Utilisation des techniques avancées

Bases de données relationnelles, catalogues en ligne sur www., logiciels expert de contrôle qualité pour l'archivage, la gestion et la diffusion des données océanographiques

Catherine Maillard

Introduction

L'Ifremer a développé plusieurs bases de données d'intérêt général, pour les données collectées lors des campagnes scientifiques, pour le suivi de l'environnement littoral, pour les ressources halieutiques, et pour les données des satellites ERS. Cet article concerne le système d'archivage mis en place au Sismer pour gérer les données des campagnes océanographiques ainsi que les données de plusieurs programmes scientifiques internationaux. Il fait appel aux techniques informatiques avancées de communication : bases de données relationnelles pouvant être interrogées directement sur www. Il inclut également des outils expert de contrôle qualité pour assurer que les données issues de sources différentes sont cohérentes et compatibles entre elles.

Avant de décrire ces outils, une brève description du centre de données et des données gérées est présentée.

Données océanographiques du centre Sismer

L'archivage et la diffusion des données océanographiques françaises datent de 1968 et couvre différents domaines de la physique et de la chimie marines, de la géophysique mesurée en route ainsi que de l'information générale. Sismer (Systèmes d'informations scientifiques pour la mer) a été créé en 1990 pour reprendre ce service dans un contexte de volume de données grandissant, de besoins scientifiques et techniques évoluant rapidement et de l'apparition de nouvelles technologies informatiques.

Il est le centre national de données océanographiques (CNDO) désigné pour la France par la commission océanographique intergouvernementale (COI) de l'Unesco, dans le cadre du programme international oceanographic Data and Information Exchange (IODE), la suite de l'ancien bureau national des données océaniques (BNDO). Il est également le centre d'archivage de plusieurs projets internationaux.

Les données et information (méta-données) gérées au Sismer se composent de :

1. Banque nationale de données océanographiques :
- Catalogue des campagnes océanographiques (ROSCOP/Cruises Summary Reports) : plus de 4 700 résumés de campagnes de tous les organismes de recherche civile, disponibles sur www avec plans de route ;

- inventaire des données marines des laboratoires (EDMED/European Directory of Marine Environmental Data) ;
- banque de géophysique marine ;
- banque de physique et chimie marines.

2. Données et information de programmes nationaux et internationaux.

3. Données de référence publiées par d'autres centres de données.

Quelques informations sont données ci-dessous sur les principaux programmes. On peut constater que les données gérées se caractérisent par leur hétérogénéité, tant par leurs types que leurs volumes : méta-données textuelles et graphiques pour les catalogues des campagnes et des bases de données, données numériques multi-paramètres pour les prélèvements biochimiques, gros volumes de données pour les données d'imagerie acoustique collectées en route.

Banque nationale de géophysique marine

La banque de géophysique marine contient les mesures en route de :
- bathymétrie verticale : 340 campagnes ;
- bathymétrie multifaisceaux : 268 campagnes ;
- gravimétrie : 202 campagnes ;
- magnétisme : 245 campagnes ;
- imagerie : 38 campagnes.

À l'issue de chaque campagne, un ou plusieurs gros fichiers de données de même type est transmis. La couverture est mondiale (fig 1.).

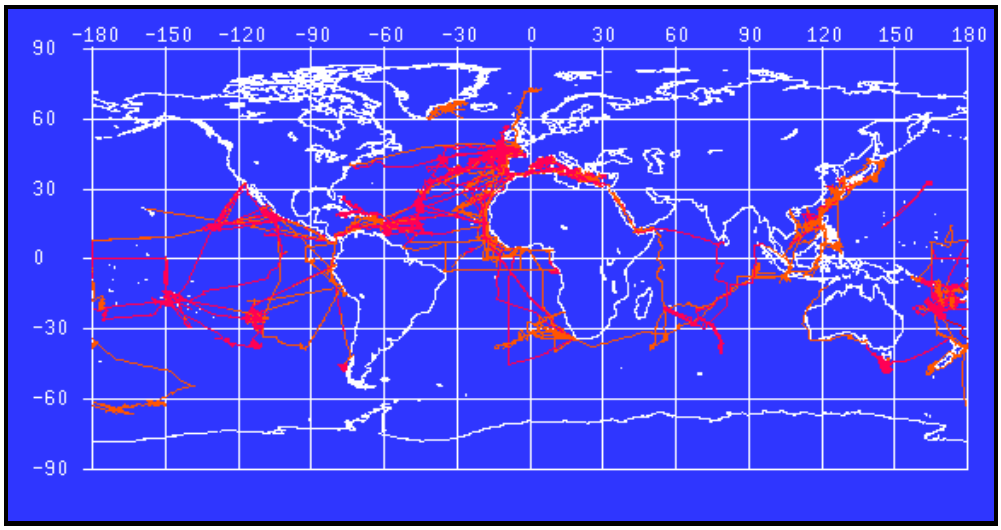


Figure 1 - Répartition géographique des données de bathymétrie multifaisceaux françaises

