A lattice-based query system for assessing the quality of hydro-ecosystems

Agnès Braud, Cristina Nica, Corinne Grac, Florence Le Ber

A lattice-based query system for assessing the quality of hydro-ecosystems

Agnès Braud[1] Cristina Nica[2] Corinne Grac[3]
Florence Le Ber[3,4]

[1] LSIIT, CNRS-UdS, Strasbourg, France
[2] University Dunărea de Jos, Galați, Romania
[3] LHYGES, CNRS-ENGEES-UdS, Strasbourg, France
[4] LORIA – INRIA NGE, Nancy, France

CLA 2011 – 17-20 octobre 2011 – Nancy, France

Abstract

Concept lattices are useful tools for organising and querying data. In this paper we present an application of lattices for analysing and classifying stream sites described by physical, physico-chemical and biological parameters. Lattices are first used for building a hierarchy of site profiles which are annotated by hydro-ecologists. This hierarchy can then be queried to classify and assess new sites. The whole approach relies on an information system storing data about Alsatian stream sites and their parameters. A specific interface has been designed to manipulate the lattices and an incremental algorithm has been implemented to perform the query operations.
Keywords: incremental lattice, lattice-based query system, classification, information system, biological quality of water-bodies.

1 Introduction

Concept -or Galois- lattices are useful tools for organising, mining, and querying qualitative data in various application domains [14, 10, 24]. However when developing a domain specific lattice-based tool -to be used by domain analysts, a main problem is to define the proper approach and tool that fit the requirements of the experts and other users involved in the project. This paper presents an application of Galois lattices to the hydro-ecological domain, focussing on how to assess and monitor the ecological state of streams or water areas. These questions are currently major problems in Europe, as underlined by the recent European Water Framework Directive (2000). Assessing the ecological quality of streams requires to take into account various data such as physico-chemical measures on sites, but also taxonomic statements or qualitative information on species. Furthermore tools are needed to summarise all these data and to provide a global and reliable information on the ecological state of streams and water areas. Following this aim we have developed an information system to collect data on Alsatian streams (North-East of France) [17] and implemented a lattice-based query system to help hydro-ecologists to compare and assess the ecological state of streams. Concepts lattices are used: (1) to organise data, i.e. stream or water area sites with similar parameters are clustered within concepts; (2) to embed expert knowledge, i.e. concepts are annotated with an expert qualification or comment; (3) to perform queries, i.e. the annotated concepts are used to help assessing new sites of streams or water areas.

The paper is organised as follows. First (Section 2) we present the application domain. Section 3 is devoted to the principles of lattice-based querying. Sections 4 and 5 describe the principles and the implementation of our proposition. Section 6 compares our approach to other lattice-based tools and the last section is a conclusion.

2 Assessing the quality of hydro-ecosystems

The European Water Framework Directive (2000) requires the development of new tools for monitoring and assessing the quality of water-bodies (i.e. rivers, lake, gravel pits,...). Such an assessment is built on various information: information about the species living in the streams and physical, chemical and biological data collected on the sites. From these information are built several numerical indices that are synthetic indicators for assessing the physico-chemical or biological quality of an hydro-ecosystem.

More precisely, in France, five biological indices have been normalised to assess the quality of running water. They are based on three faunistic groups: the invertebrate index [1], the oligochaete (small worms living in sediments) index [3], the fish index [5], and on two floristic groups: the diatom (microscopic algae) index [2], and the macrophyte (macroscopic plants living in water) index [4]. Illustrations of the taxa used for these indices are given in Figure 1.

(a) Invertebrate (b) Oligochaete (c) Fish (d) Diatom (e) Macrophyte

Figure 1: Taxa examples for the five biological indices

According to AFNOR (French organism of normalisation) [1, 3, 5, 2, 4] each of them gives a different estimation of the water ecosystem quality. The macrophyte index estimates the trophic level of water, the diatom index gives the global water quality, the oligochaete index gives an evaluation of the sediment quality, and the fish index allows to classify the chemical and physical water quality quite like the invertebrate index. Therefore, their answers on a same site, with a same undergone pressure, at the same time can be really different but the simultaneous application of these five indices is not common and work comparing their answers are not frequent [20].

