

"TIC et mer: nouveaux défis et solutions"

Compte-rendu de la conférence organisée par l'Ifremer et le Lab-STICC



CONFÉRENCE

TIC et mer: nouveaux défis et solutions
Les technologies de l'information au service de la recherche marine

Gérer des bases de données de plus en plus grandes et complexes
Partage et interactions de bases de données
Méthodes de fouille et d'analyse

9h45: ACCUEIL
10h00-16h00: PRÉSENTATIONS
MATHIAS HERBERTS, STÉPHANIE MAHÉVAS,
GUILLAUME MAZE, JOHN PUENTES, JEAN-FRANÇOIS
PIOLLÉ, THOMAS LOUBRIEU, GILBERT MAUDIRE,
PHILIPPE LENCA, ROMAN FABLET
16h00: TABLE RONDE AVEC RENÉ GARELLO

le 26 novembre 2013
Centre Ifremer de Brest
INSCRIPTION GRATUITE
MAIS OBLIGATOIRE
<http://wwz.ifremer.fr/bigdata>

Ifremer
Lab-STICC

"Ultimately, big data is more about attitude than tools; data-driven organizations look at big data as a solution, not a problem."

Roger Magoulas and Ben Lorica, from "Big Data: Technologies and Techniques for Large Scale Data", page 32.

La conférence s'est tenue le 26 novembre 2013 dans la salle Lucien Laubier du centre Ifremer de Brest, de 10h00 à 17h00. Cet événement avait trois objectifs. Le premier était de réunir le public local (Ifremer et technopôle) de chercheurs, ingénieurs et informaticiens en sciences marines et confrontés aux problématiques issues de ce que l'on appelle communément aujourd'hui le "Big Data". Le second objectif de la conférence était de fournir des informations aux participants pour les aider à mieux se situer et appréhender dans un cadre formalisé leurs problématiques spécifiques. Enfin, le troisième objectif était de réunir des intervenants sur ces problématiques et de contribuer à la fédération d'une communauté locale de spécialistes qui contribuent au développement et à la mise en oeuvre de solutions à ces problématiques. Le comité d'organisation¹ de la conférence a le sentiment que ces trois objectifs ont été atteints.

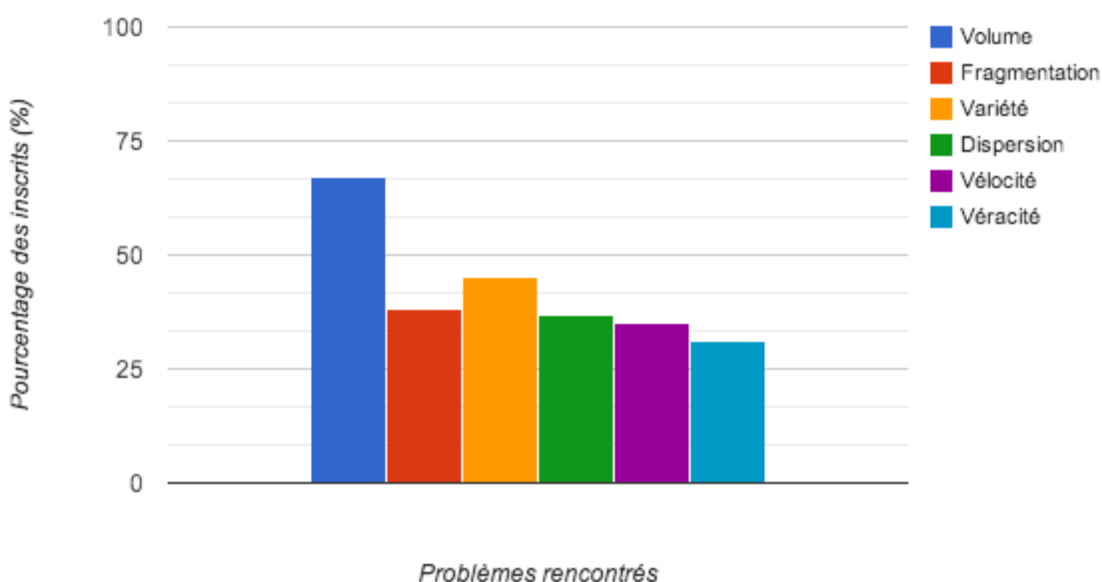
¹ Le comité d'organisation était constitué de (ordre alphabétique): J. Bonnel, B. Chapron, G. Coppin, F. Cudennec, J. Détoc, P. Dhaussy, P. Farcy, R. Garelo, M. Herberts, Y. Kermarrec, D. Le Brun, T. Le Toullec, P. Lenca, T. Loubrieu, G. Maze, H. Mercier, T. Odaka, J.F. Piollé, I. Puillat.

