

Chemical complexity and molecular topology

The DARC concepts and applications

Jacques-Émile Dubois



Nous sommes heureux de livrer aux lecteurs l'essentiel d'un article de Jacques-Émile Dubois paru fin 2004, quelques mois avant son décès, dans *Proceedings of the 2002 Conference on The History and Heritage of Scientific and Technological Information Systems, Ed. The Chemical Heritage Foundation and ASIST, 2004, p.149*. Nous remercions les éditeurs – Chemical Heritage Foundation – qui nous ont permis cette adaptation.

Résumé

Complexité chimique et topologie moléculaire : concepts et applications du DARC

Dans les années 50, il a été observé que la position relative des atomes de carbone situés dans l'environnement immédiat de la fonction carbonyle des cétones contribuait à modifier les propriétés de ces molécules. Ceci a conduit à représenter les cétones aliphatiques par un code topologique. Cette représentation a été ensuite généralisée en créant un nouveau système de nomenclature : le code topologique DARC (Documentation et Automatisation des Recherches de Corrélations), applicable à toute la chimie organique. Cette représentation originale repose sur l'appréhension du graphe associé à la molécule comme une succession d'environnements limités ordonnés. L'épopée et les atouts du DARC dans la constitution et la gestion d'importantes bases de données « online » sont analysés dans le contexte d'une informatique chimique alors en pleine évolution (1970, 1985, 1990). Des exemples physico-chimiques concrets illustrent l'apport de la topologie, et montrent comment la théorie de l'information permet de traiter avec précision des ensembles de données complexes, dans de vastes domaines de connaissance.

Mots-clés

Chimie topologique, recherche de sous-structure, base de données chimiques, graphe chimique, corrélation structure activité, système d'information chimique.

Abstract

In the 1950s it was found that all the sites of a well-defined carbon environment of the ketone function contribute to the expression of its properties. This justified a new description of ketones with a topological code. This treatment was generalized to achieve a new, original nomenclature called DARC (Documentation and Automated Research of Correlations), valid for the whole of organic chemistry. This original coding breaks down the support graph of a structure into a series of local ordered rooted trees of limited size. The DARC epic of constituting large structural databases on-line (1970, 1985, 1990) is evoked in the context of that era of computerizing chemical data. The advantage of expressing organic structures topologically is illustrated by physical-chemical examples; information theory allows handling of complex data spread over vast areas of study.

Keywords

Chemical topology, substructure search, chemical database, chemical graph, activity structure correlation, chemical information system.

The place and the impact of topology on chemistry cannot be analyzed without some preliminary reminders of the present state of the art of structural organic chemistry. One needs to recall briefly some of its basic concepts and conventions in order to comprehend fully their interactions or influence on topology. In facilitating computerization, topology is instrumental in advancing many aspects of structural chemistry. It modifies qualitatively some chemical concepts and helps to handle differences from standard representational notions.

Structural chemistry background

Two main concepts or principles constitute the backbone of structural chemistry. The first considers that molecules or structures result from joined groups of atoms. Some of them

are more active and are usually considered as functions; others build the skeleton or framework. The other important concept is that of homogeneous families of structures organized either in ascending order, for homologues, or with a common global atomic identity, for isomers.

To facilitate communication and the elaboration of reference files of the numerous structures, chemists use rules of constitution and graphic representation of structures to discuss the structures by systematic nomenclature and structural chemistry. These have developed in symbiotic ways, and the syntax of systematic nomenclature bears a close relation to the ways in which structural chemistry is shaped and taught to future chemists.

The advances come from the converging efforts of the International Union of Pure and Applied Chemistry (IUPAC),

the Chemical Abstracts Service (CAS), *Beilstein's Handbook of Organic Chemistry*, and other institutions that constantly enlarge the field of action of systematic nomenclature as science progresses. As a living science, though, chemistry continues to use many common or trivial names for each structure, mainly because of the difficulties encountered with the oral expression of systematic names, essentially derived to give a unique and complete description of structures.

With computing, new horizons appeared in chemistry, both in its information-documentation aspects and in designing molecular syntheses. One could hope to handle efficiently the explosion of new compounds being synthesized and elucidated as well as the exponentially increasing amounts of analytical data produced by new physical methods.

