Jacques-Émile Dubois, pionnier de l'informatique chimique et inventeur du DARC

Le système DARC : de la recherche fondamentale aux transferts technologiques

Daniel Laurent

C'est en 1966 que Jaques-Émile Dubois présente, dans une note à l'Académie des sciences, les principes généraux du système DARC (Documentation et Automatisation des Recherches de Corrélations) [1]. Ce fut pour plusieurs générations de chercheurs une véritable aventure scientifique à travers laquelle J.-E. Dubois a exprimé toute la mesure de son talent et de sa personnalité : créativité, imagination, vision et sens de l'organisation. En fait, la genèse du système DARC remonte à 1954, la publication de 1966 étant le fruit d'une longue maturation. J.-E. Dubois, dès le début de sa carrière universitaire, s'est préoccupé d'étudier le comportement d'importantes séries de composés comme il le fit par exemple en cinétique rapide (bromation des oléfines) ou en spectroscopie (cétones aliphatiques). Un problème l'obsédait : « *Comment relier sur une famille homogène les changements de positions relatives des atomes aux évolutions de comportement mesurées expérimentalement ?* » La position relative des atomes, c'était la topologie d'une molécule et le problème se ramenait aussi à expliciter et quantifier cette grandeur. Il le fit d'une façon originale, avec son intuition de chimiste, en veillant à ce que tous les concepts introduits soient certes définis rigoureusement pour servir de support à des traitements automatiques, mais également à ce qu'ils « parlent » aux chimistes afin que ces derniers puissent se les approprier et les utiliser dans leur raisonnement. Dès 1966, il s'assigne comme ambition de relever le défi de concevoir un système qui tende vers l'idéal : « *Les propriétés exigées d'un système idéal paraissent en effet relativement exclusives. Si ces solutions sont proposées pour l'indexation et la recherche documentaire, par contre la classification systématique et donc l'établissement aisé de corrélations « propriétés-structure » restent encore dans le domaine des recherches de base sinon des projets.* »

Tels étaient la vision et le défi. Il releva ce défi par la recherche fondamentale, l'enseignement supérieur et des travaux de recherche plus finalisés, qui conduisirent à des transferts technologiques majeurs.

La recherche fondamentale

La démarche de J.-E. Dubois lors de la conception du système DARC fut celle du chercheur fondamental qui ne se préoccupait pas d'applications concrètes mais cherchait à satisfaire sa curiosité et valider son intuition initiale. Il était très exigeant sur la définition et la cohérence des concepts, la rigueur des procédures et l'élégance des raisonnements. À partir de sa vision de chimiste, il conçut l'appréhension de la topologie d'une molécule à partir d'un foyer (un atome, une liaison, un groupe d'atomes caractéristiques d'une série de composés) et de l'environnement – comprenant le reste de la molécule – généré selon des étapes logiques et non équivoques. Ce fut un apport conceptuel majeur car les systèmes de notation existant alors décrivaient des molécules isolées, par une fragmentation analytique. Ainsi, le processus de génération induisait un ordre sur le graphe représentatif de l'environnement et permettait également d'ordonner les éléments d'une population de composés. Il systématisait en particulier les travaux qu'il avait conduits antérieurement pour définir de façon non conventionnelle des plans de synthèse de cétones aliphatiques.

Les différentes méthodes de corrélation mises en œuvre pour relier la structure d'une molécule à son comportement à travers des relations dites de « topologie-information » relèvent également d'une démarche de recherche fondamentale. La topologie est considérée et explicitée comme une véritable variable, corrélée à l'évolution du comportement d'une série de composés. L'idée de discrétiser sur chaque site de l'environnement ordonné une valeur de comportement a permis, par la prise en compte implicite de multiples interactions, d'établir des corrélations dans des domaines très variés (propriétés physiques, réactivité, spectroscopie, activité pharmacologique, radioprotection…).

L'une de ses préoccupations constantes fut l'appréhension globale d'une population de composés chimiques afin de repérer un composé sans ambiguïté et d'exprimer son environnement au sein de cette population, c'est-à-dire compter et énumérer ses voisins et évaluer sa distance avec chacun d'eux. Pour cela, il importait d'organiser la population et donc d'établir des relations conceptuelles entre des objets réels. Aussi introduisit-il le concept très riche d'hyperstructures qui exprime l'organisation formelle d'un ensemble d'objets, eux-mêmes structurés par la création de relations entre ces objets. Ce thème de l'organisation d'une population chimique transparaissait dès les années 50 dans ses premiers travaux sur la synthèse des cétones où il se préoccupait de rationaliser et de visualiser l'ensemble de cette population, et ce fut l'une de ses préoccupations constantes jusqu'à la fin de sa vie.