Il y a eu 91 inscrits via le formulaire web (voir la liste en annexe 1). Parmi ces inscrits, environ la moitié (47) était affiliée à Ifremer. L'autre moitié des participants était plus fragmentée. On pouvait compter: 7 inscrits de Télécom Bretagne, 7 du SHOM, 4 de l'ENSTA et 4 de l'UBO, puis 2 de chez Altran, idem de CLS, l'ECA Group, l'école Navale et Thalès. Enfin était inscrite 1 personne de "Code Lutin", 1 de Datalone, 1 de l'IRD, 1 de l'IRISA, 1 de l'université du Chili, 1 de Ouest Valorisation, 1 de Quiet Ocean et enfin 1 du Pôle Mer Bretagne. Cette distribution des inscrits indique un fort intérêt pour ces problématiques de la part des membres de l'Ifremer ainsi que de l'ensemble des acteurs académiques et institutionnels du Technopôle de Plouzané.



Retour sur les interventions du matin

La journée a commencé par une brève introduction de **Guillaume Maze** (chercheur Ifremer, animateur de la conférence et coordinateur du comité d'organisation) qui est revenu sur la composition du comité d'organisation puis sur l'analyse du sondage mis en place sur la page web des inscriptions à la conférence. Chaque nouvel inscrit se voyait demandé de préciser quelles étaient les problématiques auxquelles il était confronté. Les résultats sont représentés sur le diagramme en bar ci-dessous:



On peut constater que la problématique du volume de données est la plus citée par les inscrits et que toutes les autres problématiques semblent concerner une fraction plus homogène de l'audience. Notons que les discussions et commentaires des intervenants ont montré au cours de la journée que si la notion qualitative de "volume" était la plus problématique la plus courante, elle recouvrait quantitativement des valeurs très différentes suivant les disciplines et spécialités concernées.

Puis **Patrick Farcy** (direction scientifique de l'Ifremer) et **René Garello** (responsable du programme transverse TIC pour les océans -ICTO- au LabSTICC) ont fait une présentation des contextes institutionnels. P. Farcy indique que l'Ifremer stocke de plus en plus en données et que l'institut a besoin d'outils adaptés pour soutenir les recherches conduites avec ces données. En conséquence, la direction scientifique a disposé en 2013 et continuera de disposer en 2014 d'un budget pour soutenir des projets liés à cette thématique. Une école d'été est déjà prévue pour le mois de septembre 2014. R. Garello indique que cette thématique Big Data est très présente à Télécom Bretagne et que la complexité et l'étendu des problématiques qu'elle soulève a donc amené le Lab-STICC à mettre en place un axe transverse (ICTO) entre ses équipes de recherche. R. Garello précise également que l'IEEE vient de créer un journal dédié au thème du Big Data.

Après ces introductions, **Mathias Herberts** a pris la parole pour un exposé de 45 minutes intitulé "Défier un océan de données". M. Herberts est considéré comme un des pionniers du "Big Data" en France. Après avoir travaillé chez Google (sur le produit Big Table), il a implémenté de nouvelles solutions software et hardware au Crédit Mutuel Arkéa pour permettre à cette banque de faire face à leurs problématiques de type "Big Data". Aujourd'hui co-fondateur d'une startup (CityZen Data), il était une des personnes les mieux à même de pouvoir venir présenter et expliquer les problématiques que recouvrent le terme désormais trop utilisé parce que trop à la mode de "Big Data". Ce terme, introduit il y a plus de 10 ans par la compagnie américaine Gartner (cabinet de conseil et veille technologique) recouvrait initialement trois problématiques: volume, vitesse et variété. Le "Big Data" pouvait donc se définir initialement comme *l'art* de faire face (c'est-à-dire stocker et analyser par exemple) à une énorme quantité de données arrivant très rapidement et contenant une multitude de types d'information (comme des vidéos, du texte, des tableaux, etc ...). Très vite, les définitions se sont multipliées pour englober d'autres problématiques.