The information revolution opened up a variety of avenues for structural chemistry. Statistical surveys of large files helped identify substructures not considered chemically active fragments but nevertheless useful as screen or meta data in chemical design strategies. I propose here that the codes and computer-retrieval tools of structural data are instrumental in designing information management systems but should not themselves be called global information systems.

Structural information resources before DARC

In this paper I will emphasize the academic origins of the DARC (Documentation and Automated Research of Correlations) topological system.

The situation before DARC

In the 1960s, data handling was restricted to the information institutions: CAS, patent offices and industrial sectors. Scientists were users of primary and secondary information, but information research was not in the academic curriculum.

I became interested in aspects of information through my research in spectroscopy, which I pursued in the 1950s. Later I began new academic teaching areas in chemical and bioinformatics, but subsequently, we had to develop our DARC system, partly through external associations and with non-academic partners.

Spectroscopy: evidence of an extended carbonyl functional group

In the 1950s at the University of Saarbrücken, I began a research program to evaluate quantitatively the alkyl influence on the environment EX (e.g. E_1 , E_2 , etc.) of the carbonyl group, leading to red shifts of the $n \rightarrow \pi^*$ transition in the ultraviolet range. For this we had to consider synthesizing ketones belonging to the aliphatic ketone families with crowded and hypercrowded environments of the carbonyls in the α_i and β_{ij} positions, which meant changing the groups R and R1 in the general formula RCOR1.

Our first deduction, based on experience, was that one could neglect as a first approximation those long-range influences that go farther than the carbons α and β – that is, the first and second carbons of the environment, later named A and B. Moreover, rather than evaluate alkyl group actions, we could identify the influence of the carbon sites in the A and B positions (*figure 1*) and the practical absence of influence beyond these atoms.

The environment E of the $-C(CO)C-$ root of the graph or focus of a family of aliphatic ketones consists of two segments

or modules E_1 and E_2 contributions. One can identify the contributions of all atoms of the first layer of atoms of the R and R1 alkyl groups. Even the action of atoms B can be evaluated. The $\Delta\nu$ contributions are calculated by comparison with the acetone chosen as reference for the alkylacetone family. The site contribution is achieved through a correlation over a large population of ketones.

These findings resulted in precise rules (the Dubois-Maroni rules) that were confirmed and improved during the project's continuation in my laboratory at the University of Paris. Our rules probably provided the first systematic labeling of the EB carbons, although much later similar contributions on the topological side became common practice with ^{13}C nuclear magnetic resonance spectroscopy, and each carbon has a δc shift value (chemical shift).

This research developed over a number of years confirmed our initial Dubois-Maroni prediction rules. Limited but similar site interventions were provided by the Woodward-Fieser rules, and we decided to investigate them more fully with our topological concepts. Our results will be discussed later in this paper.

For our research programs we found it difficult to evaluate our various syntheses of ketones. Moreover, we had little information about crowding effects on the synthesis of encumbered species.

All these constraints and results constantly stimulated us to look for better structural representation, both for direct communication as well as for computerized operations. We looked for an original quantitative handling of alkyl environments and opened a new paradigm in this field when, for interpretation, we turned to topological theory and ordered graph representation of our ketones.

A topological representation of alkyl environments: DARC concepts

One Sunday morning in March 1964, I found a way to express our $A_i B_{ij}$ environmental sites by locating them in a special matrix (*figure 2*) corresponding to our quantitative A_i and B_{ij} validations by means of the Dubois-Maroni rules (*figure 1*).

In the special matrix adopted, called "E" for environment, we separate the three B_{ij} carbons as substituents of any A_i site (A_1 , A_2 , A_3 , or even A_4), easily designated by their

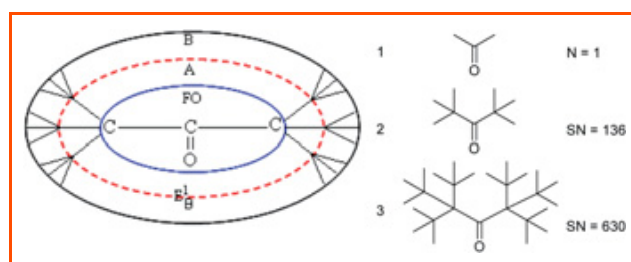


Figure 1 - α_i and β_{ij} increments in the prediction rules to identify $n \rightarrow \pi^*$ site contributions in ketones.

