LA BANQUE DE DONNEES

par

Yves ADAM, Huguette LAVAL, Pol CLOSSET, Jean-Pierre FOGUENNE, Walter KEUTGEN

Chapitre I

La gestion des données océanographiques

La tendance actuelle en océanographie est à la construction de modèles mathématiques pour représenter les phénomènes physiques, chimiques et biologiques. Pour être validés et initialisés, ces modèles demandent la connaissance d'une foule de données expérimentales, soit prises *in situ* soit déterminées en laboratoires, qui doivent être traitées, analysées et injectées ensuite dans les programmes de simulation.

A cause de l'énorme quantité de données nécessaires, la modélisation des systèmes marins implique deux obligations :

a) la réalisation d'expériences de grande envergure, étalées dans le temps et dans l'espace, qui fournissent des données cohérentes entre elles;

b) le traitement automatique des données.

Ce dernier fait l'objet du présent travail de recherche.

La seule manière d'établir un compromis entre le grand nombre de données disponibles et l'accessibilité à ces informations, est d'établir une base de données structurée selon le modèle mathématique et de développer un ensemble de programmes de gestion et d'analyse qui en rendent l'accès aisé. Les variables à étudier, à mesurer, à analyser sont de nature très diverse; ce peuvent être en effet :

a) des séries temporelles, régulières ou non, de paramètres physiques ou chimiques;

b) de séries de paramètres variant en fonction de la profondeur;

c) des séries de paramètres dont les valeurs sont mesurées aux points d'une grille régulière couvrant la zone à étudier.

Les paramètres sont eux-mêmes très différents les uns des autres, tant dans leur signification physique que dans leur méthode de mesure. Ils peuvent être engendrés par des instruments automatiques ou non, venir de différents laboratoires; une série peut n'avoir de signification que comparée à une autre.

Il est nécessaire, d'abord, de définir un *format de stockage* des données suffisamment souple pour faire face au problème de la diversité des données; ensuite, de développer des programmes généraux pour mettre en forme, écrire, retrouver et sélectionner les différents types de données, chaque programme étant construit de façon à être facilement exploitable par l'utilisateur.

La plupart du temps, et plus particulièrement lors d'expériences regroupant de nombreuses institutions, ces données doivent être échangées entre différents laboratoires, sur des supports digitaux susceptibles d'être traités automatiquement par les systèmes informatiques dont disposent les centres de recherche. Etant donné l'énorme variété des données à traiter, ainsi que la grande diversité de conception des différents systèmes de traitement et d'analyse de l'information, il est exclu de développer des logiciels particuliers à chaque type de données, et compatibles seulement avec quelques systèmes. Il faut donc définir un *format d'échange* de données sur support digital (et de préférence sur bande magnétique), autodescriptif (c'est-à-dire qui contienne toutes les informations nécessaires pour définir les données) et susceptible d'accepter n'importe quel type de données. Il faut de plus que ce format d'échange soit compatible avec la structure de la majorité des ordinateurs en fonction dans les centres océanographiques. Il faut enfin développer un logiciel qui lise et écrive ces structures et qui soit portable d'ordinateur à ordinateur. On va s'attacher, dans les chapitres suivants, à expliquer les raisons qui ont conduit à l'élaboration de tels formats et à décrire sommairement leur structure.

Chapitre II

Construction de la base de données

1.- Etat présent de la base de données

Ce chapitre est consacré à la description et au développement de la base de données. On va mettre l'accent sur les difficultés qui ont été rencontrées lors de la construction de ce système d'information. Les scientifiques impliqués dans des problèmes de traitement et d'exploitation de données ont été, jusqu'ici, bien trop occupés par d'autres recherches fondamentales (développement de modèles mathématiques, création d'algorithmes numériques) ou par des questions de technologie (réalisation de systèmes d'acquisition de données) pour pouvoir consacrer leur temps à l'élaboration d'une banque de données.

Toutefois, il existe déjà plusieurs fichiers de données aisément accessibles sur des supports lisibles par ordinateurs (le plus souvent des bandes magnétiques) et un logiciel élémentaire capable de stocker, de lire, de classer et de visualiser ces données. Le logiciel n'a pas d'unité et consiste en diverses chaînes de plusieurs programmes, chacune étant spécialisée dans le traitement d'un type particulier de données. Jusqu'à présent, les outils mathématiques développés pour un type de données ne peuvent être immédiatement adaptés à un autre type. De plus, les programmes ont été réalisés pour des ordinateurs de types particuliers, mais des travaux sont actuellement en cours pour assurer la portabilité du logiciel et des données (pour rappel, deux types d'ordinateurs sont disponibles dans le cadre du projet pour l'exploitation des données : un HP 2100A et un IBM 370/158).