Banque nationale de physique et chimie marines

La banque de physique et chimie marines contient de longues séries historiques (Fig. 2) de :
- profils verticaux CTD 15 551
- profils Bouteilles (prélèvements chimiques) 33 143

- séries temporelles de Courantomètres 1 738
- séries temporelles de Thermistances 105

Ces données sont transmises par les laboratoires à divers formats. Les volumes sont variables ainsi que le nombre de paramètres mesurés (jusqu'à 20).

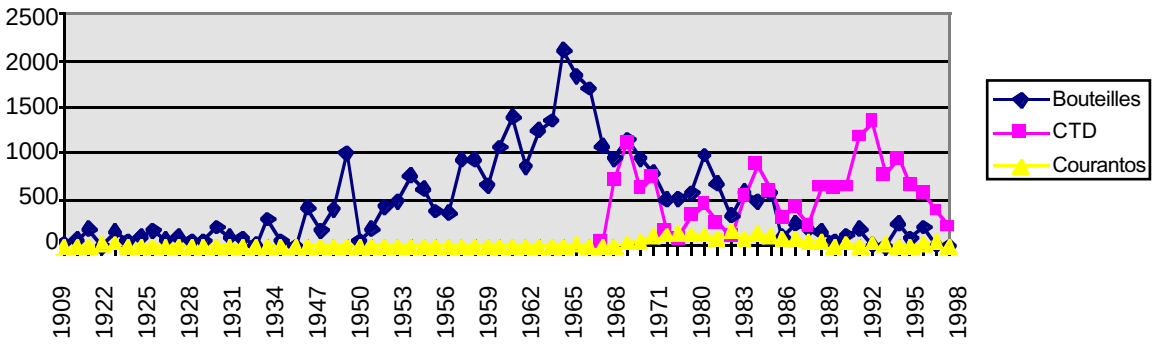


Figure 2 - Chronologie des données de la banque de physique et chimie marines.

Base de données de subsurface TOGA/WOCE/CLIVAR-CORIOLIS

Cette base de données a été développée pour les programmes internationaux TOGA et WOCE dans le cadre d'un programme conjoint Orstom-Ifremer. Elle contient des profils de température basse résolution collectés depuis 1985 sur une couverture mondiale, (Fig. 3), soit 480 000 profils de température.

Elle est maintenant utlisée pour gérer les données temps différé du programme européen MAST/MFSPP (Mediterranean Forecasting System Pilot Project) et les données Temps Réel de température et salinité du projet national CORIOLIS.
Les données de cette base sont directement accesssibles sur internet, en particulier les données temps réel.

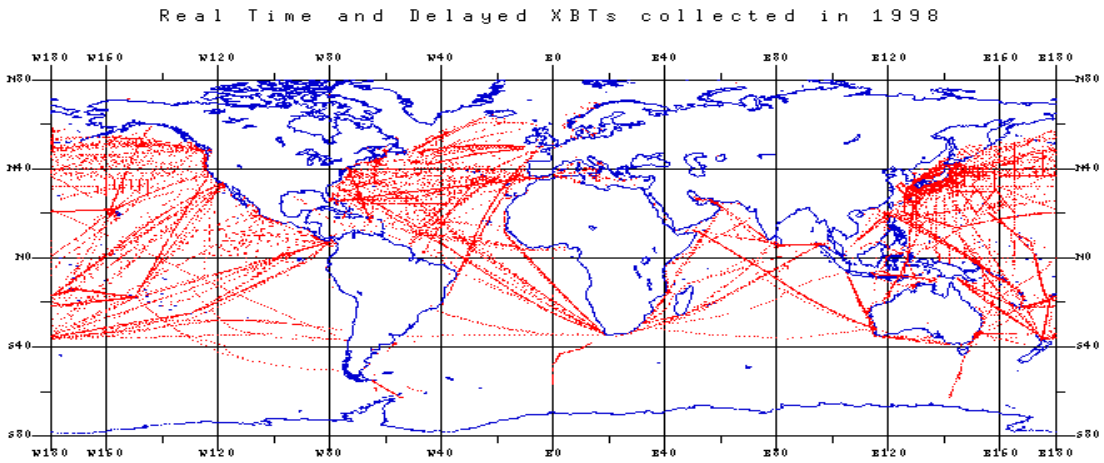


Figure 3 - Répartition géographique des profils de température TOGA/WOCE archivés en 1998.