Furthermore, indices based on physical (e.g. width and slope of the stream bed) and physico-chemical (e.g. pH, temperature, nitrates, organic matters, pesticides) data give an other estimation of the ecosystem quality.

Thus, it is necessary to combine the various indices to assess the quality of a whole water ecosystem. Such an approach, called the *ecological ambiance system*, has been proposed in [20, 21] based on the five French biological indices. Our objective is to develop this concept and to propose a concretely applicable tool. We therefore rely on a large database collecting data on Alsatian streams and water areas [18]. The database contains 38 tables and it suits the SANDRE[1] French national format for aquatic data. It is implemented within the MySQL Database Management System.

The data are either issued from samples, synthetic data or general information issued from the literature. They are qualitative and quantitative, and suit the current standards about protocol sampling and indices computation based on thresholds [1, 3, 5, 2, 4, 22, 23]. Data issued from samples correspond to raw data. Synthetic data are produced from these samples, in particular taxonomic lists are used to compute biological indices. Data issued from the literature are used for the analysis and synthesis of the preceding data (for example they provide the thresholds for the classification of physical, physico-chemical and biological results into classes ranging from 1 (very good quality) to 5 (very bad quality)). We have gathered information on 700 sites in the Alsace Plain, the oldest one being collected 20 years ago. Details on this database and how it is used are given in [17].

3 Using lattices for querying databases

Galois lattices are useful tools for organising data and building knowledge bases [7, 14, 24]. Furthermore, they are very interesting for information retrieval since they allow both direct retrieval and browsing [16]. Primarily, concept lattices have been used for information retrieval within texts [25, 11]. More recently lattice-based approaches have been used to build query or information retrieval systems on various data: e.g. information retrieval within photos or personal data [13], geographical data [8], or museum collections [26]. The underlying hypothesis is that a concept extent represents the result of a query which is defined by the conjunction of its intent. The query can be easily refined or enlarged following the edges starting from the concept into the lattice hierarchy.

Practically, the query (a A set of attributes) can be performed as follows: the lattice is looked for a matching concept that is a concept which intent equals the A set -if it exists- or the most general concept which intent is larger than A. This concept can also be characterised as the infimum (greatest lower bound) of all the concepts containing at least one of the attributes of A. This can be done with various algorithms and the queried lattice does not have to be modified. Furthermore, a local view can be displayed to the user.

However, when the query represents a new object that is to be incorporated within the lattice, an incremental algorithm has to be used [15, 10]. This is the case in our application, since the user has got data about real stream sites which she/he wants to confront to the sites represented in the existing lattice. Furthermore, she/he can add the new sites to the lattice and thus modify its structure. We have implemented therefore two incremental algorithms proposed in [10], and roughly described in section 5.1. These algorithms have been chosen

[1] http://sandre.eaufrance.fr

because they allow to build the Hasse diagram of the lattice, contrarily to most of incremental algorithms (see [19] for a comparison on these algorithms). Furthermore, we did not look for performance, since in this first step of our work only small data sets (40 sites) have been considered.

4 Using lattices for assessing hydro-ecosystems

Lattices have been used in two ways: firstly to cluster stream sites into concepts that are used by hydro-ecologists to define *profiles* of these sites; secondly, the lattices are annotated with the profiles and used into a query-system to help the assessment of new sites. The proposed tool includes the two stages (see Section 5.2).