Les données proviennent essentiellement de deux types de sources :

- les stations d'acquisition automatiques opérant dans le cadre du Programme National sur l'Environnement Physique et Biologique;
- les échanges avec d'autres institutions coopérant avec les chercheurs du Programme dans le cadre d'expériences communes.

2.- Données issues du Programme

Les chaînes d'acquisition automatique des données concernent :

a) les données des bouées,
b) les données des courantomètres,
c) les données météorologiques.

Toutes sont des séries temporelles (données toutes les 15 minutes pour les courantomètres, toutes les heures pour la bouée, toutes les 3 heures pour la météorologie).

La première chaîne a été installée à Ostende, les deux autres à Liège.

2.1.- Données météorologiques

La chaîne traitant les données météorologiques est la plus sophistiquée de toutes, à cause de la structure complexe des séries temporelles (28 échantillons et 14 variables à chaque pas de temps).

Cette chaîne traite les données météorologiques qui proviennent en temps réel de la Régie des Voies Aériennes (RDVA) grâce à un récepteur telex. Les informations météorologiques codées (exemple fig. 1) sont enregistrées telles quelles sur ruban perforé par la perforatrice automatique associée au récepteur. Un certain nombre de stations météorologiques dont les émissions sont régulières ont été sélectionnées; elles forment un réseau relativement dense couvrant le Southern Bight et ses côtes (fig. 2); elles comptent notamment des stations terrestres dont la validité des informations en ce qui concerne la représentation de la situation en mer n'est pas toujours certaine, notamment pour les données relatives au vent.

```
        vents geostrophiques

     du   26 /  1 / 1976     a 0900 z

     a  = 327 /  26 nds   n =  4
     b  = 270 /  19 nds   n = 10
     c  = 335 /  25 nds   n = 11
     d  = 291 /   8 nds   n = 16
     e  = 316 /  17 nds   n = 11
sauk22 egrr 260900
  03898 63221 98022 40607 13253 145//=

sanl40 ehwx 260929
ehvb 20004    4500         1cu025 3ci180 m01/m03 1012
 grn wht tempo grn=
sinl23 ehdb 260900
  06220 83006 98261 11104 7937/ 01713 30201 35502 4068/=
  sinl21 ehdb 260900
  06235 51906 65151 10701 39462 52810 81708 83916=


SCC sa ehts not available
sinl23 ehdb 260900
  06300 53115 98272 13304 49360 01801 30402 00/00 4073/=
  sinl22 ehdb 260900
  06310 12109 57050 13300 18400 51303 81815=


SCC sa ehgo not available

sanl eham 260925=
ehrd19005 5000 1cu022 3ci250 m01/m01 tempo 2000 9//005=

sabx50 ebwm 260925
ebfn 26002    9999          3cu020 01/m01 1014
 black wht wht inter amb=


SCC sa ebwh not available
mmxx 2609
06407 13507 70010 13902 12500 52105 81825=
mmxx 2609
06408 30204 98027 14603 32400 51205 91838=
sifr22 lfpw 260900
  07002 83412 66022 15000 885// 51312 91827=


SCC sa lfac not available


SCC sa lfaj not available
✠
41746 enviro b
25757 enviro b
```

fig. 1.
Exemple de messages codés produits par la R.D.V.A.

