# Le projet Fresqueau : exploiter les données massives concernant les cours d'eau

**Florence Le Ber** [1], **Maguelonne Teisseire** [2], **Agnès Braud** [1], **Flavie Cernesson** [2], **Corinne Grac** [3], **Pascal Poncelet** [4]

1. *ICube, Université de Strasbourg, CNRS, ENGEES — Illkirch, France*
   *florence.leber@engees.unistra.fr,agnes.braud@unistra.fr*
2. *TETIS, AgroParisTech, IRSTEA — Montpellier, France*
   *{maguelonne.teisseire,flavie.cernesson}@teledetection.fr*
3. *LIVE, Université de Strasbourg/ENGEES, CNRS — Strasbourg, France*
   *corinne.grac@engees.unistra.fr*
4. *LIRMM, Université Montpellier 2, CNRS — Montpellier, France*
   *pascal.poncelet@lirmm.fr*

RÉSUMÉ. *Le projet ANR 11 MONU 14 Fresqeau est un projet interdisciplinaire associant informaticiens et hydroécologues. Il a pour but de collecter un ensemble de données concernant l'état des cours d'eau, à l'échelle de deux grands bassins versants français, puis d'exploiter ces données avec des méthodes de fouille afin de répondre à des questionnements scientifiques. Outre leur volume, les données ainsi collectées ont des origines et caractéristiques diverses : elles sont spatiales, temporelles, plus ou moins denses, quantitatives ou qualitatives, à précision et fiabilité variables. L'ensemble de ces caractéristiques rend délicate l'exploitation de ces données massives.*

ABSTRACT. *The ANR 11 MONU 14 Fresqueau project is an interdisciplinary project involving computer scientists and hydroecologists. Its aim is to collect a large dataset about watercourses, in two large french watersheds; and then to explore it with data mining methods for answering scientific questions. The collected data are numerous and have various origins and characteristics: they have spatial or temporal characteristics, they can be quantitative or qualitative, sparse or dense, their precision and validity are variable. Due to all these characteristics, mining this big dataset appeared a complex task.*

MOTS-CLÉS : *Base de données intégrée, fouille de données, données massives, hydrologie*

KEYWORDS: *Integrated Database, Data Mining, Big Data, Hydrology*

DOI:10.3166/ISI.22.1.9-?? © 2012 Lavoisier

## 1. Introduction

Le projet Fresqueau, qui a débuté en octobre 2011 pour trois ans, réunit un consortium de quatre laboratoires de recherche et deux bureaux d'études, avec des équipes d'informaticiens spécialisés en structuration et extraction de connaissances à partir de données et des équipes d'hydrologues et d'écologues spécialistes de l'évaluation des écosystèmes aquatiques. S'inscrivant dans les objectifs de la Directive Cadre Européenne sur l'eau (The European Parliament and the Council, 2000), qui 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, le consortium s'est donné 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 prend en charge 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 (occupation du sol, entre autres). Les données collectées sont caractérisées par une grande hétérogénéité en raison de leur origine (mesures effectuées localement, synthèses nationales ou bibliographies), des objectifs qui ont conduit à leur acquisition (suivi à long terme, référentiel, rapportage européen, études ponctuelles, etc.). De plus, 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).

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, concernent le développement d'un entrepôt 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.

## 2. Les données

Les données collectées proviennent principalement des agences de l'eau et de l'ONEMA (Office National de l'Eau et des Milieux Aquatiques), mais également d'autres sources tels que l'IGN (Institut Géographique et Forestier National) pour l'information géographique et différents services de l'Etat. Elles portent sur deux grands bassins hydrographiques, correspondant aux districts Rhin-Meuse (33.000 $km^2$) et Rhône-Méditerranée et Corse (130.000$km^2$), pour une période de temps allant de 1995 à 2010 ; cinq catégories de données sont considérées, nous en donnons quelques caractéristiques ci-dessous.

– Les données relatives à la qualité de l'eau, bioindicateurs et paramètres physico-chimiques, permettant de qualifier de façon détaillée la qualité des cours d'eau ; il y a plus de 4000 paramètres physico-chimiques et biologiques suivis à des échelles temporelles et spatiales variables et qui donnent lieu à des processus longs de vérification.