4.1 A lattice-based clustering of Alsatian stream sites

Stream sites are described by different numerical attributes, biological indices on the one hand, physico-chemical data on the other hand. Those attributes are converted into ordinal scales leading to quality classes. The whole context contains about 40 stream sites, described with 5 biological indices, 10 physico-chemical indices and 5 physical indices. In the following, we focus on the biological indices. Table 1 gives the values of these five indices restricted to seven sites. Each site is denoted by a code: for example, the BW2 site (Brunnwasser downstream) has a good quality (class 2) for the IBGN (invertebrate), IBD (diatom) and IPR (fish) indices, a bad quality (class 4) for the IBMR (macrophyte) index and an average quality (class 3) for the IOBS (oligochaete) index. The multi-valued context represented in table 1, denoted *C7* in the following, can be converted into a binary one by using a linear scale [14].

Site code	IBGN	IBMR	IOBS	IBD	IPR
BW2	2	4	3	2	2
IL1	3	3	3	2	3
MO1	1	4	3	3	4
MS2	2	4	5	2	2
RT2	2	5	4	2	2
ST1	1	3	4	3	2
ZN4	1	4	4	3	2

Table 1: Quality classes of the five biological indices for 7 stream sites

The general idea is to gather similar sites and to allocate them a profile describing their ecological state, combining the quality estimations of all compartments, with respect to the different classes of indices. This work is based on the approach described in [20]. The process is as follows:

- Step 1: Lattice construction on the data. To facilitate the expert analysis, the context size is reduced by focussing on a small number of indices or by identifying sub-lattices with respect to classes of indices. For example,

Figure 2 presents the lattice obtained from the context *C7* (the lattice was built with ConExp[2]) .

- Step 2: Analysis by the experts of the lattice hierarchy and its implication rules in order to select relevant concepts (or site profiles). In this step, the expert may identify profiles which are not present in the lattice and create virtual sites to be represented in the lattice.
- Step 3: Qualification of the concepts by the experts. For example, the concept ({IBGN 2, IBD 2, IPR 2, IBMR 4, IOBS 3},{BW2}) (down on the lattice, Figure 2) is interpreted as follows: *Brunnwasser downstream: low sediment degradation, high eutrophication, good general potential of resilience and possible resilience for sediments, various habitats.*

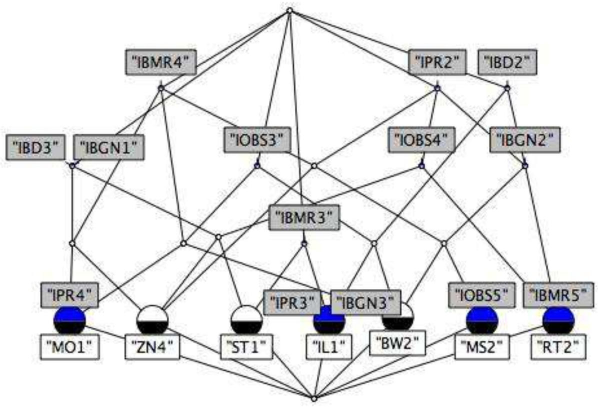

Figure 2: The lattice based on the context of table 1 (linear scale)

Once a suitable annotated lattice has been built following this process, it can be used to determine the profile of a new site based on its values for the corresponding indices. This is explained in the next section.

4.2 Assessing a stream site from the lattice

According to the *ecological ambiance system* described in [20], several lattices have been built for clustering sites with similar average values (or *alteration* degrees[3]) on the five biological indices. The underlying hypothesis is that global state of an hydro-ecosystem can be assessed on the basis of the five biological indices and synthesised by the alteration degree. Sites with similar alteration degrees can be compared even if they represent various profiles. The intervals of similarity have been defined by the hydro-ecologists [18]. For example, the

[2] `http://conexp.sourceforge.net/`

[3] The alteration degree is computed as the average value of the five biological indices, e.g. the alteration degree of BW2 equals 13/5. Currently the physico-chemical parameters are not taken into account.

lattice in Figure 2 was obtained from a set of sites with an alteration degree belonging to [2.5 ; 3] (see *C7* context in table 1). The classes of indices in the lattice vary between 1 and 5. Each site is represented alone in an atom of the lattice, which is coherent with the choices done in the project, trying to represent all the variety of streams or water areas in the Alsace plain.