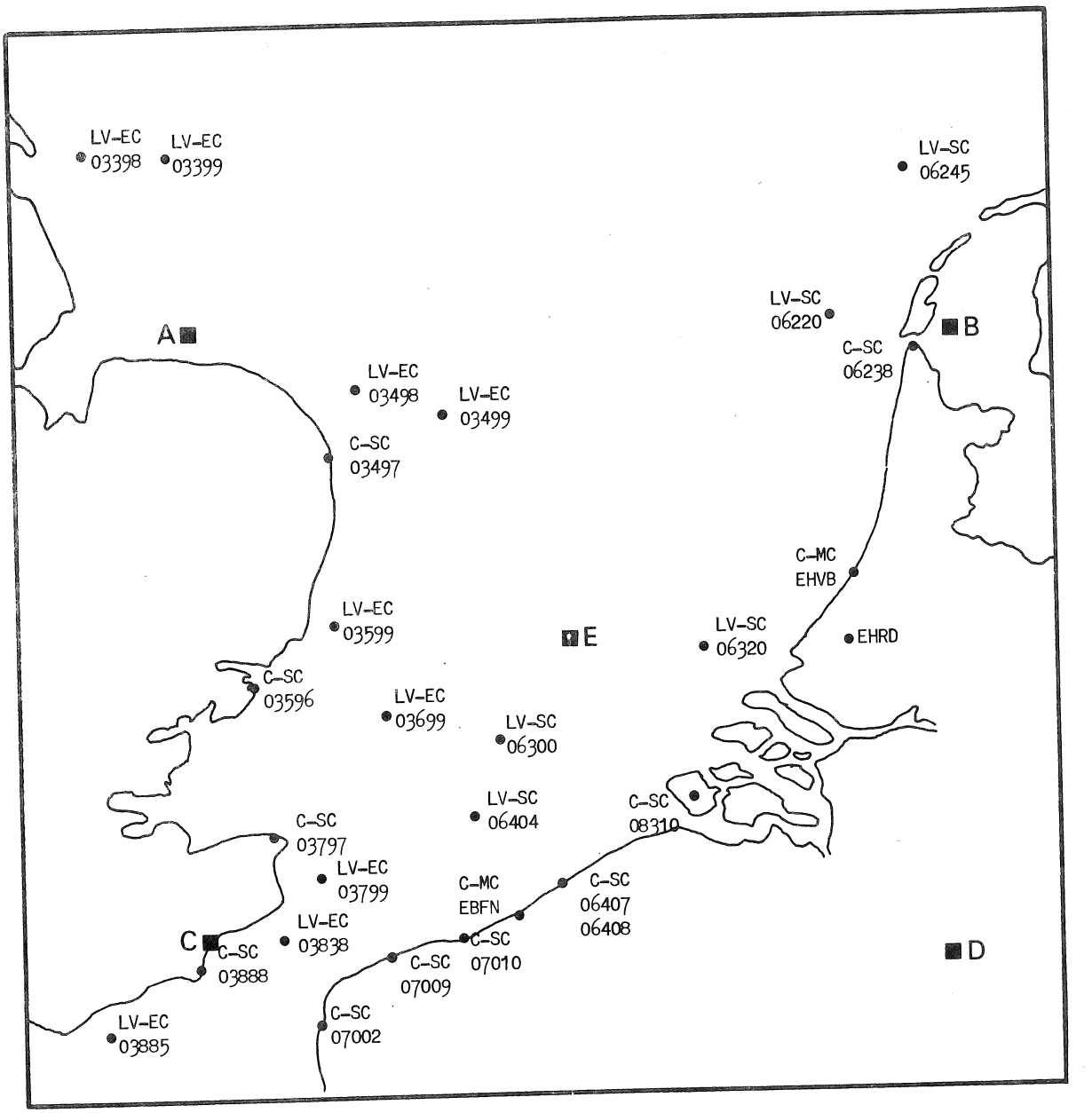


fig. 2.
Position des stations météorologiques dans le sud de la mer du Nord

La plupart des stations émettent toutes les 3 heures, mais certaines sont plus irrégulières. La liste des paramètres mesurés en ces stations et retenus dans les fichiers est donnée au tableau 1. En plus

Tableau 1

Liste des données synoptiques stockées dans la base de données

Ce tableau montre les divers types de données stockées par le second programme de la chaîne de traitement. Quelques stations fournissent les informations en code "anglais", d'autres en code "Synop" ou "Metar". L'ensemble des informations varie d'après le type de code. Dans le but d'enregistrer les données sous un format commun, les informations inexistantes dans certains codes sont remplacées par une valeur fictive 9999 . Quelques stations, bien qu'utilisant un code donné, n'envoient pas un ensemble complet d'informations : les astérisques dans le tableau indiquent les informations qui manquent dans les messages de certaines stations. Une croix indique une information toujours présente.

Type de données	Unités	Code anglais	Code Synop	Code Metar
Nom de la station	-	x	x	x
Visibilité horizontale	mètres	x	x	x
Hauteur	demi-mètres	*	9999	9999
Période } des vagues	secondes	*	9999	9999
Direction de propagation } des vagues	dizaines de degrés	x	9999	9999
Température de l'air	°C	x	x	*
Température de la mer	°C	x	9999	9999
Température de rosée	°C	9999	x	*
Pression barométrique	dixièmes de mb	*	x	*
Vitesse du vent	noeuds	x	x	x
Direction du vent	dizaines de degrés	x	x	x
Temps présent	code	x	x	9999
Temps passé	code	x	x	9999
Tendance barométrique	code	9999	x	9999
Caractéristique de la tendance barométrique	code	9999	x	9999

de mesures, la R.D.V.A. fournit les vents géostrophiques calculés et prédits (toutes les 6 heures) en 5 points : A,B,C,D,E sur la figure 2. Le contenu des rubans perforés (support d'information primaire) est transféré sur une bande magnétique; chaque fois que le volume d'informations est suffisant (tous les quinze jours) ou que le besoin en est exprimé (en cas d'urgence), un premier programme de la chaîne

a) traduit le code SIEMENS du ruban en code EBCDIC et transfère le résultat de ce décodage sur une autre bande magnétique;

b) optionnellement produit une liste des codes transmis.