$$\Delta\nu_{RRI} = \Delta\nu_R - \Delta\nu_{Ri} - a\Delta\nu_R\Delta\nu_{Ri} = S - aP$$

$$\Delta\nu_R^{\text{calc}} = \sum_{i=1}^3 n_i \alpha_i + p_i \beta_i$$

α_i and β_i are the increments corresponding to groups bonded to C_{α} and C_{β} carbons, respectively; n_i and p_i are the corresponding numbers of these substituents. SN are the theoretical numbers of molecules corresponding to the different graphs.

developments, we named the system DARC (Documentation and Automated Research of Correlations). To cover the diversity of concepts discovered in our research, three theories were conceived and presented together, as they are indeed interwoven.

The DARC system: three basic theories

The DARC system encompassed three main theories closely linking structural representation and its use in different fields of chemistry (figure 5):

1. Theory of generation-description: this deals with the development and application of the language of DARC-DUPEL (Description that is Uniline by Propagation of an Environment that is Limited in B) which generates the representation of structures and substructures for a chemical database management system.

2. Theory of population-correlation: this involves the organization of spaces of states, such as the representation of structural populations and of the physical property spaces of those populations. Such organizations imply hybrid and complex representations, including classical and topological descriptors to use for QSPR (“quantitative structure-property relationship”) and QSAR strategies.

3. Theory of topo-information correlations: DARC is a topological theory of environment effects in physical chemistry. The E matrix is used to derive topological validation of active sites within a structure-property space.

	DB-KB	AI/Dgn	CONCEPTS
• General-Description Theory	+++	+	S, HS, SS
• Population-Correlation Theory	++	++	P-S/HS; P-I
• Topo-Information Theory	+	+++	P-S/HS/I

Figure 5 - The DARC system: three basic theories and applications.

DB-KB = data and knowledge bases; AI/Dgn = artificial intelligence/design; P = properties (1965, Documentation and Automatic Research of Correlation; 1970, Documentation Acquisition, Retrieval and Design).

From DARC concepts to structural on-line services

Generation-description theory: the DARC-DUPEL generator

The first theory and its common language DARC-DUPEL are better known than the others in the information profession. We will therefore elaborate on it more fully and only briefly present the two others that primarily concern scientific and academic applications.

To code a graphical formula $G(X)$ in DARC language, one uses its topomodel described as: $T(G) = \text{focus (FO)} + \text{environment (E)}$. Both E and FO are coded similarly, but in documentation FO = 1 because it refers to one unique site only.

Thus the coding of G is essentially the choice of unique FO and the coding of E. This is done by a combined or generating process segmenting E in successive EB starting from the focus up to the terminal sites of G. The process obeys priority rules. It starts by creating a spanning tree (EXS), expressing the existence of a spanning S graph associated to G but with no ring closures. This underlying graph of all further information is called “existence graph”, $G(\text{EX})$.

This basic topological organization of $T(G) = T(\text{EX})$ in segments is based on ordering rules of local connectivity. To order the sites, ordering decisions are based on the absolute priority given to the highest connectivity of all neighboring sites.

All chromatic information is added within the ELCOs of (EXS). This information includes the nature of bond-links (LI) of atoms (NA, also called atomic numbers), geometry, chirality, conformations, electronegativities and the ring closures (figure 6).

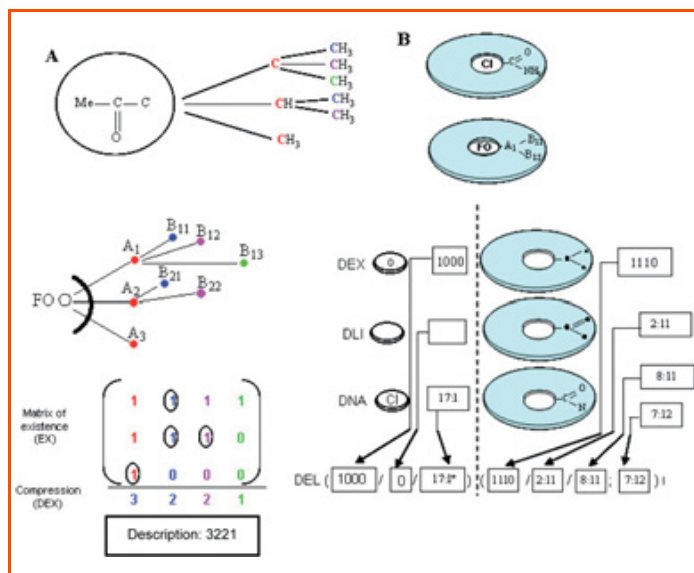


Figure 6 - A) E matrix of methylketone and its DEL: FO/3221; B) chromatic description of a structure, including its focus labeling.