Let us now suppose that we have got a partial information on a new stream site, denoted Q, defined by the following values: IBGN 2 IBMR 4 IOBS 3 IPR 2 (IBD missing). Its alteration degree is $2.75 \in [2.5 ; 3]$, Q can thus be compared to the stream sites represented in the *C7* lattice. This is done by classifying Q within this lattice, as shown in Figure 3.

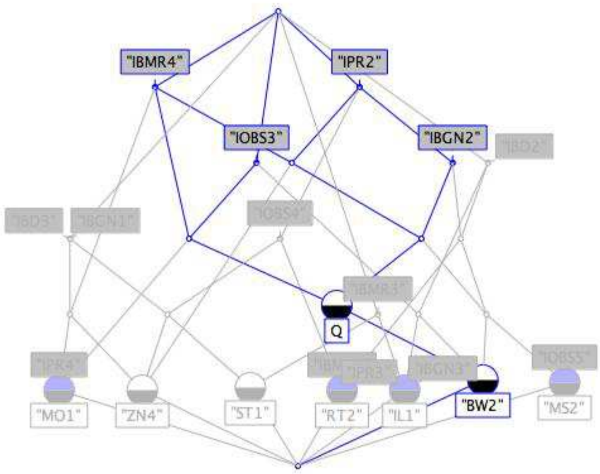

Figure 3: The *C7* lattice with the Q site-query inserted

Looking at the lattice in Figure 3, one can see that the Q site-query has four common values with only the BW2 site (Brunnwasser downstream). The expert qualification of BW2 (except for the IBD index) can thus be used to assess the Q site. The Q site could thus be assessed as follows: *the habitat quality and the water physico-chemical quality are good, expect for nutriments (nitrate and phosphor mineral forms) which quality is medium; the sediment quality is medium, the resilience potential of the general ecosystem is good, while the resilience potential of sediments is deteriorated.*

5 Implementation

5.1 Algorithms

As explained before, the built lattices have to be queried for assessing new sites. Furthermore, they could have to be updated, by adding a new site, or by modifying an existing site. The new/updated object is described by attributes which can exist in the context of the lattice or not. In this paper we only consider the case where the attributes already exist. Two algorithms described by Carpineto and Romano [10] have been implemented, the first one allows to add a new object in a lattice, while the second one allows to delete an object from a lattice.

The first algorithm allows to add a new object into an existing Galois lattice, which can be interpreted as classifying a new object. It takes as input a Galois lattice and the new object with its attributes. The output is the updated Galois lattice of the new context. The mechanism of the algorithm is as follows. The set of the concepts is divided into subsets according to their intent cardinality, and then analysed in ascending order. For each concept of a subset, if the intent is included in or equal to the set of the new object attributes then the current concept extent is augmented by the new object; otherwise a new concept is created, after verifying that such a concept is not in the initial set of concepts or among the new added ones. The intent of this new concept is determined by the intersection of the current concept intent and the new object attributes; its extent is defined by the current concept extent augmented with the new object. After the addition of a new concept a new link between this concept and the current concept is created. The links with neighbouring concepts are also updated.

The second algorithm allows to delete an object from a lattice. It takes as input a Galois lattice and the object to be removed. The output is the updated Galois lattice of the new context. The mechanism of the algorithm is as follows. For each concept, if the object to be deleted is included or equal to the current concept extent, then it is removed from this extent. If the modified concept has then the same extent as one of its children, it is deleted. When a concept is removed the links among the concepts are updated.

The modification of an existing object in a Galois lattice is performed in two steps: (1) deleting this object using the second algorithm; (2) adding the updated object using the first algorithm. The whole process could be improved with a third algorithm for adding attributes into the lattice context, allowing to enrich the initial lattice with new information.