Bases de Meta-données-catalogues

Les bases de données « résumés de campagnes » et « résumés de bases de données marines » sont publiées sur le site www du Sismer. Elles sont également accessibles à partir du catalogue européen Sea-Search (projet MAST/EURONODIM) (Fig. 4).

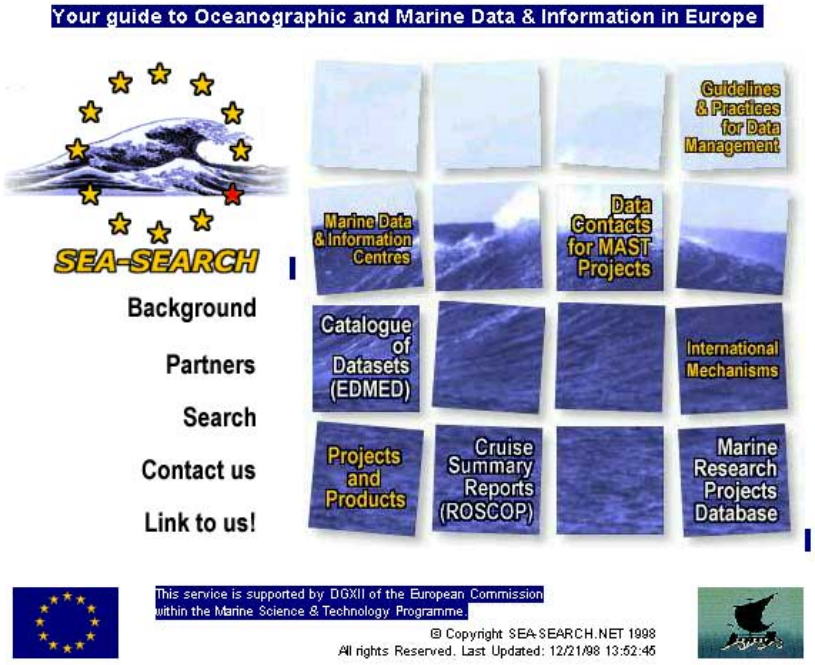


Figure 4 - Site www Sea-Search.

Bases de données Méditerranée

Sismer contribue à trois programmes européens MAST en Méditerranée :

- MFSPP - Mediterranean Forecasting System Pilot Project - Centre d'archivage temps différé (MAS3-CT98-0171) ;
- MTP II- MATER - Mediterranean Targeted Project II, MAss Transfer and Ecosystem Response (MAS3-CT96-0051) En collaboration avec le CNDO Grec et un centre de données italien, gestion des données du principal programme scientifique présent - 55 laboratoires, plus de 400 paramètres identifiés ;
- MEDAR/MEDATLAS II - Mediterranean Data Archaeology and Rescue (MAS3-CT98-0174 & ERBIC20-CT98-0103) : en collaboration avec 20 partenaires européens et non-européens, développement d'une base de données marines aussi complète que possible pour les paramètres hydrologiques de base : température, salinité, oxygène, nitrate, nitrite, NH4, azote total, phosphate, phosphore total, silicate, H_2S, pH, alcalinité, chlorophylle. Ces paramètres étant considérés comme essentiels pour la modélisation des écosystèmes, le suivi climatique ainsi que pour de nombreux besoins techniques.

La base de données doit inclure des données historiques aussi bien que récentes et les trois programmes sont fortement imbriqués : les données historiques servent à qualifier les données récentes et permettent de suivre des évolutions possibles. Les expériences récentes permettent de développer de nouvelles technologies et de nouveaux types d'observations.

Circulation des données

Sismer acquiert les données et méta-données, soit directement des navires, soit des laboratoires scientifiques. Ces données sont validées par les laboratoires scientifiques ou les responsables des capteurs.

Afin d'assurer la cohérence et la compatibilité de données provenant de sources diverses, Sismer transcode les données à un format unique pour chaque type de données et opère des contrôle de qualité.