De la recherche fondamentale à la recherche finalisée

De la description d'une molécule à partir d'un foyer et de la génération progressive de son environnement résultait une codification topologique originale de la molécule, le code

DARC qui décrivait toutes ses caractéristiques (cycles, stéréochimie) [2]. Une telle représentation était au cœur des préoccupations du domaine qualifié alors de « documentation chimique ». Chaque semaine, 8 000 molécules nouvelles apparaissaient dans la littérature scientifique. À l'époque, le Chemical Abstracts Service (CAS) avait le monopole de la collecte et de l'enregistrement de ces informations, il utilisait pour cela un code dérivé de la représentation d'un graphe par une matrice de connectivité. D'autres institutions utilisaient des codes fragmentaires qui représentaient une molécule comme un agencement de fragments identifiés. Deux problèmes étaient au cœur de la documentation chimique et s'inscrivaient dans un contexte mathématique plus large.

Le premier était de décider l'isomorphisme de deux graphes, c'est-à-dire dans le contexte de la documentation chimique, de décider qu'une molécule était réellement nouvelle au sein de la base de données et n'avait pas été déjà enregistrée. Pour cela, il convenait de définir une codification standard, unique de la molécule, le problème de l'isomorphisme se ramenant à la comparaison des codes.

Le deuxième problème relevait de décider de l'homorphisme de deux graphes, c'est-à-dire de savoir si une molécule contenait une sous-structure particulière. La résolution de ce problème était essentielle. Elle permettait d'exploiter intelligemment une base de données chimique en répondant à des questions du type « *Quels sont les composés de la base de données qui possèdent telle ou telle sous-structure ?* » Pour cela, on recourt à des écrans structuraux ou à des filtres contenus dans une molécule qui permettent d'accélérer la recherche en décidant très rapidement qu'une molécule ne répond pas à la question car elle ne contient pas un écran appartenant à la sous-structure.

La recherche se ramène à la génération des écrans associés à la sous-structure d'interrogation et à l'élimination des composés de la base de données qui ne possèdent pas ces écrans. Un système d'écrans performants permet de filtrer considérablement la base de données et d'apporter des réponses pertinentes [3]. Les travaux de recherche plus finalisés avec un objectif précis permirent d'apporter des solutions élégantes et performantes qui firent du code DARC un code de référence. Un algorithme de génération de la codification standard unique fut conçu. Il repose sur la logique de l'appréhension de la molécule à partir d'un foyer, et de la génération ordonnée de l'environnement, et conduit ainsi à une description unique. L'algorithme de génération repose sur le choix, à partir d'une série de critères, d'un foyer monoatomique.

Quant aux écrans, c'était un atout majeur du système DARC pour la recherche de sous-structures car ils ne reposaient pas sur des choix fixés *a priori* mais étaient générés très simplement à partir du code de la molécule en considérant tour à tour chaque atome de connectivité supérieure à 3 comme un foyer et en décrivant un environnement limité à deux couches d'atomes à partir de ce foyer, dénommé « FREL » ou fragment d'environnement limité. Ce concept résulte directement des principes généraux et permet par le recouvrement des FREL d'appréhender la richesse et la diversité topologique d'une molécule, des cycles notamment. La simplicité du système d'écrans et les algorithmes de conversion mis en œuvre pour assurer la convertibilité avec d'autres codes, notamment celui utilisé par les CAS, est à l'origine des développements industriels que je mentionnerai plus loin [4].

Je citerais Bill Milne : « *In 1966, Jacques-Emile Dubois published the first of a number of papers on what became the DARC system, truly the first structure search system. The software was solid, grounded in chemical graph theory, and was ingenious and clever*. »

Parmi les recherches finalisées, j'évoquerais celle conduite dans le domaine de la communication que l'on qualifiait à l'époque d'homme-machine, aujourd'hui totalement banalisé, dont J.-E. Dubois fut l'un des pionniers dans le domaine de la chimie.

