de qualité, « Très bon », « Bon », « Moyen », « Mauvais » et « Très mauvais », représentées par cinq couleurs Bleu, Vert, Jaune, Orange et Rouge. La discrétisation des paramètres biologiques et physico-chimiques est déterminée par des normes différentes. La biologie est discrétisée comme préconisé dans les normes AFNOR des indices biologiques. Par exemple, le tableau 3 donne les intervalles de discrétisation pour l'indice biologique IBGN selon sa norme (AFNOR, 2004a). La physico-chimie est discrétisée selon la norme SEQ-eau (MEDDE, 2003) qui regroupe les paramètres physico-chimiques en macro-paramètres. Par exemple, le macro-paramètre AZOT (matières azotées hors nitrate) regroupe les paramètres $NH_4^+$ (ammonium), NJK (azote Kjeldahl) et $NO_2^-$ (nitrites). Le macro-paramètre PHOS (matières phosphorées) regroupe les paramètres $PO_4^{3-}$ (phosphate) et phosphore total. La classe de chaque macro-paramètre est la pire des classes de ses constituants [12].

Tableau 3. Définition des classes de qualité de l'IBGN selon les intervalles de notes (AFNOR, 2004a)

| Indice | Bleu | Vert | Jaune | Orange | Rouge |
|---|---|---|---|---|---|
| IBGN | [20,17] | ]17,13] | ]13,9] | ]9,5] | ]5,0] |

Après discrétisation, les séquences à fouiller sont construites de la façon suivante. Pour chaque valeur d'indice (par exemple $IBGN^B$), mesurée à une date $t$, on considère les valeurs physico-chimiques mesurées sur le même site dans une période de 6 mois avant la date $t$. En général, on trouve dans cette période 2 ou 3 mesures effectuées à des dates différentes, $t_1 < t_2 < \ldots < t$. Les valeurs mesurées à la même date sont assemblées dans un itemset, par exemple ($AZOT^B$) pour la date $t_1$ et ($AZOT^V$, $PHOS^B$) pour la date $t_2$ ; les itemsets sont ensuite ordonnés dans une séquence selon l'ordre temporel, par exemple $\langle(AZOT^B)(AZOT^V, PHOS^B)\rangle$. Chaque séquence de valeurs de macro-paramètres est affectée à une sous-base de qualité biologique en accord avec la valeur de l'indice biologique associé. Ainsi la séquence $\langle(AZOT^B)(AZOT^V, PHOS^B)\rangle$ fait partie de la sous-base associée à la valeur $IBGN^B$ (voir tableau 4).

Les motifs ont été extraits puis filtrés selon une combinaison de critères d'intérêt adaptés aux besoins des hydro-écologues. Nous présentons ici quelques exemples de motifs obtenus pour les indices IBGN (macroinvertébrés) et IPR (poissons (AFNOR, 2004b)). La figure 10 présente deux CPO-motifs extraits des sous-bases associées aux états « Très bon » ($IBGN^B$) et « Très mauvais » ($IBGN^R$) de l'IBGN. Les deux macro-paramètres MINE et MOOX représentent respectivement la minéralisation de l'eau et les matières organiques et oxydables présentes dans l'eau. Les deux motifs signifient respectivement que l'on a mesuré des très bonnes (très mauvaises) valeurs de ces deux paramètres simultanément et jusqu'à six mois avant un relevé IBGN bleu (rouge).

12. http://sierm.eaurmc.fr/eaux-superficielles/fichiers-telechargeables/grilles-seq-eau-v2.pdf

Tableau 4. Sous-bases de séquences associées à trois valeurs de l'indice IBGN – les notations $^{B}$, $^{V}$ ... correspondent respectivement aux valeurs Bleu, Vert, etc.