Les données et méta-données sont ensuite sauvegardées dans le système d'archivage et diffusées selon le schéma donné fig. 5.

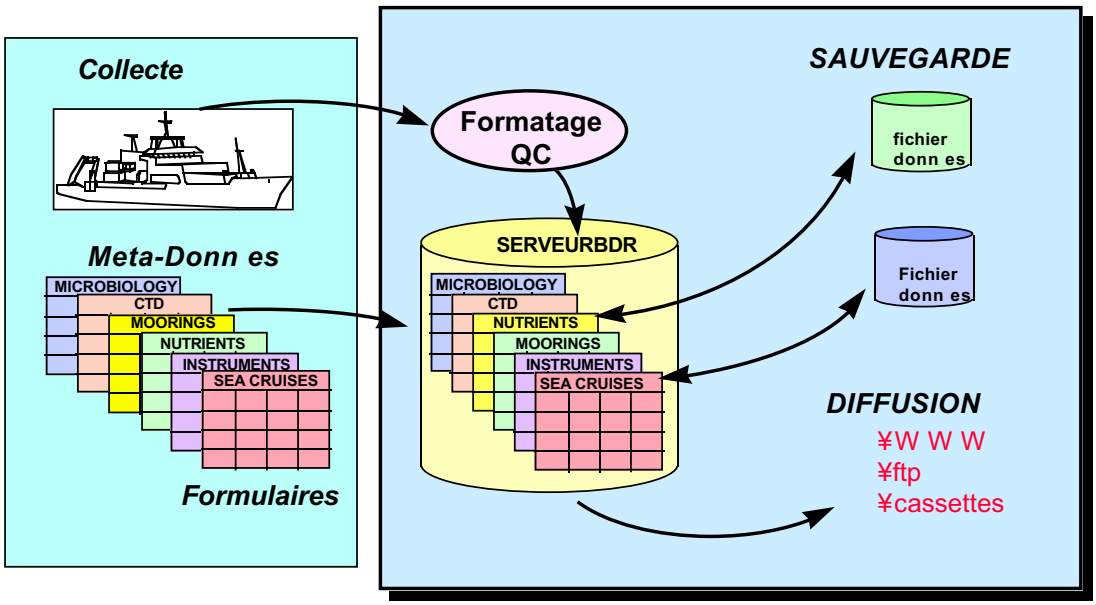


Figure 5 - Circulation des données des campagnes océanographiques.

Les principaux formats d'archivage des fichiers sont de l'ASCII autodescriptif (MEDATLAS) et du binaire (NetCDF).

Techniques et outils logiciels

Les outils logiciels utilisés sont des systèmes de gestion de bases de données relationnelles du commerce munis d'interfaces de chargement et d'édition (papier et www) développées par Ifremer et des outils de traitement comme les systèmes expert de contrôle qualité.

Le serveur Sismer

Le système matériel et logiciel s'appuie sur une base de données relationnelle sous ORACLE V8, le serveur Sismer, implanté sur un serveur centralisé de l'Ifremer. Dans le serveur sont chargées toutes les méta-données dont comptes-rendus de campagne et résumés de fichiers,

ainsi que les adresses logiques des fichiers de données. Des outils logiciels ont été développés autour du serveur :

- des logiciels de chargement permettent soit la saisie manuelle de l'information, soit le chargement automatique des résumés d'un ensemble de fichiers ;
- des logiciels d'édition de rapports comme le Recueil annuel des campagnes ;
- un logiciel d'extraction en ligne (fig. 6) permet d'accéder aux informations et aux fichiers de données en fonction de plusieurs critères : campagne, type de données, date, zone géographique.