La bande magnétique créée par ce premier programme sert d'entrée à un second programme, qui interprète les codes synoptiques en valeurs numériques, utilisable par les programmes de traitement subséquents. Le second programme produit :

a) une bande magnétique comprenant 3 types de fichiers :
   - fichiers de résultats d'observations,
   - fichiers de vents géostrophiques calculés,
   - fichiers de vents géostrophiques prédits;

b) optionnellement une liste de messages en clair.

Ces deux programmes doivent absolument traiter les données avant que celles-ci soient effectivement utilisables. Aucun filtrage, aucune épuration, aucune interpolation ou extrapolation n'est réalisé à ce stade : les valeurs numériques fournies sont l'exacte et stricte traduction des messages codés (à condition toutefois que le sens de ceux-ci n'ait pas été altéré par des erreurs de transmission); de la sorte on peut après examens répétés, déterminer les éventuelles erreurs systématiques, repérer les stations dont l'information est peu sûre et prendre les dispositions nécessaires dans les programmes d'exploitation des données. Plusieurs de ceux-ci ont été développés, notamment :

a) un programme fournissant, par interpolation sur un réseau régulier couvrant le Southern Bight, la valeur du champ de vent (vent observé et vent géostrophique calculé). La figure 3 montre plusieurs répartitions du vecteur-vitesse du vent (moyenne sur 24 h) pendant quelques jours de la campagne JONSDAP 73. Les mêmes valeurs interpolées sont utilisées pour le calcul de la tension due au vent dans des programmes de simulation hydrodynamique. Les mêmes calculs seront effectués tout au long de la campagne JONSDAP 76;

b) un programme permettant de faire le bilan thermique du Southern Bight en tenant compte de la température de la mer et de l'air, du taux d'humidité et de la couverture nuageuse.

Les données stockées sous le format actuel peuvent facilement être mises sous le format standard décrit plus loin.

fig. 3.

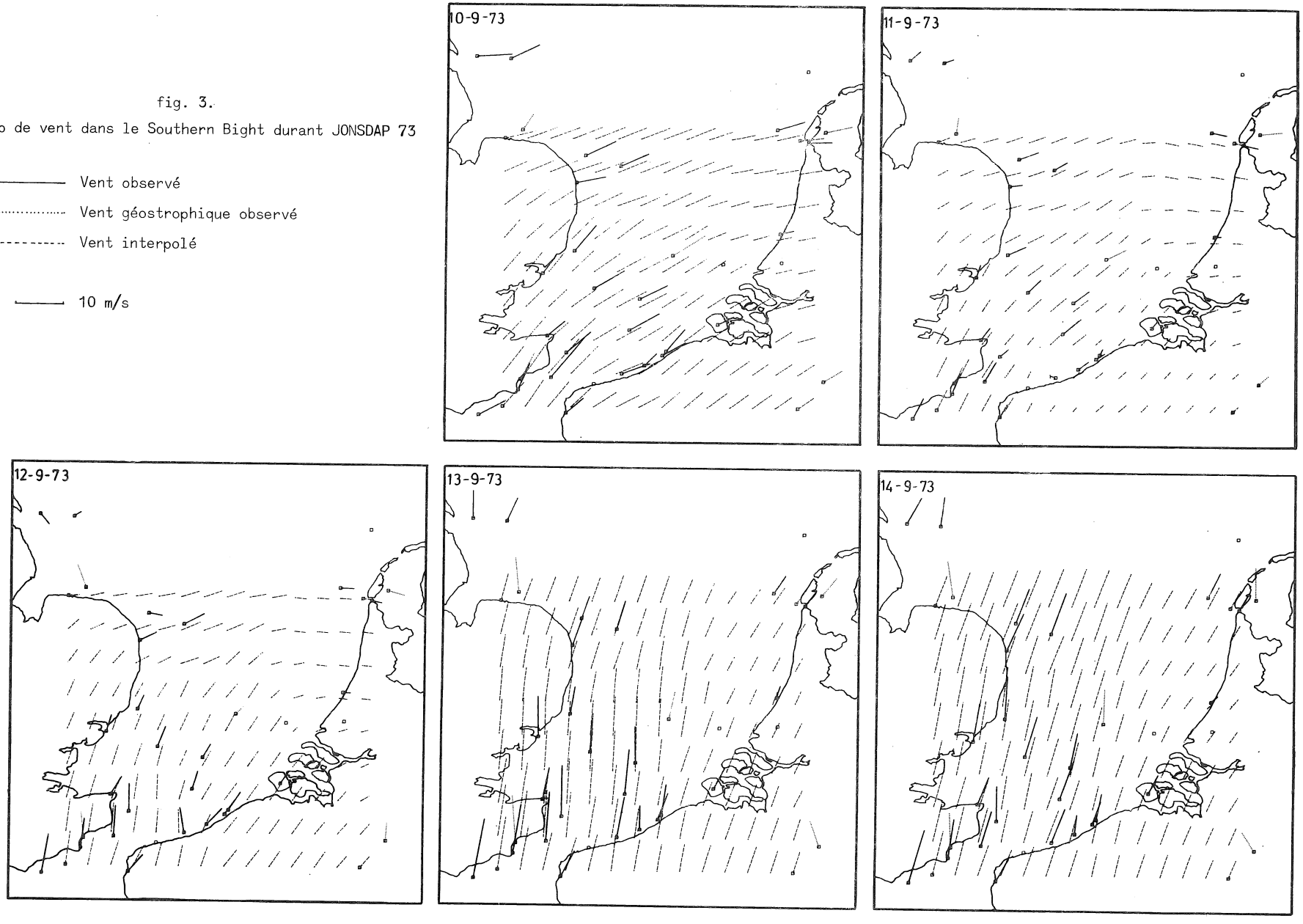
Champ de vent dans le Southern Bight durant JONSDAP 73