| Sous-bases | Séquences |
|---|---|
| IBGN$^{B}$ | $\mathfrak{h}$(AZOT$^{B}$)(AZOT$^{V}$,PHOS$^{B}$)$\mathfrak{i}$ |
| | $\mathfrak{h}$(AZOT$^{B}$,PHOS$^{V}$)(PHOS$^{V}$)(AZOT$^{J}$,PHOS$^{B}$)$\mathfrak{i}$ |
| IBGN$^{J}$ | $\mathfrak{h}$(AZOT$^{V}$,PHOS$^{V}$)(AZOT$^{B}$,PHOS$^{V}$)$\mathfrak{i}$ |
| | $\mathfrak{h}$(PHOS$^{O}$)(AZOT$^{O}$,PHOS$^{J}$)(AZOT$^{V}$,PHOS$^{J}$)$\mathfrak{i}$ |
| | $\mathfrak{h}$(AZOT$^{V}$,PHOS$^{J}$)(AZOT$^{V}$,PHOS$^{B}$)(AZOT$^{V}$)$\mathfrak{i}$ |
| IBGN$^{R}$ | $\mathfrak{h}$(AZOT$^{O}$,PHOS$^{J}$)(AZOT$^{R}$,PHOS$^{O}$)(AZOT$^{V}$,PHOS$^{J}$)$\mathfrak{i}$ |
| | $\mathfrak{h}$(PHOS$^{O}$)(AZOT$^{O}$,PHOS$^{J}$)(AZOT$^{V}$)$\mathfrak{i}$ |

On remarque ici que les classes des macro-paramètres correspondent exactement aux classes de l'indice biologique.

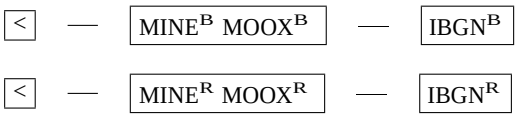


Figure 10. Deux CPO-motifs extraits des sous-bases IBGN$^{B}$ et IBGN$^{R}$ associant les mêmes macro-paramètres avec des classes différentes

La figure 11 présente un CPO-motif caractéristique de la sous-base IPR$^{R}$ mais qu'on trouve aussi fréquemment dans la base IPR$^{O}$. On observe ici une séquence de trois relevés du macro-paramètre PAES (matières en suspension) de bonne qualité associée à un relevé du macro-paramètre PHOS (matières phosphorées) de mauvaise qualité. Ce motif n'est pas forcément explicatif du très mauvais état mesuré par l'IPR, car les poissons sont aussi sensibles aux paramètres physiques du milieu. Il faut donc approfondir l'étude en intégrant d'autres paramètres présents dans la base Fresqueau (état des berges, vitesse du courant, forme du lit, etc.).

Nous n'avons présenté ici que quelques résultats. Le lecteur intéressé trouvera plus de résultats dans l'article (Fabrègue et al., 2014) ; des listes exhaustives des motifs trouvés pour chaque classe d'indice biologique sont disponibles en ligne [13].

## 6. 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

13. http://engees-fresqueau.unistra.fr/patterns/patterns.html

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 en section 2, 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 » et d'un outil d'interrogation de cette banque, le SEEE (Système d'évaluation de l'état de l'eau), capable de fournir des évaluations des masses d'eau consultées. Naïades rassemblera toutes les données non encore présentes dans les bases des agences de l'eau : données de qualité hydromorphologiques, biologiques (par exemple, listes floristiques et faunistiques, ensemble des indices biologiques et métriques de calculs intermédiaires), données spécifiques des stations de mesures biologiques. Cette banque interrogera à distance les bases de données des agences de l'eau pour les données de qualité physico-chimiques et les caractéristiques générales des stations de mesures. Comme actuellement dans les bases de données des agence, Naïades ne prendra pas en compte les données concernant les réseaux hydrographiques, les activités humaines ou les variables de forçage ou de contexte.

On citera également la base de données constituée dans le cadre du PIREN Seine (Programme interdisciplinaire de recherche sur l'environnement de la Seine) [14]. 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 17000 km$^2$. Si la base est générale,

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

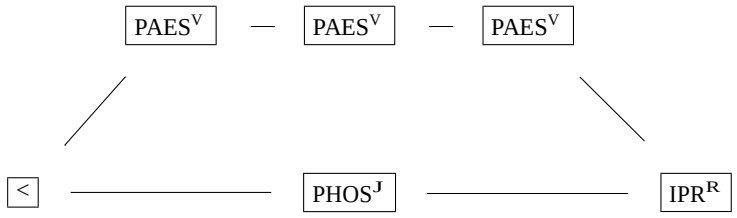


Figure 11. Un CPO-motif de la sous-base IPR$^{R}$, associant les macro-paramètres « matières phosphorées » et « matières en suspension »