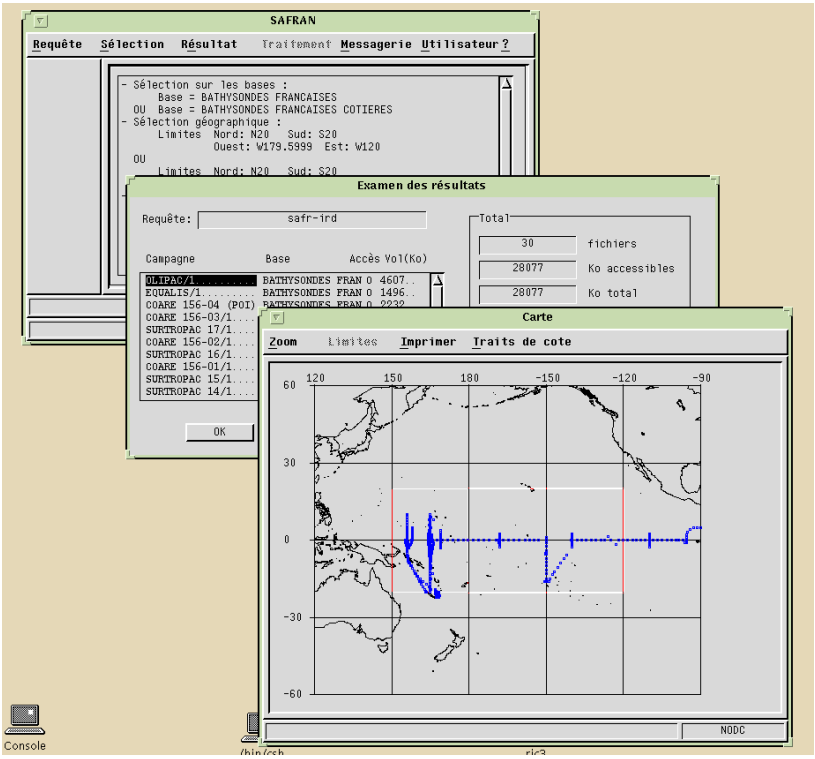


Figure 6 - Interface de recherche et d'extraction de données.

Le serveur Internet www

Le système comporte également un serveur internet (www et ftp) et des outils logiciels permettant d'éditer dynamiquement le contenu de la base ORACLE, en particulier les résumés de comptes rendus de campagne (Fig. 7).

Il permet aussi de faire directement des requêtes de données. L'accès en ligne aux données est toutefois soumis aux règles habituelles de diffusion avec vérification de la confidentialité scientifique, économique, militaire, ces dernières n'étant accessibles qu'au cas par cas *via* un opérateur.

Bien que l'accès direct aux données soit possible, sur les 170 requêtes reçues en 1998, la plupart des utilisateurs ont préféré faire appel à l'aide d'un opérateur.

Figure 7 - Home page du catalogue www des campagnes océanographiques françaises.

Les outils de diffusion

En dehors de l'accès aux données sur requête, des produits de synthèse sont également développés et publiés sur CDrom pour faciliter la diffusion. Les Cdroms sont alors munis d'outils logiciels spécifiques pour sélectionner et visualiser les données du disque. Il faut être attentif au fait qu'un tel type de publication ne suffit pas à garantir la pérennité des données car la durée de vie des supports et surtout des lecteurs s'est toujours avérée limitée dans le temps.

Le système de sauvegarde

Sismer s'appuie sur le système d'archivage centralisé de l'Ifremer avec des sauvegardes quotidiennes automatiques, ainsi que des recopies de sauvegarde sur un deuxième site. Chaque fois que la technologie évolue, l'ensemble des archives est basculé sur le nouveau système.

Contrôles de qualité

La cohérence et la compatibilité de données issues de différentes sources n'est pas évidente. De plus, les manipulations de données opérées au centre sont des sources potentielles d'erreur. Il est donc important de vérifier les données avant l'archivage. Les contrôles qui sont opérés au Sismer suivent les normes internationales édictées par Unesco/COI/IODE (International Oceanographic Data & Information Exchange), le CIEM (Conseil international pour l'exploration de la mer) et MAST (European Marine Science & Technology Programme).

Les contrôles de qualité (QC) comprennent les trois étapes suivantes, quel que soit le type de donnée :

QC0 : contrôle du format et de la complétude de l'information, de l'utilisation d'unités du système international etc. ;

QC1 : contrôle de la position (y compris verticale pour les séries temporelles au point fixe) et de la date (Fig. 7) ;
QC2 : contrôle des valeurs d'observation (Fig. 8).

Ces contrôles s'opérent automatiquement puis visuellement à l'aide de logiciels experts. En résultat, les données ne sont pas modifiées mais un indicateur ou flag de qualité est ajouté à chaque valeur numérique. Le fournisseur des données est contacté pour valider ou éliminer le données anormales, mais dans le cas des données anciennes où le retour aux sources n'est pas possible, le flag est conservé.