Overlapping substructures: FREL

Whereas ELCOB were used as joint segments, it is interesting to create a larger number of different EB over the structure by choosing all sites as local foci. One then generates overlapping segments, which we prefer to name FRELs (Fragments Reduced to a Limited Environment). The operation of overlapping fragments, here FRELs, was not used in past documentation strategies, although topological overlapping can lead to interesting observations. The FRELs on the sites are fascinating formal fragments for the query system of any database.

In the DARC system we consider that the union of all the overlapping FRELs of a $G(X)$ constitutes another DARC name of the structure.

Canonicalization procedure

Extended and local connectivity

In the 1960s there was an urgent need to represent structures making use of the computer facilities that were developing so rapidly. The goals were clear, but the problem of transforming a graphic structure into a machine representation had no obvious solutions. Competition and stimulation during these years were too important to be summarized in a few lines. Readers interested in this period will find two excellent books in my reference list of suggested reading (Lynch *et al.*, 1971; Davis & Rush, 1974).

In early theoretical work topology played an important role, for example, in quantum chemistry, but ordering was usually associated with a specific project. Things became more complex with the intent to produce unique structure identifiers. Methods had to be found to designate, by convention, a unique canonical matrix among the many matrices resulting from all possible numberings of a structure.

In the early 1960s two methods were found simultaneously. L.H. Morgan at CAS used an "extended connectivity" program to determine the unique focus of the underlying graph G . With it came a sequential and unique ordering of all atoms of S .

With the DARC system we developed a canonical search algorithm based on "local connectivity" using the DARC segmentation of the entire underlying graph $G(EX)$ in successive ELCOB.

From its inception the "local connectivity" EB concept was oriented to deal with chromatic graphs. The "existence graph" $[G(EX)]$ led to a sequential ordering of all vertices.

In both strategies, any chromatic information is localized on the topological coordinates of the defining spanning tree graph. In both methods one ends with the equivalent rooted tree description, including a choice of a specific root.

DARC-DUPEL and chemical finalization and optimization

Chemistry is complex and deals with many functional classes described by conventions. A formal language such as DARC needs to be adapted to integrate those conventions. To deal with the complexity of these numerous aspects, one needs to access large working files and to interact with their information specialists. For this I created a small private association called the Association for Research and Documentation in Chemical Informatics (ARDIC) and hired professional computer people to develop and implement DARC structural tools and programs.

Simultaneously, with the DEL structural coding of a whole structure, the ARDIC group proposed a similar descriptor, FREL for substructures, handled as finite and ordered subgraphs or fragments. ARDIC became the normal interface for different exchange projects that we had launched before its existence. We reinforced our cooperation with the Basel Group of Pharmaceutical Industries, which began in the mid-1960s, and with CAS, begun early in 1970. The Basel Group lent us some connection table files, and we were able to start with ten thousand, then with five hundred thousand connection tables. We then began to create structural inter-conversion programs and to investigate various topological searches. We carried out a stimulating task of interconversion of the CAS chemical and biological activities file (more than five hundred thousand connection tables) at the time of its computerization from the CAS program to DARC-DUPEL. We put a chemical and biological activities-DARC version on-line around 1975 to test it with industry users (*figure 7*). Query strategies are diversified, but the two main approaches are based either on FREL screening or on a topological search of an organized file tree. Answers include a final graphic edition of formulas.

During those years of rapid change, extensive work in the IUPAC Interdivisional Committee on Machine Documentation (that it was my privilege to chair from 1968 to 1976) created an international community of interest that benefited all its member experts.