——— Vent observé

..... Vent géostrophique observé

- - - - - - Vent interpolé

——— 10 m/s



2.2.- Données des bouées automatiques

Le logiciel développé pour le traitement des données transmises par les bouées automatiques comprend 2 chaînes de programmes; toutes deux se limitent à construire des séries temporelles de données prétraitées, c'est-à-dire d'où l'on a éliminé les erreurs dues à la transmission et au décodage. Les données résultantes sont prêtes à être analysées et exploitées.

a) La première chaîne est l'ensemble des programmes qui effectuent le prétraitement en temps différé (*off-line data acquisition system* : DAS) des informations fournies par le système d'acquisition de données sur ruban perforé. La bouée émet ses informations (canal par canal) qui sont reçues à Ostende par le système d'acquisition; celui-ci traduit les signaux radio en code digital et envoie les résultats vers une perforatrice automatique qui fournit un ruban de papier portant les données brutes écrites en code-caractère ASCII. Le ruban est lu par un premier programme qui stocke les données sur une bande magnétique, sous un format fixe; cette bande est alors traitée par une série de programmes qui

- produisent une liste des données;
- corrigent les erreurs de décodage;
- éliminent les cycles de données trop mal transmis.

La plupart des corrections (caractères incorrects, transmissions tronquées, ...) se font à l'aide de programmes standard du système d'exploitation. La bande purifiée est interprétée par un programme final qui, suivant les instructions de l'opérateur, exécute différents traitements des informations correspondant aux divers canaux de la bouée et qui construit les séries temporelles définitives. On ne décrira pas davantage cette chaîne étant donnée qu'elle ne sert plus que de système de sécurité en cas de défaillance de la seconde chaîne, beaucoup plus automatisée et facile à manipuler.

b) La seconde chaîne est un ensemble de programmes d'acquisition de données en temps réel (*on-line data acquisition system* : OLDAS). Grâce à elle, le prétraitement des données se fait de manière quasi-automatique, ne nécessitant l'intervention d'un opérateur que pour la

définition exacte des traitements à opérer (qui peuvent éventuellement différer d'une bouée à l'autre) et en cas d'erreur grave; l'acquisition elle-même peut se faire sans intervention humaine pendant des périodes fort longues. Pour réaliser ce logiciel très élaboré, il a fallu faire appel à toutes les ressources du système d'exploitation en temps réel (*Real Time Executive System*) qui gère actuellement l'ordinateur d'acquisition de données.