5.2 User interface and manipulation

The user interface allows to use a lattice either stored in the database or stored in a `XML` file with the structure used in the software Galicia[4]. Three main functional views are provided to the user. The first one allows to qualify concepts, i.e. to describe the profile of a set of sites. The second one allows to define a query, i.e. a new site to be assessed according to an existing lattice. The third view allows to explore the result of the query, i.e. to compare the characteristics of the new site to those of the already assessed sites. Currently texts appearing on the interface views are written in French since the target users are French. Other languages could be used in the future.

The functional view for qualifying concepts is presented on Figure 4. Once a lattice is chosen, it is possible to select a given concept in a list and to see its description (intent, extent, and comment). The lists of the parents and children of that concept are also shown, and by a click on one of them, we see its related information. These information may help the experts in qualifying the concept.

[4] `http://www.iro.umontreal.ca/~galicia/`

The comment is then stored in the database.

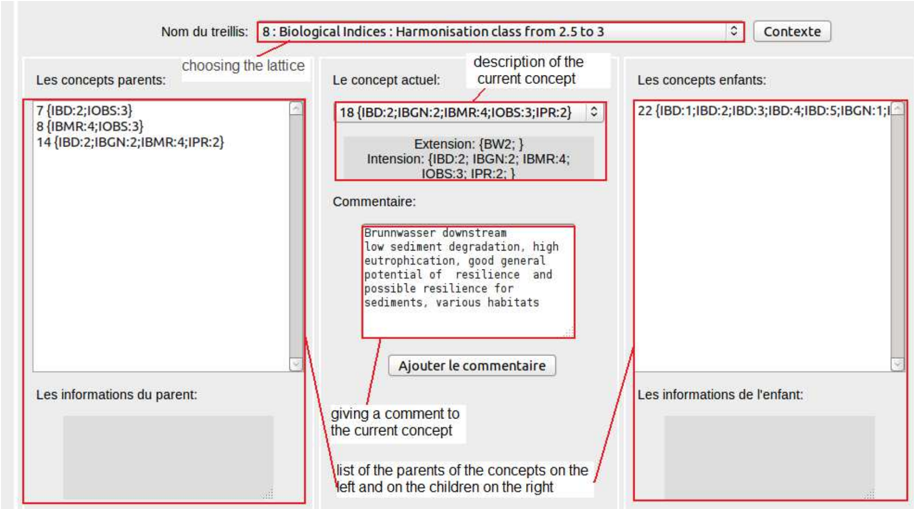

Figure 4: Qualifying the concepts of the site lattice

The functionality for classifying a new site based on its values (for one or several indices) is presented on Figure 5. One has first to select a lattice and to give a name for the new site, and then to provide a description of this new site by choosing indices and their values. Once this is done, it is possible to classify the site, that is to integrate it in the lattice, either temporarily or to save it in the lattice. The button "Classer" allows this classification. To interpret the result, the button "Visualiser le résultat" can be used to see the new lattice with the modifications shown in a specific colour. The button "Explorer le treillis" also helps in the interpretation by giving access to a third view (Figure 6) where it is possible to navigate within the concepts and see the description of the parents and children of the current concept.