les informations collectées sont limitées au cours d'eau étudié et aux exigences du modèle utilisé, 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. Plusieurs études ont également montré l'intérêt d'utiliser des entrepôts de données et les technologies d'analyse associées pour stocker et analyser des données environnementales (Alexandru et al., 2010 ; McGuire et al., 2008 ; Pinet, Schneider, 2010). Par exemple, dans (Alexandru et al., 2010) il est question d'un entrepôt pour l'analyse de données sur les polluants industriels et les risques environnementaux liés. Dans (Le Gal et al., 2002) il est question de la mise en place d'un SI dédié à la maintenance des réseaux hydrauliques au Niger. Dans (Mimouni et al., 2007) est décrit un système d'information géographique permettant de collecter et visualiser des données spatiales concernant la géologie et l'environnement au Maroc. Plus récemment, les travaux décrits dans (Vernier et al., 2013) portent sur un système décisionnel, incluant un entrepôt de données et permettant de caractériser les activités agricoles dans des bassins versants, en relation avec des indicateurs de présence de pesticides.

Ces différents 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. À 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 et de décision, grâce aux analyses permises par les cubes et les méthodes de fouille de données développées, pour répondre à l'ensemble des questions posées par les hydrologues et hydro-écologues.

## 7. Conclusion

Nous avons décrit dans cet article un système d'information pour l'étude et l'évaluation de l'état des écosystèmes aquatiques. Le développement de ce système 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 des métadonnées, des dictionnaires des données et d'un journal de bord et doivent également les partager. 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 des hydro-écologues. A chaque question est associé un sous-jeu de données, extrait de la base intégrée.

La diversité des données collectées et de leurs formats, associée à la complexité des accès, nous ont conduits à 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 avons mises en œuvre nécessitent de disposer des données ensemble pour effectuer les prétraitements et sélectionner des sous-jeux adaptés aux questions posées.

Pour permettre l'analyse des données selon leurs différentes dimensions, nous avons développé un système ROLAP à base d'outils libres et composé de deux cubes de données : (1) un cube pour l'analyse OLAP des données physico-chimiques et (2) un cube pour l'analyse OLAP des données hydrobiologiques. Des indicateurs spécifiques au domaine, impliquant des agrégations complexes, ont été implantés pour permettre aux utilisateurs de disposer des informations nécessaires à leur analyse. De plus les problématiques inhérentes aux données hétérogènes ont été prises en compte (Boulil et al., 2014). Parallèlement, nous avons exploré les données au moyen de différentes techniques de fouille de données, avec des résultats nouveaux et intéressants selon les hydro-écologues impliqués dans le projet. Pour permettre aux utilisateurs finaux d'appliquer ces méthodes sur les jeux de données de leur choix et d'analyser facilement leurs résultats, nous développons une interface de visualisation (Accorsi et al., 2014) connectée à la base de données Fresqueau.

En perspectives, nous envisageons une mise en relation des cubes et des méthodes de fouille, pour faciliter la sélection des données à fouiller. D'autres cubes sont également à construire pour couvrir les différents thèmes (réseau hydrographique, activités humaines ...). De plus, les données étudiées ne concernent que 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, la procédure d'intégration, les cubes mis en place et aussi de tester à plus grande échelle les méthodes de fouille développées.

Remerciements

Ce travail a été financé dans le cadre du projet ANR 11 MONU 14 Fresqueau. Nous remercions les différentes personnes impliquées dans le projet et les organismes fournisseurs des données.

## Bibliographie

Abelló A., Samos J., Saltor F. (2006). A multidimensional conceptual model extending UML. Information Systems, vol. 31, n° 6, p. 541–567.

Accorsi P., Fabrègue M., Sallaberry A., Cernesson F., Lalande N., Braud A. et al. (2014). HydroQual: Visual Analysis of River Water Quality. In IEEE VIS 2014 Conference, Paris.

Agrawal R., Srikant R. (1994). Fast algorithms for mining association rules in large databases. In International conference on very large data bases (vldb'94), p. 487-499.

Agrawal R., Srikant R. (1995). Mining sequential patterns. In International conference on data engineering (icde'95), p. 3-14.