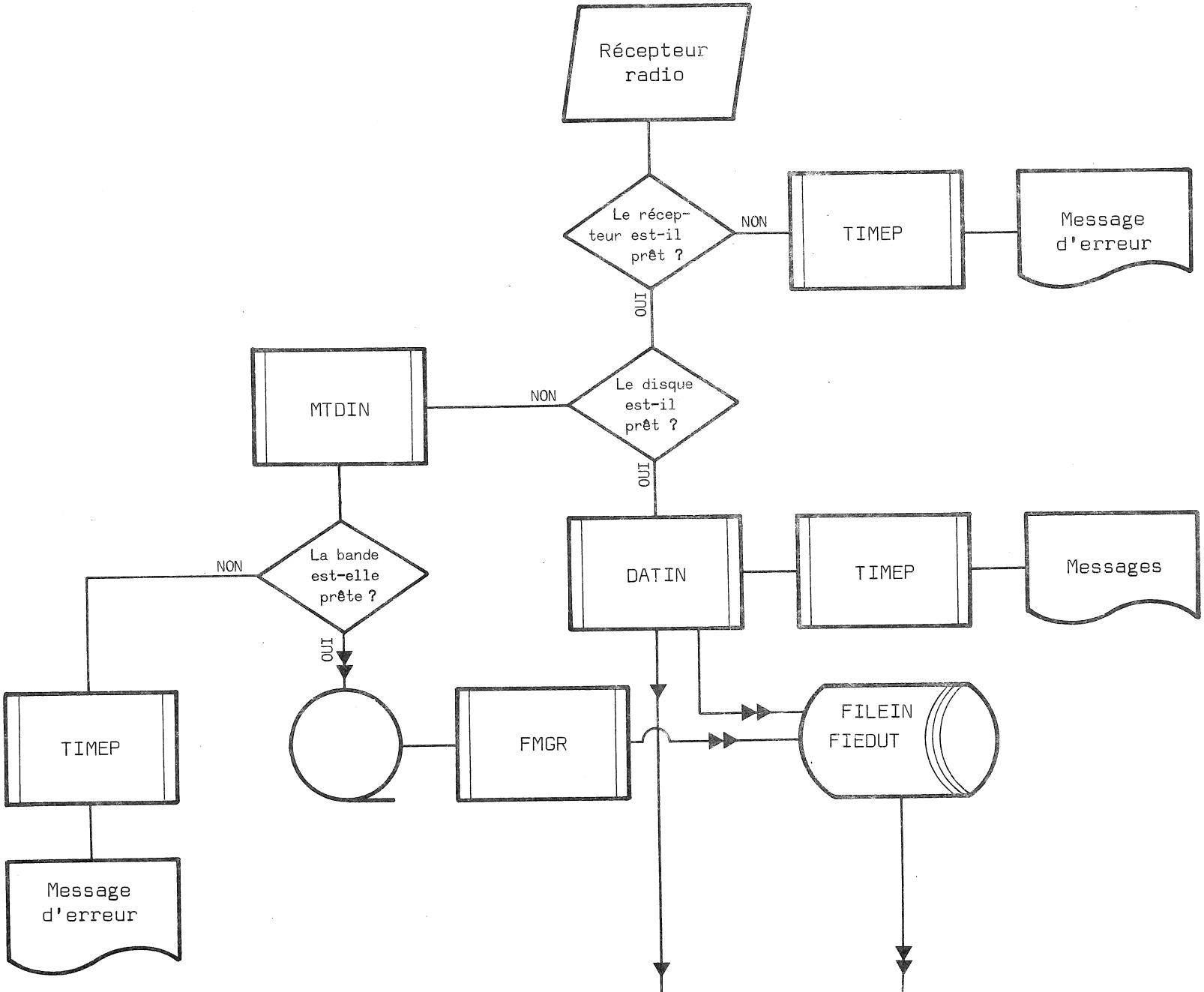
Les différents programmes sont activés automatiquement et selon les besoins du traitement par leur prédécesseur dans la chaîne.

La première phase est l'acquisition des données elle-même : dès que le récepteur capte une émission, un premier programme (DATIN) est activé automatiquement par le système-récepteur; selon que le disque est capable de stocker les données ou non, ce programme "lit" les données une à une et les sauve sur un fichier ou transfère le contrôle à un autre (MTDIN) qui lit les données et les sauve sur bande magnétique; cette bande peut soit être transformée en fichiers disques par la suite et être traitée par les phases suivantes de la chaîne OLDAS, soit être analysée par la chaîne DAS. Dès que l'acquisition est terminée, le contrôle passe à la phase de prétraitement. Pendant toute la durée de l'acquisition, le récepteur est logiquement déconnecté de sorte qu'il ne puisse y avoir d'interférences entre les émissions de plusieurs bouées. Bien entendu, des sécurités sont prévues pour éliminer les transmissions erronées (par exemple les messages sur la même fréquence envoyés par des émetteurs autres que la bouée). Les programmes DATIN et MTDIN restent en permanence dans la mémoire centrale de l'ordinateur, parce que :

i) ils doivent pouvoir être exécutés même en cas de panne ou d'arrêt du disque,

ii) ils doivent pouvoir être activés avec la plus grande rapidité possible.