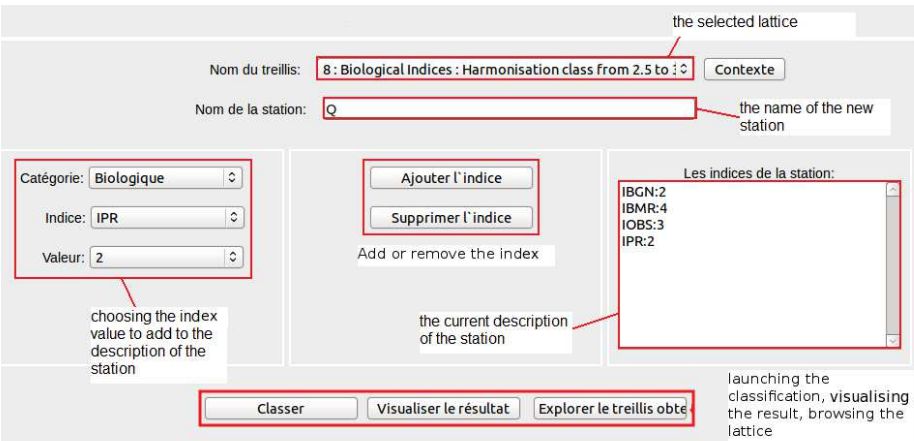

Figure 5: Definition of the Q site-query

More precisely, the third view allows to explore only the modified or new concepts of the lattice, i.e. the concepts where the site-query is represented. These concepts can be commented and the modified lattice can be stored in the database. Eventually, the commented lattices can be exported in various formats to be further analysed.

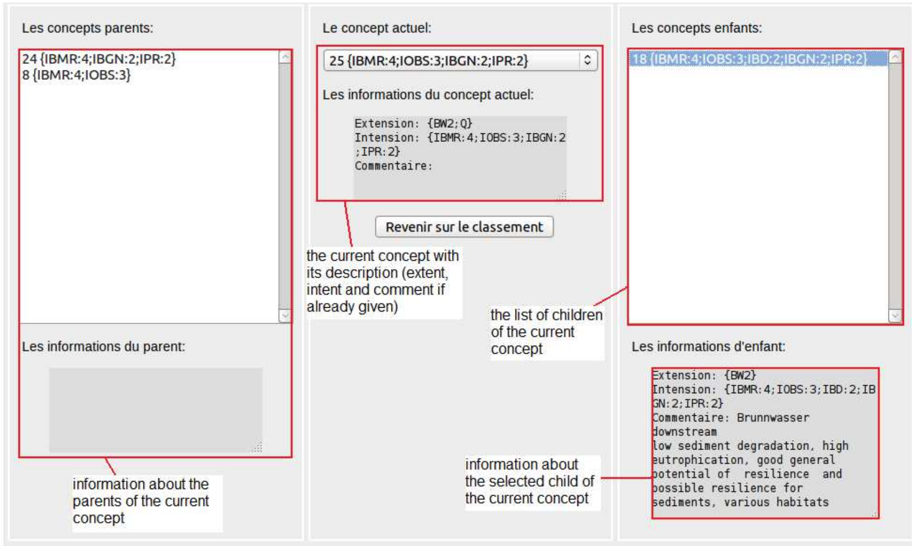

Figure 6: Analysing the classification result of the Q query

6 Discussion

We decided to implement a specific tool for several reasons:

1. the tool has to be interconnected with a database and to offer a user-friendly interface for hydro-ecologists, allowing them to annotate the concepts;

2. the purpose of the tool is not navigating throughout the whole database;

3. this is a two-stage tool: the first stage organises a specific information within a lattice; the second stage allows the user to explore and possibly modify this lattice.

Regarding the first point, lattice-builder tools like Galicia, ConExp, or the Toscana suite[5] cannot be used, since they do not fit the requirements of hydro-ecologists. Actually, as said before, we have used Galicia to build the lattices which are then recorded in the database to be annotated and explored by hydro-ecologists. Besides, the lattices built through our tool can be exported into a Galicia format.

Regarding the second point, our approach differs from those used in search or browsing tools like Camelis [13], Abilis [6], D-SIFT [12] or in the Virtual Museum of the Pacific [26]. Indeed we did not try to implement a lattice-based

[5] `http://toscanaj.sourceforge.net/`

approach to explore the whole database, but only specific information from this database. This information was chosen by hydro-ecologists as a synthetic view of the database. Furthermore, the lattice is used as a basis to record expert knowledge (the annotations) that can be involved in further investigations.