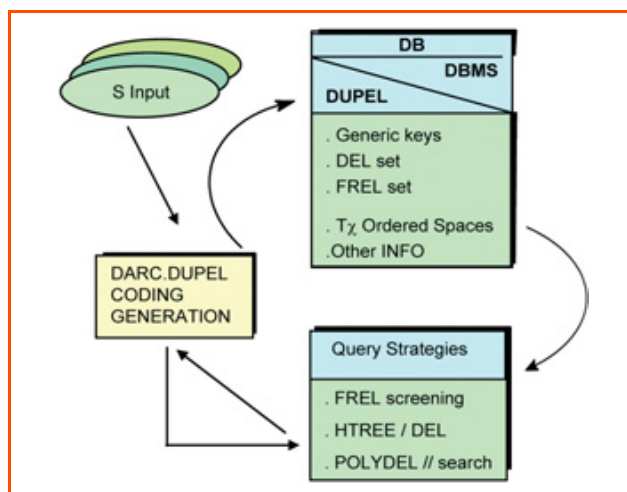


Figure 7 - DARC logistics for structured data management systems.

ACS-CAS-CNIC convention

For the DARC the final stage of this epic came in 1978 when the Centre National d'Information en Chimie (CNIC, National Center for Chemical Information), a newly created French chemical organization, signed a cooperative agreement with CAS. We then received and transcoded the entire CAS registry and produced EURECAS (the European DARC version of the CAS file) as an on-line service monitored for two months by the development team of ARDIC. We then transferred the EURECAS and chemical and biological activities files and other by-products to the CNIC for QUESTEL, which became the developer vendor agent. In our EURECAS database, the management system was based on an artificial intelligence tool called DARC-TSS (Topological Screen System). Partial extensions were managed both for specific populations (e.g. nuclear magnetic resonance) and for proper DARC-Markush generic structures⁽¹⁾.

Used worldwide for patent searching, the DARC system is useful wherever a large number of graphs are involved. Further DARC theories were to broaden our topological vision of chemistry and add new potentialities. The production and nature of topological chemical knowledge helped us to later develop elucidation and synthesis methodologies in parallel.

Conclusion

Topology is a natural constituent of structural and theoretical chemistry. In explicit developments it has been instrumental in the computerization of chemistry and the birth of numerous information products.

Over the past fifty years, topology has had positive effects on the representation and management of different subsets of chemistry: structures, substructures, large collections and files of structural data. In particular, it has been used to simplify the computer interfaces used in chemical computer-aided design and correlation work. Our work has convinced us of the necessity of handling synchronously the topological expressions of s (a single atom considered as a substructure, e.g. Cl, whereas C has other atoms in its minimal connectivity, e.g. C-CH₃; ss is a limited form of substructure), substructure, structure and hyperstructure; their interplay is needed to deal with chemistry in all its complexity. The three components of

data, information and knowledge are general and are valid for all disciplines. They are the communicating poles of the trilogy shown in figure 8. Each of these poles has specific tools and products. They all fertilize and modify their neighbors, even as they themselves undergo modifications.

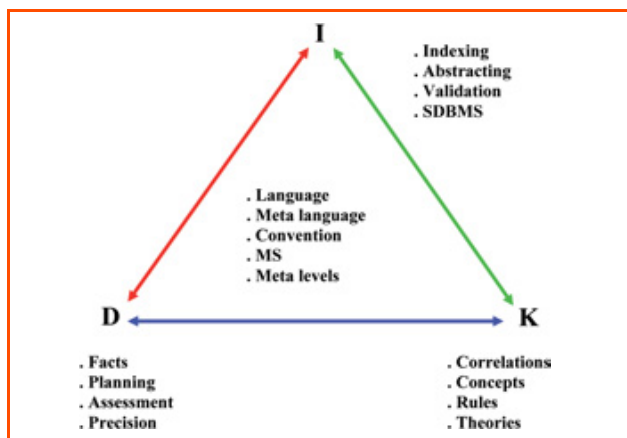


Figure 8 - Data-information-knowledge (DIK) interplay: interfaces or synchronous effects.

Each improvement of any of these poles can benefit the others. New data flows to databases structured by fragmentation and can produce better information on which knowledge can be based. Thus each pole, besides its own logistics, depends on topology to maintain an easy and harmonious cyclic data-information-knowledge flow, which results in a continuous and dynamic evolution, enhanced by faster, clearer communication. This is likely to remain the best hope for progress in chemistry and science in general.

Acknowledgments

I do not want to conclude without expressing my sincerest congratulations and my warmest gratitude to Mary Ellen

Bowden – congratulations for her having organized this highly successful conference and gratitude for her hard work, patience, competence and invaluable help in the editing of this conference paper. I also thank the Chemical Heritage Foundation for the invitation kindly extended to me, which I was more than happy to accept.