La phase de prétraitement comprend les programmes PLINK et PRCES; le premier peut être activé par la première phase (DATIN) et dans certains cas spéciaux par PRCES; le second est activable par l'opérateur ou par PLINK. PLINK est un programme interface dont le but principal


Récepteur radio
Le récep-teur est-il prêt ?
NON
TIMEP
Message d'erreur
OUI
MTDIN
NON
Le disque est-il prêt ?
OUI
La bande est-elle prête ?
NON
OUI
DATIN
TIMEP
Messages
TIMEP
FMGR
FILEIN
FIEDUT
Message d'erreur

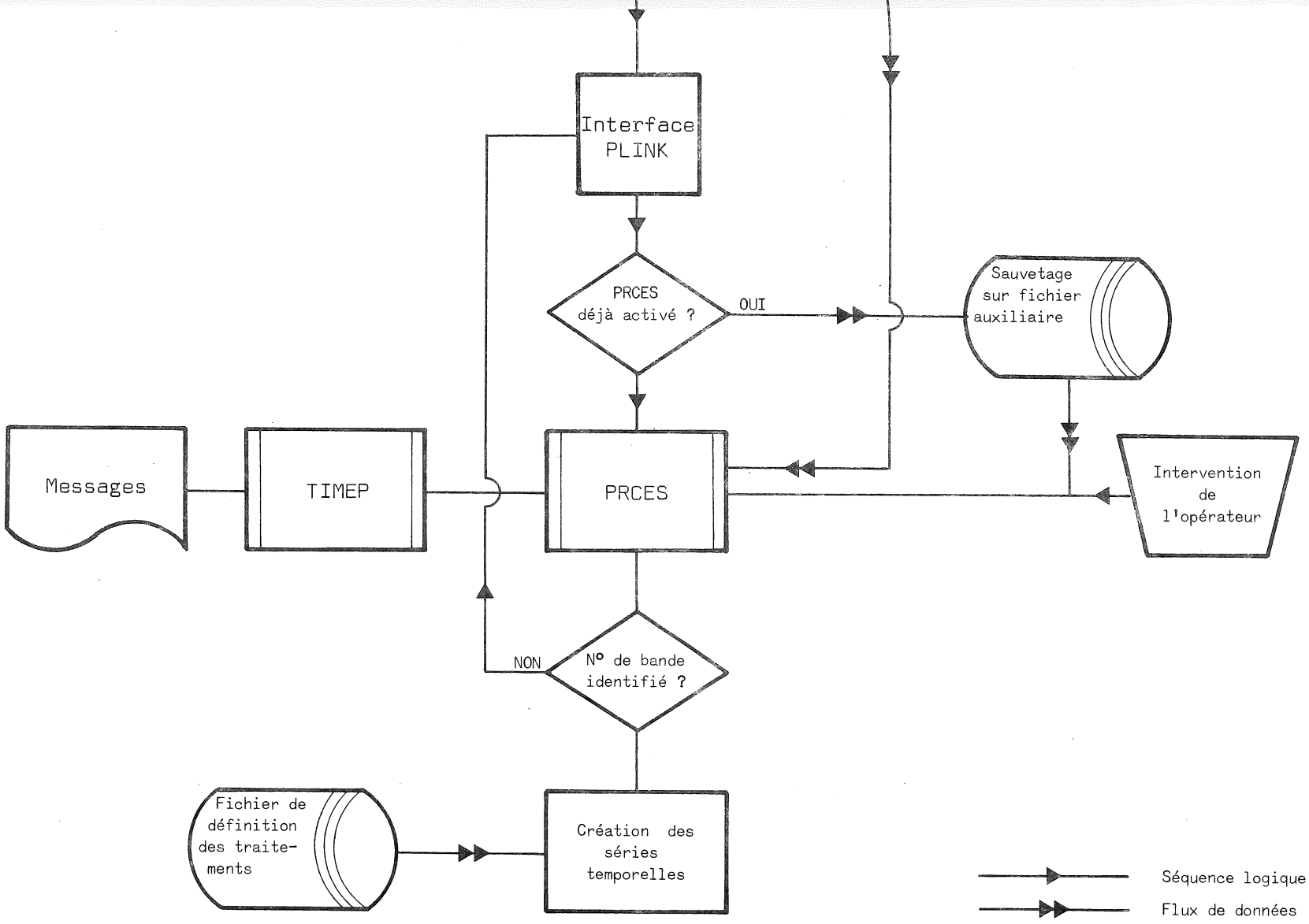


fig. 4.