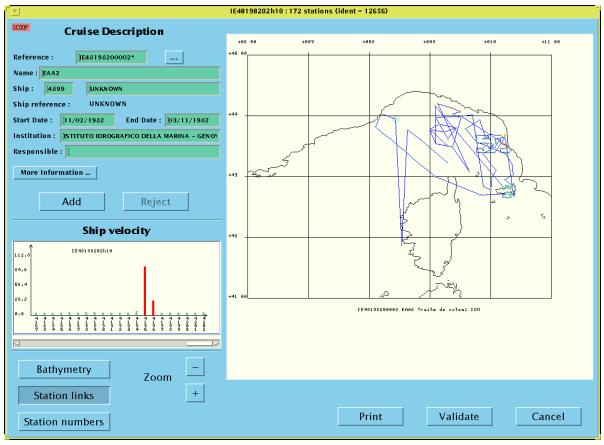


Figure 8 - Contrôle de la position géographique des observations.

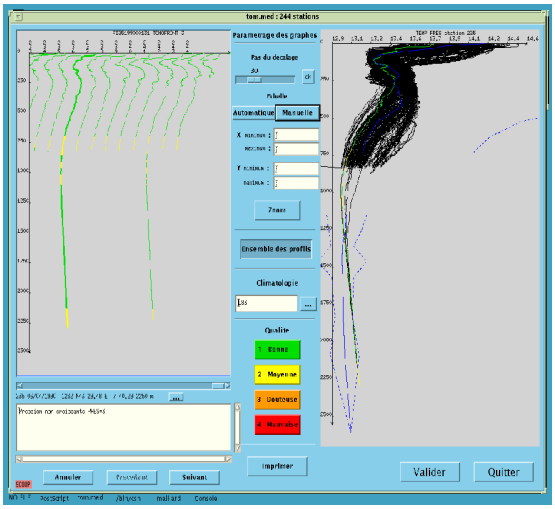


Figure 9 - Contrôle des observations.

Conclusion

Un tel système centralisé permet de faciliter l'accès aux données, d'assurer leur pérennité et un niveau minimum de qualité.

Faciliter l'accès aux données vise à valoriser l'effort de collecte auprès de la communauté scientifique et au-delà, c'est-à-dire dans le monde de l'éducation, de l'industrie et du grand public.

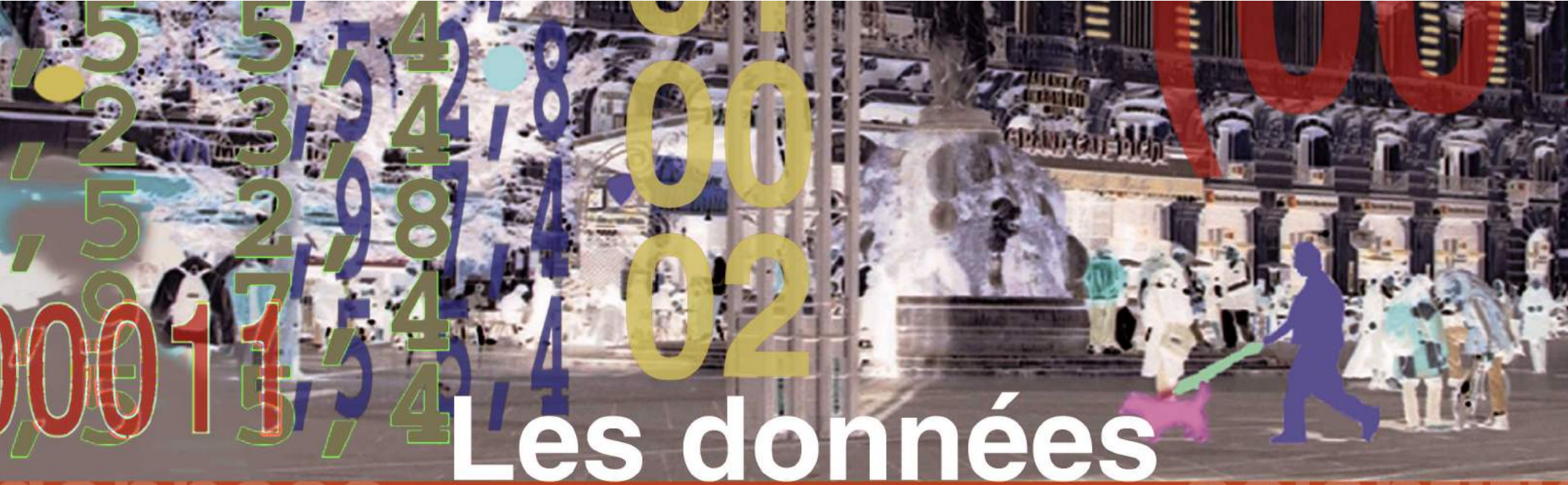
Assurer la pérennité permet d'économiser du temps navire et éviter de perdre les observations collectées dans un environnement variable qui ne peuvent plus jamais être refaites. Or la gestion des ressources marines, la surveillance de l'environnement et du changement climatique nécessitent le suivi de longues séries d'observations, en particulier :