Alexandru A., Gorghiu G., Nicolescu C. L., Alexandru C.-A. (2010). Using OLAP Systems to Manage Environmental Risks in Dambovita County. Bulletin UASVM Horticulture.

Association Française de Normalisation. (2003a). Qualité de l'eau : détermination de l'Indice Biologique Diatomées (IBD). NF T90-354.

Association Française de Normalisation. (2003b). Qualité de l'eau : détermination de l'Indice Biologique Macrophytique en Rivière (IBMR). NF T90-395.

Association Française de Normalisation. (2004a). Qualité de l'eau : détermination de l'Indice Biologique Global Normalisé (IBGN). NF T90-350.

Association Française de Normalisation. (2004b). Qualité de l'eau : détermination de l'Indice Poissons Rivière (IPR). NF T90-344.

Boulil K., Bimonte S., Pinet F. (2015). Conceptual Model for Spatial Data Cubes: A UML Profile and its Automatic Implementation. Computer Standards & Interfaces, vol. 38, p. 113–132.

Boulil K., Le Ber F., Bimonte S., Grac C., Cernesson F. (2014). Multidimensional modeling and analysis of large and complex watercourse data: an OLAP-based solution. Ecological Informatics, vol. 24, p. 90–106.

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 EGC'2013: 13e Conférence Francophone sur l'Extraction et la Gestion des Connaissances, Toulouse, p. 121-132. Hermann-Éditions.

Fabrègue M., Braud A., Bringay S., Grac C., Le Ber F., Levet D. et al. (2014). Discriminant temporal patterns for linking physico-chemistry and biology in hydro-ecosystem assessment. Ecological Informatics, vol. 24, p. 210–221.

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, IDA 2013, London, vol. LNCS 8207, p. 186–197. Springer.

Fayyad U., Piatetsky-Shapiro G., Smyth P., Uthurusamy R. (1996). Advances in knowledge discovery and data mining. AAAI Press/The MIT Press.

Geng L., Hamilton H. J. (2006). Interestingness measures for data mining: A survey. ACM Computing Survey, vol. 38, n° 3.

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.

Inmon W. (2005). Building the data warehouse. John Wiley & Sons.

Lalande N. (2013). Impacts multi-échelles de l'occupation du sol sur l'état écologique des cours d'eau: élaboration et test d'un cadre d'analyse et de modélisation. Thèse de doctorat d'AgroParisTech. Montpellier.

Le Ber F., Teisseire M., Braud A., Cernesson F., Grac C., Poncelet P. (2014). Le projet Fresqueau : exploiter les données massives concernant les cours d'eau. Revue Ingénierie des Systèmes d'Information, vol. 19/3, p. 169–174.

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 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. Cirad – Cemagref – IRD.

Malinowski E., Zimányi E. (2008). Advanced Data Warehouse Design. Springer.

McGuire M., Gangopadhyay A., Komlodi A., Swan C. (2008). A user-centered design for a spatial data warehouse for data exploration in environmental research. Ecological Informatics, vol. 3, nº 4-5, p. 273–285.

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.

Ministère de l'Ecologie, du Développement Durable et de l'Energie. (2012). Guide technique : évaluation de l'état des eaux de surface continentales (cours d'eau, canaux, plans d'eau).

Ministère de l'Ecologie, du Développement Durable et de l'Energie et Agences de l'Eau. (2003). Système d'évaluation de la qualité de l'eau des cours d'eau (SEQ-Eau), version 2. Etude inter-agences de l'eau, n° 52.

Parlement européen et Conseil. (2000). Cadre pour une politique communautaire dans le domaine de l'eau. Directive 2000/60/EC.

Pinet F., Schneider M. (2010). Precise design of environmental data warehouses. Operational Research, vol. 10, nº 3, p. 349–369.

Ruelland D. (2004). SENEQUE, logiciel SIG de modélisation prospective de la qualité de l'eau. Revue Internationale 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, nº 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.

Strahler A. (1957). Quantitative analysis of watershed geomorphology. Transactions of the American Geophysical Union, vol. 38, p. 913–920.

Usseglio-Polatera P., Bournaud M., Richoux P., Tachet H. (2000). Biological and ecological traits of benthic freshwater macroinvertebrates. Hydrobiologia, vol. 516, p. 173-18.

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.