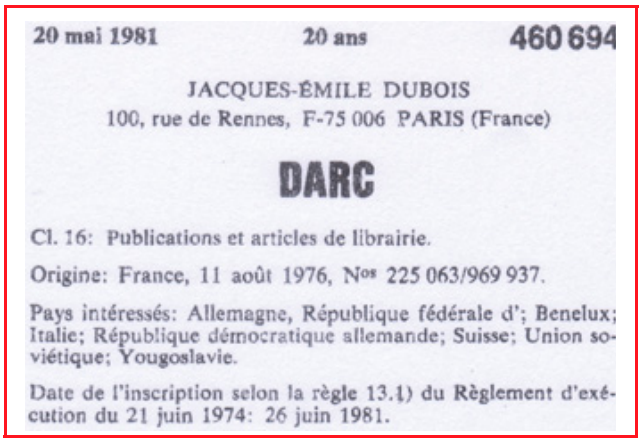
Tout d'abord, la saisie de la molécule : afin de disposer d'un système de codification automatique pour cette dernière, deux dispositifs furent développés. Le premier, conçu en coopération avec le LIMSI (Laboratoire d'informatique pour la mécanique et les sciences de l'ingénieur) de la Faculté des sciences d'Orsay permettait de saisir la molécule en la dessinant sur une tablette analogique. Le second, conçu au laboratoire et breveté, consistait à éditer sur un écran la formule développée de la molécule et à mémoriser les opérations d'édition sur une bande magnétique, afin de pouvoir les exploiter en différé sur un ordinateur pour ensuite générer automatiquement le code.

Dans le domaine de la visualisation sur console graphique, ce fut une série de recherches conduites sur une quinzaine d'années donnant lieu à de nombreuses thèses, démarrée en coopération avec le département d'Informatique de l'Université de Grenoble. Ce département était, à l'époque, le pionnier de l'informatique graphique dans notre pays. Il s'agissait, à partir de la codification DARC, d'éditer de la façon la plus réaliste possible la molécule (2D) sur un écran, la difficulté résidant dans le fait qu'il fallait détecter et représenter harmonieusement toutes les caractéristiques topologiques qui ont une signification visuelle, les cycles par exemple. Ensuite avec les évolutions technologiques, nous avons abordé la visualisation 3D des molécules pour en faire un véritable support de conception pour les chimistes. Ces travaux ont valu à J.-E. Dubois de siéger au Comité de rédaction de *Visual Computer*, référence internationale dans le domaine de l'infographie, et ont été publiés en 1985 dans un article de ce même journal en collaboration avec J. Weber de l'Université de Genève au titre fort évocateur : « Chemical ideograms and molecular computer graphics » [5].

Enfin, parmi les recherches finalisées qui ont conduit à des résultats intéressants, je citerais les travaux entrepris dès 1972 sur les formules dites « de Markush ». Ces formules sont utilisées depuis fort longtemps dans l'industrie chimique et pharmaceutique lors de dépôts de brevets car elles permettent par exemple de délimiter l'ensemble des entités chimiques que couvre une méthode de synthèse. Les concepts à la base du code DARC, foyer et environnement, se sont révélés particulièrement adaptés à la description des formules de Markush bien que la mise en œuvre ne fut pas simple car il convenait de décrire l'environnement d'une façon générique et de concevoir des algorithmes permettant de répondre à des interrogations comme : « *Telle entité ou telle famille d'entités se situe-t-elle dans le champ délimité par une formule de Markush ?* » Le traitement des formules de Markush par le DARC est encore utilisé aujourd'hui par l'Office Européen des Brevets.

Alors que les travaux de recherche fondamentale furent conduits par J.-E. Dubois au sein du laboratoire sur ses ressources internes, les travaux de recherche finalisés que je viens d'évoquer le furent avec l'appui de la Délégation à la Recherche Scientifique et Technique qui apportait des moyens de financement souples avec une grande liberté de manœuvre pour prendre des initiatives, saisir très rapidement des opportunités, associer à notre équipe – par

exemple en décidant quasiment dans la journée son recrutement pour une période limitée – un post-doc de passage à Paris en provenance de Stanford ou de Cornell. Ceci peut laisser rêveur les directeurs des laboratoires qui, engagés aujourd'hui dans la compétition scientifique internationale, doivent remplir en plusieurs exemplaires de nombreux formulaires et fiches diverses destinés à permettre à des responsables administratifs et de comités d'experts hexagonaux d'apprécier, en quelques mois, de la pertinence de la demande !



20 mai 1981 20 ans 460 694

JACQUES-ÉMILE DUBOIS
100, rue de Rennes, F-75 006 PARIS (France)

DARC

Cl. 16: Publications et articles de librairie.

Origine: France, 11 août 1976, Nos 225 063/969 937.

Pays intéressés: Allemagne, République fédérale d'; Benelux; Italie; République démocratique allemande; Suisse; Union soviétique; Yougoslavie.

Date de l'inscription selon la règle 13.1) du Règlement d'exécution du 21 juin 1974: 26 juin 1981.

Le brevet DARC.