Aujourd'hui, Mathias nous a proposé la définition suivante du "Big Data":

Ensemble des problématiques liées à une telle quantité et diversité de données que l'on ne peut plus les gérer et les exploiter avec les méthodes et outils traditionnels.

Mathias Herberts détaillera ensuite pour une liste non exhaustive de ces problématiques qui ne sont pas que techniques (le stockage, l'analyse, la sécurité, l'organisation et les partenariats) les idées reçues, les écueils à éviter et les solutions possibles. Mathias terminera son intervention par quelques conseils:

- commencer petit mais voir grand pour le long terme où le Big Data est inévitable
- s'entourer de personnes et de partenaires qui ont un intérêt commun
- mettre en place un socle de travail commun (Hadoop+Yarn par exemple)
- identifier les rêves et les frustrations pour pouvoir y répondre en misant sur un cycle vertueux de “cuisine/récolte”.

Les questions de l'assemblée à Mathias ont tourné autour de la notion de calcul de coût des solutions de type “cloud” (cad sous-traitées) vs solutions locales et des distances physiques entre puissance de calcul et stockage de la donnée.

Après une présentation générale du thème du “Big Data” par un intervenant extérieur au monde de la mer (intervention de Mathias Herberts) le comité avait souhaité faire parler des scientifiques, océanographes, de leurs difficultés liées aux problématiques de type “Big Data”. Ainsi **Stéphanie Mahevas** (chercheuse Ifremer au département Ressources Biologiques et Environnement dans l'Unité Ecologie et Modèles pour l'Halieutique) a présenté les difficultés d'analyse d'un modèle statistique de gestion des pêches adoptant une approche écosystémique du problème. Puis **Guillaume Maze** (chercheur Ifremer au département océanographie et dynamique des écosystèmes dans le laboratoire de physique des océans) a présenté les difficultés à conduire des diagnostics complexes et à analyser la structure verticale océanique à l'aide des données in-situ (type Argo). Enfin **Pierre Tandeo** (chercheur à Télécom Bretagne au département Image et Traitement Information) a présenté les difficultés rencontrées dans l'utilisation des données historiques, dans le contexte des données satellites, pour par exemple développer des produits sans “trous”. Cette session s'est terminée par une intervention de **Bertrand Chapron** (chercheur Ifremer au département océanographie et dynamique des écosystèmes dans le laboratoire d'océanographie spatiale) qui a présenté “Nymphaea”, le prototype de solution “Big Data” développé par le LOS pour faire face au stockage et à l'analyse des données satellites.

La matinée s'est achevée avec seulement 15 minutes de retard sur l'agenda. Toute l'assemblée a ensuite pris son déjeuner au restaurant de l'Ifremer.

Retour sur les interventions de l'après-midi

Dans l'après-midi, les interventions de Gilbert Maudire, Thomas Loubrieu, Philippe Lenca et Ronan Fablet ont permis de présenter un peu plus en détails les solutions possibles pour trois des problématiques que le comité avait jugé des plus importantes: l'interopérabilité (échange des données), le stockage et la fouille.

Gilbert Maudire (informaticien à l'Ifremer, responsable de l'unité de service Informatique et Données Marines du département Infrastructures Marines et Numériques) a fait une présentation intitulée “Partage et interaction des bases de données”. S'appuyant sur l'exemple de l'infrastructure de recherche SeaDataNet² qui permet via un portail web unique d'interroger 80 centres de données en Europe, G. Maudire a présenté les nécessités, difficultés mais aussi avantages de ce type

² <http://www.seadatanet.org/>

d'interopérabilité entre centre de données. En effet, l'observation in-situ de l'océan, qui reste complexe et coûteuse, est menée par de nombreuses équipes ayant des objectifs, des modes d'organisation et des sources de financement propres. Or, intégrer des jeux de données de référence aussi larges et cohérents que possible est une nécessité pour analyser la variabilité spatiale et temporelle d'un milieu aux larges interactions : côte (domaine des états riverains) au large ; atmosphère, colonne d'eau et fond ; physique, chimie et biologie (voire activités économiques). Dans ce contexte, mettre en place des réseaux distribués pour la gestion des observations présente de nombreux avantages : rapprocher la gestion des données des équipes effectuant les mesures en mer qui garde ainsi le contrôle et la visibilité de leur travail tout en permettant des échanges et une intégration à grande échelle. La recherche de l'équilibre entre architecture répartie et centralisée admet des réponses différentes en fonction des besoins à l'exemple de MyOcean (et Goos) ayant une composante temps-réel et de SeaDataNet, plus axé vers le temps différé.