Regarding the last point, our tool can be compared to Ulysses [9] which is a visual interface allowing to access a lattice structure organising information from a database. Ulysses allows the user to search the retrieval space both by browsing or querying, whereas our tool only allows querying. Nevertheless, the originality of our tool is the user possibility of modifying and annotating the lattice concepts.

Finally, the underlying aim of our approach is to build an ontology, gathering the knowledge of various experts on hydro-ecosystems. Each expert indeed focuses on a specific compartment of the hydro-ecosystems (e.g. fishes, macrophytes, diatoms...) and a generic tool is needed to combine their expertises and produce a global assessment of the ecological state of a stream site.

7 Conclusion

This paper presents a lattice-based query system for helping the assessment of hydro-ecosystems. The approach relies on a database storing various information on stream sites of the Alsace plain. These data are summarised within qualitative indices, biological indices or physico-chemical and physical indices. Based on these indices and their own expertise, hydro-ecologists can perform a global evaluation of the functioning of a stream ecosystem. Furthermore, they want to define quality profiles of streams or water areas that could be used to assess new sites. Eventually a tool is needed to help the whole process.

Our work aims at building such a tool. Concept lattices appeared as a good approach since they allow both to build hierarchical clustering of sites, to navigate through the clusters, and to perform queries for helping the assessment of a new site. The clustering aspects already proved to be interesting, and the user interface allowing to comment and query the lattices is currently being experimented by hydro-ecologists. In the future, several lattices have to be built including various sets of indices (physico-chemical and physical indices). Furthermore, the whole approach will be tested with stream or water area data from other regions in France.

Regarding the implementation aspects, the system should be improved in two ways: allowing the integration of new attributes in an existing lattice and allowing the navigation through bigger lattices. Finally improvements can be done to provide self-building comments on the site-queries, based on the comments of the neighbouring concepts.

Acknowledgements

The Indice project (2006-11) was supported by the Agence de l'Eau Rhin-Meuse. We also acknowledge the scientific and technical help of the Cemagref Centre in Lyon, the Gabriel Lippmann Public Research Centre in Luxembourg and the regional delegation of ONEMA (Office National de l'Eau et des Milieux Aquatiques). Cristina Nica's stay in France was supported by the Erasmus

European program. We acknowledge the anonymous reviewers who helped us to improve our paper.

References

[1] AFNOR: Qualité de l'eau : détermination de l'indice biologique global normalisé (IBGN). NF T90-350 (1992), révision 2004

[2] AFNOR: Qualité de l'eau : détermination de l'indice biologique diatomées (IBD. NF T90-354 (2000), révision 2007

[3] AFNOR: Qualité de l'eau : détermination de l'indice oligochètes de bioindication des sédiments (IOBS). NF T90-390 (2002)

[4] AFNOR: Qualité de l'eau : détermination de l'indice biologique macrophytique en rivière (IBMR). NF T90-395 (2003)

[5] AFNOR: Qualité de l'eau : détermination de l'Indice poissons rivière (IPR). NF T90-344 (2004)

[6] Allard, P., Ferré, S., Ridoux, O.: Discovering Functional Dependencies and Association Rules by Navigating in a Lattice of OLAP Views. In: Kryszkiewicz, M., Obiedkov, S. (eds.) Proceedings of CLA 2010, Sevilla, Spain. pp. 199–210 (2010)

[7] Barbut, M., Monjardet, B.: Ordre et classification – Algèbre et combinatoire. Hachette (1970)

[8] Bedel, O., Ferré, S., Ridoux, O., Quesseveur, E.: GEOLIS: a logical information system for geographical data. Revue Internationale de Géomatique 17, 371–390 (2007)

[9] Carpineto, C., Romano, G.: ULYSSES: A Lattice-based Multiple Interaction Strategy Retrieval Interface. In: Blumenthal, B., Gornostaev, J., Unger, C. (eds.) Human-Computer Interaction, 5th International Conference, EWHCI'95, Moscow, Russia. LNCS, vol. 1015, pp. 91–104. Springer-Verlag (1995)