Note and suggested reading

(1) A *Markush structure* is a generic structure describing specific and nonidentified structures. One such structure may cover hundreds of thousands of identified candidates within a single patent coverage. It corresponds to a complex hyperstructure.

I recommend the following books for a good introduction to this field:

- Lifschutz S., Lifson M.L., *Discrete Mathematics*, McGraw Hill, New York, 1992, chapters 5, 6 & 7.

- Balakrishnan V.K., *Graph Theory*, McGraw Hill, New York, 1997.

- Valiente G., *Algorithms on Trees and Graphs*, Springer, New York, 1998 (an overview of LEDA, Library of C++ data structures and algorithms for graph handling).

Structural coding history: "en route" to computers:

- Davis C.H., Rush J.E., *Contributions in Librarianship and Information Science*, 1974, N8, p. 230.

- Lynch M.E., Harrison J.M., Town W.G., Ash J.E., *Computer Handling of Chemical Structure Information*, American Elsevier (Computer Monographs), New York, 1971.

DARC system: concepts and theories:

- Dubois J.-E., Laurent D., Viellard H., *C.R. de l'Académie des sciences*, 1966 & 1967, 263C, p. 764; 263, p. 1245; 264, p. 348.

DARC generation: description theory and databases:

- Dubois J.-E., *Chemical Applications of Graph Theory*, A.T. Balaban (ed), Academic Press, New York, 1976, p. 333.

- Dubois J.-E., Hennequin F., *Bull. de la Société Chimique de France*, 1966, 11, p. 3572.

- Dubois J.-E., Hennequin F., Boussu M., *Bull. de la Société Chimique de France*, 1969, 10, p. 3615.

- Dubois J.-E., *Molecular Similarity and Reactivity from Quantum Chemical to Phenomenological Approaches*, R. Carbo (ed), Kluwer Academic, Dordrecht, Netherlands, 1995, p. 203.

- Dubois J.-E., Sicouri G., Picchiottino R., *Modern Approaches to Chemical Reaction Searching*, P. Willett (ed), Gower, Brookfield, VT, 1985, p. 240.

- Dubois J.-E., Sobel Y.-J., *Chemical Information and Computer Science* (25th Anniversary Issue), 1985, 25, p. 326.

DARC topo-information correlations:

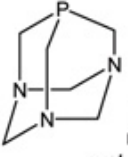
- Dubois J.-E., *3rd IUPAC Conference in Physical Chemistry, Pure and Applied Chemistry*, Pergamon Press, Cambridge, 1977, 49, p. 1027.

- Dubois J.-E., Bienvenue A., Chastrette M., *Chemical Communications*, 1968, p. 439.

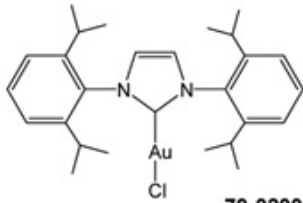
- Dubois J.-E., Loukianoff M., *SAR and QSAR in Environmental Research*, 1993, 1, p. 63.

- Maroni P., Dubois J.-E., *C. R. de l'Académie des sciences*, 1963, 256, p. 5351.

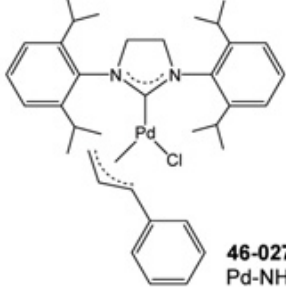
STREM Chemicals, Inc.
 Since 1964



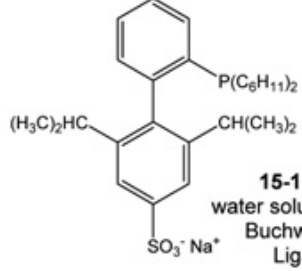
15-5710
non-pyro,
water soluble
PMe₃ analog



79-0200
Au-NHC



46-0274
Pd-NHC



15-1135
water soluble
Buchwald
Ligand

15, rue de l'Atome, Zone Industrielle
 67800 BISCHEIM, France

Tel.: (33) 03 88 62 52 60 • Fax: (33) 03 88 62 26 81
 email: strem.europe@wanadoo.fr • www.strem.com