Les questions de l'assemblée ont porté sur les limitations de l'infrastructure SeaDataNet, notamment l'absence d'une API qui permettrait l'interrogation du système par une machine. Par ailleurs, suite à une question sur les méta-données, G. Maudire a eu l'occasion de souligner leur rôle central dans le système actuel et de noter cependant qu'avec le développement des “smart sensors” (capable de s'identifier seul, à la manière d'un objet connecté via un port USB), cette importance des méta-données pourrait se voir réduite à l'avenir.



Thomas Loubrieu (ingénieur systèmes d'information à l'Ifremer du département Infrastructures Marines et Numériques) a ensuite fait une présentation intitulée “Small to Big Data” qui visait à faire un état des lieux pragmatiques des solutions possibles de stockages des données en fonction de la quantité à traiter. À partir des expériences en gestion de données du département Informatique et Données Marines de l'Ifremer, T. Loubrieu a donc passé en revue différentes problématiques liées à la volumétrie des données. Sur ces problématiques, Thomas a passé en revue les solutions possibles d'une part avec les technologies traditionnelles (formats, indexation, interopérabilité et partage de données) et d'autre part avec les systèmes dématérialisés. Cette présentation a permis de mieux comprendre et de cerner la complémentarité des solutions informatiques traditionnelles et avec celles spécifiques du Big Data.

Les questions de l'assemblée ont porté sur les détails techniques comparatifs de systèmes comme Hadoop et MPI ou sur la meilleure manière de favoriser le “data discovery” via la génération de produits intermédiaires comme des index des données.

Enfin **Philippe Lenca et Ronan Fablet** (professeurs et chercheurs à Télécom Bretagne) ont fait une présentation intitulée “Méthodes de fouille et d'analyse”. Ils ont présenté brièvement la fouille de données, ses aspects méthodologiques et les principaux algorithmes d'analyse. Ils ont illustré les nouveaux problèmes posés lorsque, par exemple, la volumétrie et la volatilité des données deviennent

torrentielles et où il est bien souvent nécessaire de faire du traitement par flots de données. Cette intervention visait à fournir un cadre au processus d'analyse des données pour l'audience non-initiées à ce type de formalisation. Ils ont ainsi rappelé ce que recouvrait les notions d'extraction et de gestion des connaissances (sélection, pré-traitement, transformation, fouille, schémas, évaluation, connaissance). Jusqu'à présent cette démarche pouvait se faire dans un mode "temps différé" avec un jeu de données figé. Désormais le processus d'analyse doit se faire en "temps réel" car le flot de données est *torrentiel*. De manière relativement surprenante, les 10 algorithmes les plus utilisés pour la fouille (pour la prévision, classification, classement et recherche de co-occurrences) sont relativement anciens car tous antérieurs aux années 2000. Quelques exemples ont permis d'illustrer certains de ces algorithmes et les spécificités du traitement des flots par des systèmes aux capacités limitées, notamment la nécessité de ne lire les données qu'un petit nombre de fois.

Les questions de l'assemblée ont porté sur les raisons du succès de chacune des méthodes et sur la différence entre les méthodes supervisées et non-supervisées.

Après une pause-café, l'assemblée était conviée dans la même salle à participer à une table ronde animée par **René Garelo**. Tous les intervenants de la journée y participaient ainsi que Tina Odaka (responsable du Pôle de Calcul Intensif de l'Ifremer). R. Garelo a souhaité soulever quelques problématiques qui n'avaient pas été abordées pendant la journée et a sollicité l'avis des intervenants sur ces questions.



La première question portait sur les approches multi-capteurs: quoi en attendre et comment les gérer ? Les exemples fournis le matin par B. Chapron et P. Tandeo ainsi que l'après-midi par R. Fablet montrent que ces questions sont déjà d'actualité pour l'océanographie spatiale et commencent à trouver des réponses. Des travaux en cours tendent à allier les données issues des radars et des capteurs infra-rouge et optique des satellites pour reconstruire les champs de vents ou de courants à la surface océanique. Plus généralement, c'est une approche quasiment écosystémique qui émerge: des capteurs et systèmes d'observations qui fournissent des informations complémentaires de l'état d'un système sans forcément caractériser la même variable ('physique ou biogéochimique par exemple). Le processus d'extraction de connaissance peut alors prendre deux voies: une première basée sur des méthodes de fouille s'appuyant sur la masse de données pour extraire des signaux synthétiques caractéristiques du système (par exemple une tendance plus ou moins visible dans les différentes variables), et une seconde basée sur le développement d'un modèle dans lequel toutes les variables sont représentées et qui permet de donner un cadre théorique d'interprétation des signaux.