De la recherche fondamentale et de l'enseignement supérieur

Les travaux conduits par J.-E. Dubois qui ont donné naissance au système DARC illustrent combien sans recherche fondamentale il n'y a pas d'enseignement supérieur digne de ce nom. J.-E. Dubois était profondément universitaire, chercheur reconnu internationalement ; il se préoccupait en permanence du transfert des avancées les plus récentes de ses travaux vers l'enseignement. Il l'avait fait dès 1957 où, jeune professeur chargé d'un enseignement de premier cycle à la Faculté des sciences de Paris, il délivrait un enseignement entièrement nouveau sur les mécanismes réactionnels. Il le fit pour les cinétiques rapides. Il le fit évidemment pour les travaux liés à la conception et au développement du système DARC. D'abord par les thèses : de nombreuses thèses d'État, d'ingénieurs docteurs, de troisième cycle furent soutenues, qui couvraient l'ensemble des domaines que j'ai évoqués. Ces docteurs, grâce à la diversité des problèmes traités, aux compétences acquises, aux échanges multidisciplinaires au sein d'un groupe dont la moyenne d'âge était inférieure à 25 ans, ont poursuivi ensuite leur carrière à des postes de responsabilité au sein de l'Université ou dans l'industrie, tant en France qu'à l'étranger.

La liaison étroite recherche fondamentale-enseignement supérieur donnera naissance dès 1971, au sein de l'Université naissante de Paris 7, à une série d'enseignements très originaux. Ainsi fut créé dans le cadre de la maîtrise de chimie-physique un certificat d'informatique chimique qui présentait comment l'informatique – qui n'était pas encore enseignée dans certaines grandes écoles – permettait d'aborder des problèmes traditionnels de la chimie sous un angle nouveau (la dynamique chimique, la synthèse, l'apport de la théorie des graphes, les systèmes d'information…).

J.-E. Dubois présenta cette démarche « Informatics and new concepts in Chemistry » au congrès international « Computers in Education » en 1975. Cette initiative se prolongera ensuite par un DEA entièrement dédié à l'informatique chimique ainsi qu'un DEA beaucoup plus large consacré à l'information scientifique et technique [6].

Cette liaison recherche fondamentale-enseignement supérieur a d'autant plus pu donner toute sa mesure qu'elle s'inscrivait dans le contexte de la création de l'Université Paris 7, présidée par le talentueux Michel Alliot, qui a encouragé toutes ses initiatives et rendu possible leur réalisation. Il est vrai aussi que le Conseil scientifique de cette université, dans cette phase de création, était un véritable « dream team scientifique », largement ouvert à des démarches innovantes, avec pour meneur de jeu Claude Allègre, et comptant parmi ses membres Pierre Aigrain, Jean Bernard, François Bruhat, Jean-Jacques Bernier, Antoine Culioli, Françoise Gaillard et Jacques-Émile Dubois.

De la recherche fondamentale au transfert technologique

L'ensemble de ces travaux de recherche conduisirent, dès le début des années 70, à des applications de nature industrielle qui constituaient l'armature d'un véritable système d'information chimique dont le DARC fut rapidement une référence internationale.

Pour cela et face à la carence de l'industrie chimique nationale, J.-E. Dubois créa ou suscita la création de structures de transfert, l'ARDIC (Association pour la Recherche et le Développement de l'Informatique Chimique) et le CNIC (Centre National de l'Information Chimique). Grâce à cette action, il put démontrer la faisabilité opérationnelle du système et comme en témoignait dans sa conclusion Michaël F. Lynch lors de la conférence de 2002 sur « The history and heritage of scientific and technological information systems » : « *In France, Jacques-Émile Dubois had already been working on file structures for substructure search on direct access rather than serial systems – the DARC system, using the notion of the FREL. This was an astonishing accomplishment; the DARC system operated on a single mainframe rather than on an array of minicomputers as used later by CAS. As part of international arrangements between countries, CAS was providing registry system files to various countries. The French mounted a system that rapidly became successful with industry around the globe…* » Michaël F. Lynch note en particulier : « *Eurecas, an early subset of the CAS registry system, which was inaugurated on the DARC system in 1978 and which anticipated the advent of CAS on line by several years had a huge impact, besides being an outright imaginative tour de force in computational terms. Its appeal to industry was immediate and widespread. Once again, when the Markush DARC system was introduced at a later date, it was equally valued*. »