– Les données relatives aux stations de mesures, traduisant la complémentarité des informations issues des différents réseaux : ces réseaux sont très nombreux (plusieurs centaines en France), d'extensions temporelles et spatiales diverses.

– Les données décrivant le réseau hydrographique : BD TOPO®, BD Carthage® et réseau Syrah. Le rattachement des stations de mesures à ces réseaux nécessite des requêtes spatiales et attributaires.

– Les données relatives aux activités humaines pour estimer les pressions anthropiques ponctuelles et diffuses qui s'exercent sur les cours d'eau : registre parcellaire graphique et données du programme Corine Land Cover pour l'agriculture ; BD TOPO® pour le bâti.

– Les données relatives aux variables de forçage climatique ou de contexte afin de caractériser l'environnement des rivières et des points de prélèvements : données hydrologiques (débits), données climatiques, hydro-écorégions (régions homogènes pour les processus physiques dominants) mais aussi données administratives.

Après la mise au point d'un modèle de données, largement appuyé sur celui des bases sources, les données couvrant les deux districts ont été intégrées *via* l'ETL Talend (http://fr.talend.com) dans une base `PostgreSQL/PostGIS`. Cette base contient 80 tables, dont certaines ont un nombre de lignes important. 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 possèdent au moins un attribut représentant une géométrie.

## 3. Exploration des données

L'approche développée est fondalement conduite par les besoins des experts. Pour leur permettre d'explorer la masse de données disponibles selon différentes dimensions, thématiques, spatiales et temporelles, nous avons construit deux cubes OLAP, l'un s'appliquant aux relevés physico-chimiques, l'autre aux relevés biologiques. Par exemple, les valeurs des paramètres physico-chimiques mesurés sur une station peuvent être agrégées selon trois hiérarchies spatiales, celle des bassins versants emboîtés, celle des masses d'eau (une masse d'eau est une partie d'un cours d'eau et peut contenir des cours d'eau plus petits) et celle des découpages administratifs (commune, département, région) (Boulil *et al.*, 2014).

La fouille des données est également conduite par les questionnements des experts. Un ensemble de questions opérationnelles a été établi, chaque question étant spécifiée par une sous-base de données et des prétraitements à appliquer (explicitation de rela-

tions entre les objets spatiaux, calcul d'indicateurs, discrétisation des données). Pour chaque question des méthodes de fouille ont été développées ou adaptées aux volumes de données traitées : recherche de motifs dans des séquences temporelles (Fabrègue *et al*., 2013), apprentissage relationnel supervisé, analyse de concepts formels sur des tables relationnelles (Dolques *et al*., 2013), statistiques spatiales (Lalande, 2013). Des combinaisons de ces méthodes sont également étudiées. Les résultats sont en cours d'interprétation avec les hydroécologues impliqués dans le projet.

## 4. Conclusion et perspectives

L'acquisition et l'intégration des données concernant l'état des cours d'eau sur les deux districts considérés a été un long travail, qui nécessite encore de nombreuses vérifications. De plus, chaque question opérationnelle, et chaque méthode pouvant y répondre, induisent des choix de prétraitement relativement complexes, liés à l'hétérogénité et au volume des données à considérer. Les possibilités d'analyses sont donc encore largement ouvertes pour exploiter complètement cette masse de données.

La dernière étape du projet portera sur le développement d'un outil opérationnel, incluant la base de données, les méthodes de fouille et des interfaces d'interrogation et de visualisation des données et des résultats de fouille. L'outil doit permettre (1) de repérer les anomalies et défauts des données (2) d'aider à la mise en relation et à l'interprétation des données sur un cours d'eau (3) de tester et d'appliquer des méthodes de diagnostic de l'état et de l'évolution d'un cours d'eau.

*Remerciements*

*Ce travail est financé par l'agence nationale de la recherche dans le cadre du projet ANR 11 MONU 14 Fresqueau. De nombreuses personnes contribuent ou ont contribué au projet, qu'elles soient ici remerciées.*

## Bibliographie

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)

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*, p. 121-132. Toulouse, France, 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 (IDA 2013), London*, vol. LNCS 8207, p. 186–197. Springer.

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 AgroParisTech.

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