Ces deux voies, probablement complémentaires, en sont à leur balbutiement dans le domaine des sciences marines.

La seconde question posée par R. Garelo portait sur l'obsolescence des données: faut-il stocker et garder toutes les données ? La question, volontairement naïve, a trouvé une réponse rapide. D'une part: acquérir des données en sciences marines, ou plus précisément des mesures, est un processus au coût encore trop élevé pour envisager de ne pas archiver toutes les données. D'autre part: les données, en particulier in-situ, sont trop éparpillées pour ne pas être toutes prises en compte dans les analyses. Enfin, les données de sciences marines sont des données acquises pour la recherche et la surveillance, une démarche en constante évolution qui impose de ne pas se dessaisir d'informations sur la base des connaissances actuelles. De futures découvertes, scientifiques ou techniques, pourraient nécessiter des informations considérées inutiles au temps de leur acquisition. Cette question soulève par ailleurs la problématique de la rétroactivité des analyses: en effet, comment porter des diagnostics développés pour les caractéristiques d'un système d'observation à un instant donné vers ses caractéristiques passées ? Des exemples sont donnés pour les données in-situ de température des sondes XBT par exemple, mais cette question reste sans réponse. Elle appelle donc forcément au stockage exhaustif de toutes les informations disponibles.

D'autres points seront rapidement abordés sans être approfondis comme les liens entre centres de calculs et centre de données (proximité nécessaire ? quid des solutions cloud ?) et les modes de distributions des données (quels sont les services attendus des centres de données ?).

La conférence s'achève à 17h15 sur le remerciement des intervenants et de l'audience.

Conclusion

L'évolution naturelle de nos disciplines scientifiques allant toujours vers plus de complexité et d'exhaustivité: développer et mettre en oeuvre des solutions adaptées à cette explosion de données en sciences marines est inexorable.

La phase préparatoire de cette conférence a montré un grand besoin d'échanges entre d'une part ceux qui mettent en oeuvre des solutions et d'autre part ceux à qui ces solutions s'adressent. En effet, il apparaît clairement que les contraintes techniques (qu'elles soient sur les méthodes statistiques de fouille ou sur l'équipement matériel) sont telles qu'aujourd'hui plus que jamais tous les acteurs (chercheurs, ingénieurs et informaticiens) doivent communiquer pour avancer ensemble.

Ces besoins et envies se sont trouvés confirmés par le résultat des inscriptions à la conférence. Avec 90 inscrits indique clairement le besoin d'informations sur cette thématique du "Big Data" dans la communauté des sciences marines à Brest. L'intervention de Mathias Herberts a permis de démystifier ce terme et de contribuer à clarifier ce que le monde de l'industrie du web (Google, Yahoo, etc ...) pouvait avoir développé - pour leurs besoins considérables liés à l'explosion du web social - et qui pourrait se retrouver utilisé progressivement dans d'autres domaines, tel que les sciences marines. Les autres interventions de la journée ont permis d'informer le public sur des problématiques et des solutions déjà mise en oeuvre dans le monde de l'océanographie.

Il apparaît clairement que ce virage technologique est un rendez-vous que notre communauté ne souhaite pas manquer.