[10] Carpineto, C., Romano, G.: Concept Data Analysis. Theory and Applications. John Wiley & Sons Ltd (2004), 201 pages

[11] Carpineto, C., Romano, G.: Using concept lattices for text retrieval and mining. In: Ganter, B., Stumme, G., Wille, R. (eds.) Formal Concept Analysis, LNCS, vol. 3626, pp. 3–45. Springer Berlin / Heidelberg (2005)

[12] Ducrou, J., Wormuth, B., Eklund, P.: D-SIFT: A Dynamic Simple Intuitive FCA Tool. In: Dau, F., Mugnier, M.L., Stumme, G. (eds.) Conceptual Structures: Common Semantics for Sharing Knowledge – Proceedings of ICCS 2005. vol. LNAI 3596, pp. 295–306. Springer-Verlag (2005)

[13] Ferré, S.: Camelis: a logical information system to organise and browse a collection of documents. International Journal of General Systems 38(4), 379–403 (2009)

[14] Ganter, B., Wille, R.: Formal Concept Analysis: Mathematical Foundations. Springer Verlag (1999)

[15] Godin, R., Missaoui, R., Alaoui, H.: Incremental concept formation algorithm based on Galois (concept) lattices. Computational Intelligence 11(2), 246–267 (1995)

[16] Godin, R., Missaoui, R., April, A.: Experimental comparison of navigation in a Galois lattice with conventional information retrieval method. International Journal of Man-Machine Studies 38, 747–767 (1993)

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

[18] Grac, C., Le Ber, F., Braud, A., Trémolières, M., Bertaux, A., Herrmann, A., Manné, S., Lafont, M.: Programme de recherche-développement *Indices* – rapport scienfique final. Contrat pluriannuel 1463 de l'Agence de l'Eau Rhin-Meuse, LHYGES – LSIIT – ONEMA – CEMAGREF (2011)

[19] Kuznetsov, S.O., Obiedkov, S.A.: Comparing performance of algorithms for generating concept lattices. J. Exp. Theor. Artif. Inelligence 14(2-3), 189–216 (2002)

[20] Lafont, M.: A conceptual approach to the biomonitoring of freshwater: the ecological ambience system. Journal of Limnology 6, 17–24 (2001)

[21] Lafont, M., Jézéquel, C., Vivier, A., Breil, P., Schmitt, L., Bernoud, S.: Refinement of biomonitoring of urban water courses by combining descriptive and ecohydrological approaches. Ecohydrol. Hydrobiol. 10, 3–11 (2010)

[22] MEDD: Système d'évaluation de la qualité de l'eau des cours d'eau (SEQ-Eau), version 2. Ministère de l'écologie et du développement durable et agences de l'eau (2003), Étude inter-agences de l'eau, no 52

[23] MEDD: Circulaire dce 2007/22 du 11 avril 2007 relative au protocole de prélèvement et de traitement des échantillons des invertébrés pour la mise en œuvre du programme de surveillance sur cours d'eau. Ministère de l'écologie et du développement durable (2007)

[24] Napoli, A.: A smooth introduction to symbolic methods in knowledge discovery. In: Cohen, H., Lefebvre, C. (eds.) Categorization in Cognitive Science. Elsevier (2006)

[25] Priss, U.: Lattice-based information retrieval. Knowledge Organization 27(3), 132142 (2000)

[26] Wray, T., Eklund, P.: Exploring the Information Space of Cultural Collections Using Formal Concept Analysis. In: Valtchev, P., Jäschke, R. (eds.) Proceedings of 9th International Conference on Formal Concept Analysis, ICFCA 2011, Nicosia, Cyprus. LNAI, vol. 6628, pp. 251–266. Springer-Verlag (2011)