- des modifications du contenu de la mer en oxygène dissous et sels nutritifs qui peuvent conduire à l'altération de la production primaire, de la biodiversité, développer des pathologies diverses dans les organismes vivants et finalement avoir un effet négatif sur l'aquaculture et la pêche ;
- de la température et de la salinité, qui couramment utilisées par les physiciens et les ingénieurs, sont également les indicateurs primaires du changement climatique.

Bien que validées à la source, toutes ces données doivent néanmoins subir un minimum de contrôle à l'archivage car des données mal documentées ou douteuses, ou encore des erreurs de manipulation les rendent inexploitables.

Enfin, ce qui concerne la centralisation de l'archivage, il faut souligner qu'elle reste nécessaire malgré l'évolution technologique. D'une part, les outils sophistiqués et coûteux, d'autre part, les compétences particulières nécessaires à la gestion de données ne sont pas à la portée de toutes les équipes scientifiques. Plus fondamentalement, ni la facilité d'accès aux données, ni la pérennité ne peuvent être assurées dans un système totalement dispersé. Des études récentes ont d'ailleurs montré qu'au bout de 10 ans, 30 % des données non archivées dans un centre de données sont perdues. Par ailleurs, l'émergence de l'océanographie opérationnelle est liée à l'existence de structures d'archivage et d'exploitation de type « opérationnel » (aval) difficiles à mettre en place dans des organismes de type « recherche » (amont).

Finalement on peut noter la prise de conscience de plus en plus forte de l'importance de l'information et des données dans la société, et il est à prévoir que les centres de données, comme autrefois les bibliothèques, continueront à se développer rapidement.



Les données scientifiques

Bases de progrès des connaissances

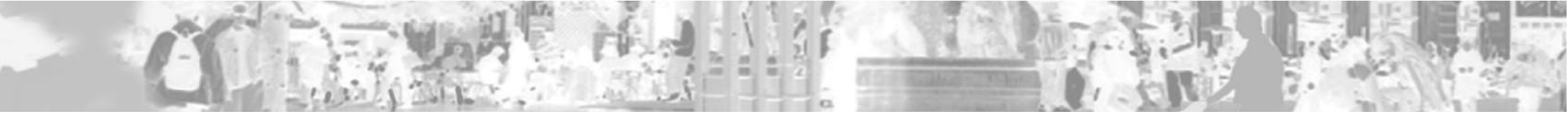
Séminaire tenu à l'IRD Ile-de France, Bondy les 4 et 5 mai 1999

Éditeurs scientifiques Jean-Michel Kornprobst, Marcel Raffy

IRD Éditions

sommaire adresses des auteurs

Sommaire



Adresse des auteurs

Jean-Luc **Boichard,** informaticien, Météo-France/Médias, BP 2102, 18, avenue E. Belin, 31401 Toulouse cedex 4.

Jeanne **Bossi**, secteur santé, CNIL, 21, rue St-Guillaume, 75007 Paris.
e-mail : jbossi@cnil.fr

Danièle **Boucon,** ingénieur CNES, 18, av. Edouard Belin, 34401 Toulouse cedex 4.

Eliane **Cubero-Castan**, informaticienne, Médias-France, BP 2102, 18, avenue Edouard Belin, 31401 Toulouse cedex 4.

Pierre **Chavance** IRD, BP 1386, Dakar, Sénégal.
e-mail : Pierre.Chavance@ird.sn

Pascal **Dallot**, assistant informatique, CEA/DAM, Analyse, surveillance, environnement, B.P. 12, 91680 Bruyères–le-Châtel.
e-mail : dallot@dase.bruyeres.cea.fr

Michel **Hoepffner**, hydrologue, IRD-Médias, BP 2102, 18, av. E. Belin, 31401 Toulouse cedex 4.
e-mail : Michel.Hoepffner@medias.cnes.fr

Claude **Huc**, ingénieur, département Valorisation et gestion des données spatiales, CNES, 18, av. Edouard Belin, 31401 Toulouse cedex 4.
e-mail : claude.huc@cnes.fr