Vidéos et présentations disponibles sur: <http://wwz.ifremer.fr/bigdata>

Annexe 1: liste des inscrits

	Nom	Prénom	Affiliation
1	Aboudib	Ala	Telecom Bretagne
2	ALIX	Anne-Sophie	Ifremer REM/GM/LGM
3	Ardhuin	Fanny	Ifremer
4	Bacher	Cedric	IFREMER/DYNECO
5	Bernard	Pascal	Datalone
6	Blandin	Jérôme	Ifremer /REM/RDT/I2M
7	Borremans	Catherine	Ifremer Brest
8	BOUDIERE	Edwige	ifremer
9	Bourbigot	Matthieu	CLS
10	Boussidi	Brahim	Télécom Bretagne
11	Bude	Gaël	SHOM
12	cabanes	cecile	CNRS
13	Carbonniere	Aurelien	Ifremer - Europe et International
14	CHARPENTIER	LIONEL	THALES SYSTEMES AEROPORTES
15	Charria	Guillaume	IFREMER/DYNECO/PHYSED
16	CHEVRIAUX	Yann	SHOM
17	conquet	florence	IMN/IDM/WEB
18	coppin	gilles	Lab-STICC / Télécom Bretagne
19	Cordier	Karen	CLS Brest
20	Couvelard	Xavier	LPO-CNRS
21	DAGORNE	Dominique	IRD
22	Demanèche	Sébastien	RBE/STH/LBH
23	Desmare	Stéphanie	SHOM
24	Detoc	Jérôme	IDM-ISI
25	DURIEUX	YVES	THALES SYSTEMES AEROPORTES
26	Dutreuil	Vincent	Ifremer
27	ETIENNE	Laurent	Institut de Recherche de L'Ecole Navale
28	FARCY	PATRICK	DS
29	Feucher	Charlène	IFREMER
30	Fichaut	Michèle	IFREMER-SISMER
31	Franco Contreras	Javier	TELECOM Bretagne
32	GABARRON	Nicolas	IUEM - CNRS UMS 3113
33	Garello	René	Telecom Bretagne
34	Gautron	Pascale	ENSTA Bretagne Brest
35	Gendronneau	Alexis	Altran
36	GONTIER	Olivier	IFREMER
37	Guelton	Serge	Télécom Bretagne
38	Herberts	Mathias	CityZen Data
39	KERMABON	Catherine	Ifremer
40	KPOGO NUWOKLO	KOMLAN	ifremer
41	Lagadec	Catherine	IFREMER/LPO
42	lample	michel	UBO/ AMURE
43	Le Bot	Philippe	LPO/Ifremer
44	le bras	yvan	CNRS
45	LE BRUN	Mikael	SHOM
46	LE BRUN	Dominique	Ifremer
47	Le Gentil	Sylvie	Ifremer
48	Le Guellec	Anne	SHOM
49	LE JEUNE	Denis	ENSTA Bretagne
50	LE MEUR	Chloé	IFREMER/RBE/STH/LBH
51	LE ROUX	JEAN-FRANCOIS	IFREMER

52	Le Toullec	Tristan	LPO - CNRS
53	Leblond	Isabelle	Ensta Bretagne
54	LEBLOND	Emilie	Ifremer
55	Lehuta	Sigrid	Ifremer
56	Letort	Cédric	SATT Ouest Valorisation
57	Leynaert	Aude	CNRS
58	LOUBRIEU	BENOIT	Ifremer
59	Maudire	Gilbert	Ifremer
60	Maze	Guillaume	Ifremer, LPO
61	MAZURIER	Alain	ALTRAN OUEST
62	Ménesguen	Claire	LPO
63	MERCIER	HERLE	cnrs
64	Mousset	Sylvain	ifremer
65	Odaka	Tina	Ifremer
66	OLLIVIER	Benjamin	doctorant Télécom Bretagne
67	PAUL	Frédéric	LOS
68	Pertuisot	Cécile	IFREMER
69	PETIT	Tristan	Ifremer/ODE/DYNECO/AG
70	Peyrot	Rémi	Institut Mines-Telecom
71	Pineau-Guillou	Lucia	Ifremer
72	Poussin	Benjamin	Code Lutin
73	PREVOSTO	MARC	IFREMER
74	QUEMENER	Gael	IMN/NSE
75	Quentel	Elise	Atout cadre
76	Raillard	Nicolas	IFREMER
77	salmon	loic	ecole navale
78	Schmitt	Thierry	SHOM
79	Schneider	Jean-Philippe	ENSTA Bretagne
80	Sepulveda	Andres	University of Concepcion, CHILI
81	SINQUIN	Jean-Marc	Ifremer
82	SIX	Lancelot	Quiet-Océans
83	Tandeo	Pierre	Institut Mines-Telecom
84	Tarot	Stéphane	IMN/IDM/SISMER
85	TAUVRY	Sébastien	ECA Robotics
86	THIERRY	VIRGINIE	IFREMER
87	Thomas-Doaré	Anne	IFREMER
88	Tourolle	Julie	Ifremer
89	Tréguier	Bruno	SHOM
90	VERBEQUE	Vincent	Pole Mer Bretagne
91	Woillez	Mathieu	Ifremer STH/LBH