Le système DARC n'était pas seulement un système novateur : c'était une arme politique. Il constituait le pivot de la coopération française avec le CAS, dans la mesure où, étant plus performant, notamment plus rapide, plus précis et offrant des possibilités plus vastes que le système du CAS, il donnait à la France la possibilité d'apporter une réelle valeur ajoutée aux bases de données du CAS. Ainsi, sous l'influence de J.-E. Dubois, le Centre National de l'Information Chimique (CNIC) fut constitué en 1979 à partir de l'ARDIC, basée à l'Université Paris 7. Le CNIC reçut mission de coordonner l'effort national pour le développement de banques de données complémentaires à CAS ainsi que de promouvoir le

système DARC « *afin de ne pas être un simple client…* ». Dans la pensée de J.-E. Dubois, sans doute n'y avait-il même jamais eu la notion d'un rapport fournisseur-client, mais plutôt d'emblée celle d'un partenariat. Conscient que le DARC, système autrement souple et performant que ce qui existait jusqu'alors, permettrait d'exploiter de manière beaucoup plus riche les bases de données du CAS, il pensait qu'il était de leur intérêt mutuel de s'associer, ce qui, vu les tailles respectives des organisations, a pu à l'époque paraître à certains relever de la métaphore de David et Goliath. Et de fait, le CNIC devint effectivement partenaire de CAS pendant plusieurs années.

Il est clair, pour tous ceux qui ont eu le privilège de participer à cette aventure, que J.-E. Dubois donna toute la mesure de son talent. Pour obtenir d'institutions américaines aussi prestigieuses et dominantes que le CAS la mise à disposition de leur base de données pour les exploiter d'une façon plus performante qu'elles, la négociation n'était pas facile. Ses armes ? La force de sa conviction, la supériorité scientifique et technique du système DARC, son intelligence des hommes et des situations, son humour et accessoirement sa maîtrise parfaite de l'argot américain qui, dans des situations difficiles, lui permettait de déstabiliser, avec esprit et un non-conformisme profondément latin, des interlocuteurs plus prisonniers des procédures et de la lourdeur de grandes institutions.

En 1981, les pouvoirs publics décidèrent de transférer la commercialisation et l'évolution du produit Eurecas et autres produits élaborés à Questel, filiale de Telesystemes, elle-même composante de ce qu'était la nébuleuse constituée autour de la Direction Générale des Télécommunications. Avec vingt-cinq ans de recul, on ne peut pas dire que ce transfert au sein d'une entreprise très hexagonale fut un succès et si le DARC perdure encore, c'est au sein de l'Office Européen des Brevets et par les concepts introduits qui ont été exploités fort intelligemment par des systèmes concurrents. Comme l'observe Johnny Gasteiger : « *It is sad that the database and retrieval systems based on the DARC system as put up for commercial use by Telesystem-Questel were eventually put on hold and discontinued. The reasons were clearly not in the performance of the system but have to be due to non-scientific considerations.* »

Jacques-Émile Dubois était un universitaire au plein sens du terme. Homme de conviction, comme en témoigne son engagement au sein de la Résistance, c'était un esprit original et passionné par la science, qui cherchait en permanence à étendre les frontières du savoir en brisant les barrières entre les disciplines. Par son charisme, il communiquait cette passion et son enthousiasme à ses jeunes élèves qu'il savait motiver, écouter et conforter. De 1947 à la fin de sa vie, il resta un « jeune leader » scientifique.

Références

[1] Dubois J.-E., Laurent D., Vieillard H., Système de documentation et d'automatisation des recherches de corrélation (DARC). Principes généraux, *C.R. Acad.Sci.*, **1966**, *263C*, p. 764.

[2] Dubois J.-E., *DARC System in Chemistry Computer Representation and Manipulation of Chemical Information*, John Wiley, New York, **1974**, p. 239.

[3] Attias R., DARC substitute search system: a new approach to chemical information, *J. Chem. Inf. Computer Sci.*, **1983**, *23(3)*, p. 102.

[4] Sobel Y., Dagane I., Carabédian M., Dubois J.-E. Specific features of scientific data banks, *Proceedings of the 9th International CODATA Conference*, Glaeser, Jérusalem, **1985**.

[5] Dubois J.-E., Laurent D., Weber J., Chemical ideograms and molecular computer graphics, *The Visual Computer*, Springer Verlag, **1985**, *1*, p. 49.

[6] Dubois J.-E., Laurent D., *Informatics and new concepts in chemistry. Computer in Education,* O. Lecarne, R. Lewis (eds), IFIP, North Holland Publishing Company, **1975**.

Daniel Laurent
est professeur d'informatique et fondateur de l'Université de Marne-la-Vallée*.

* Université de Marne-la-Vallée, Cité Descartes, 5 bd Descartes, Champs sur Marne, 77454 Marne-la-Vallée Cedex 2.
Courriel : daniel.laurent@univ-mlv.fr