Jean-Michel **Kornprobst**, professeur université de Nantes, vice-Président de la CS7, ISOMer, Laboratoire de chimie marine, BP 92208, 2, rue de la Houssinière, 44322 Nantes celex 3.
e-mail : jean-michel.kornprobst@wanadoo.fr

Sébastien **Lafargue,** juriste, Ifremer, Technopolis 40, 155, rue J.J. Rousseau, 92138 Issy-les-Moulineaux.
e-mail : Sebastien.lafargue@ifremer.fr

Francis **Laloë**, IRD, Halieutique et Écosystèmes Aquatiques, BP 5045, 34032 Montpellier cedex 1.
e-mail : laloe@mpl.ird.fr

Michel **Lang,** hydrologue, Cemagref, Division hydraulique, 3 bis, quai Chauveau, CP 220, 69009 Lyon cedex.
e-mail : michel.lang@cemagref.fr

Alain **Laponche**, ingénieur Sismer, Ifremer, centre de Brest, BP 70, 29280 Plouzané.

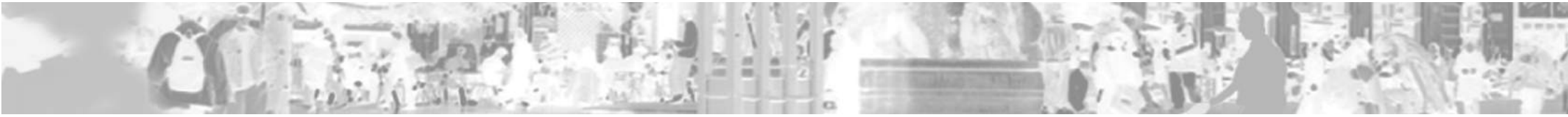
Didier **Lavoine**, ingénieur réseau, 2 bis, rue R. Le Ricollais, 44000 Nantes.

Fabrice **Lecornu**, ingénieur informaticien, Ifremer, centre de Brest, BP 70, 29280 Plouzané.
e-mail : Fabrice.Lecornu@ifremer.fr

Joseph **Le Foll**, informaticien, CEA/DSM/DAPNIA, CE Saclay, 91191 Gif-sur-Yvette cedex.
e-mail : lefoll@hep.saclay.cea.fr

François **Le Verge**, chef du service de la documentation, Ifremer, centre de Brest, BP 70, 29280 Plouzané.
e-mail : fleverge@ifremer.fr

Catherine **Mailllard**, ingénieur de recherche Ifremer, centre de Brest, BP 70, 29280 Plouzané.
e-mail : Catherine.Maillard@ifremer.fr



Bernard **Marx**, INPI, service DDI, 26 bis, rue de Saint-Pétersbourg, 75008 Paris.

Jean-Pierre **Muller**, pédologue, directeur général de l'IRD, 209-213, rue La Fayette 75480 Paris cedex 10.

Michel **Passouant**, statisticien Cirad, Campus International de Baillarguet, Bât. F, 34398 Montpellier cedex 4.
e-mail : michel.passouant@cirad.fr

Pierre **Peltre**, géographe, IRD, 32, avenue Henri-Varagnat, 93143 Bondy cedex.
e-mail : peltre@clarke.bondy.ird.fr

Marcel **Raffy**, professeur, université de Strasbourg, président de la CS7, ULP-CNRS, Parc d'innovation, 5, bd S. Brandt, 67400 Illkirch-Graffenstaden.

Armel **Rué**, ingénieur réseau, Ifremer, centre de Brest, BP 70, 29280 Plouzané.

Patrick **Séchet**, informaticien, IRD, 209-213, rue La Fayette, 75480 Paris cedex 10.
e-mail : sechet@paris.ird.fr

Éric **Simon**, directeur de recherche en informatique, Inria, BP 105, 78153 Le Chesnay.
e-mail : eric.simon@inria.fr

Françoise **Souyri**, directeur de recherche, MENRT-CSDR, 5, rue Descartes, Paris cedex 05.
e-mail : francoise.souyri@dr.education.gouv.fr

Dominique **Vuillaume**, économiste de la santé, Service du partenariat pour le Développement, Inserm, 101, rue de Tolbiac 75654 Paris cedex 13.
e-mail : vuillaume@tolbiac.inserm.fr

Philippe **Waldteufel,** climatologue, CNRS-IPSL, 10-12, avenue de l'Europe, 78140 Vélizy.
e-mail : Philippe.Waldteufel@ipsl.uvsq.fr