Organigramme de la chaîne OLDAS

est de mettre provisoirement à l'abri des données en attendant que PRCES les manipule (ce dernier programme peut ne pas être disponible losqu'il est déjà en train de traiter les données d'une autre émission). PRCES identifie la bouée qui a émis les données contenues dans le fichier qu'il est en train d'examiner, va chercher dans un fichier auxiliaire des informations concernant les traitements qu'il a à effectuer (interprétation des informations des canaux) et opère ceux-ci. Les résultats sont stockés sur des fichiers-disque puis sur bande magnétique sous forme de séries temporelles dont la structure est semblable à celle des fichiers en format standard de traitement décrits plus loin. Les fichiers descriptifs mentionnés plus haut doivent être élaborés avant tout traitement effectif et ne contiennent que des informations concernant les bouées actives. Le diagramme de la figure 4 résume la logique des opérations effectuées par la chaîne OLDAS.

2.3.- Données courantométriques

Cette chaîne d'acquisition est la moins sophistiquée, car le décodage des cassettes provenant des instruments se fait chez le fabricant. La forme sous laquelle ses données sont reçues dépend du type de l'instrument.

Les données des VACM arrivent sous une forme prétraitées de la WHOI, où elles ont été décodées, nettoyées et mises en forme (avec notamment, l'interpolation effectuée pour des valeurs manquantes. Les bandes magnétiques fournies sont lisibles par les programmes de traitement directement, mais les données sont malgré tout mise sous un format intermédiaire, qui est le même pour tous les types de courantomètres, de manière à unifier les programmes subséquents.

Les données brutes NBA arrivent sur rubans perforés en caractères ASCII. Aucune édition ni prétraitement n'a été fait. Elles doivent être filtrées avant d'être mise sous le format intermédiaire.

Les données Plessey et Anderaa arrivent encodées selon un code binaire sur rubans perforés ou sur bandes magnétiques. Elles doivent être traduites en ASCII ou EBCDIC, avant d'être disponibles pour les

programmes d'exploitation. Les données doivent également être filtrées préalablement à leur insertion dans la base de données.

Pour chaque type d'instrument, il existe un programme qui établit les séries temporelles, les édite et les stocke sur des fichiers d'entrée pour les programmes d'analyse et de tracé.

3.- Données reçues par échange

La base de données comprend, en plus des données qui proviennent d'instruments de mesure automatiques, des informations qui nous sont parvenues dans le cadre d'expériences communes.

3.1.- Données de l'expérience JONSDAP 73

3.1.1.- Historique

JONSDAP 73 est une expérience entreprise en collaboration par 4 centres océanographiques riverains de la mer du Nord. Son but était l'étude approfondie de l'hydrologie et l'hydrodynamique du Southern Bight pendant une période relativement étendue, en l'occurrence les mois d'automne. Les institutions qui prirent part à cette campagne de mesures étaient :

- le Fisheries Laboratory de Lowestoft,
- l'Institute of Oceanographic Sciences (I.O.S.) de Bidston,
- le Koninglijk Nederlands Meteorologisch Instituut (K.N.M.I.) De Bilt,
- les chercheurs du Projet Mer, et plus particulièrement le groupe Math. Modelsea.

Les données acquises au cours de la campagne étaient essentiellement de nature physique, à savoir :

- des relevés de bathysondes,
- des élévations de marées (côtières et pélagiques),
- des mesures courantométriques,
- des données météorologiques diverses, dont on a déjà parlé précédemment.

L'expérience avait été soigneusement préparée en ce qui concerne la phase d'acquisition de données et de fait, la moisson fut abondante.

Il avait été convenu que les données acquises par chaque institution devaient être échangées avec les autres participants, de sorte que chaque groupe de recherche puisse appliquer à l'ensemble des mesures ses méthodes d'analyse propres et nourrir les modèles mathématiques qu'il développe avec toutes les données disponibles relatives à la région qu'il étudie.

L'échange des informations n'avait pas été préparé, chacun ayant estimé que le transfert des données d'un centre à l'autre ne poserait guère de problèmes.

3.1.2.- Difficultés de l'échange

Or la quantité de données à échanger (notamment les séries temporelles issues de courantomètres à enregistrement automatique) est telle que seuls peuvent être utilisés des supports de données susceptibles d'être traités automatiquement par des systèmes informatiques (en effet, les risques d'erreurs dues à d'éventuelles transcriptions manuelles deviennent rapidement prohibitifs).

Les supports utilisables sont :

- les bandes magnétiques,
- les bandes de papier perforé,
- les cartes perforées

(ce dernier support est peu recommandé quend le volume des informations est très grand car les risques de mélange et le poids des cartes fixent rapidement une limite à la quantité d'informations qu'il peut transférer). Seules quelques données, comme les élévations horaires du niveau de la mer ont été échangées par ce moyen, sans problème majeur heureusement. Par contre, il est exclu de l'utiliser pour l'échange des données courantométriques, beaucoup plus volumineuses. Les 2 autres supports se sont avérés indispensables mais ont posé de graves problèmes de compatibilité entre systèmes de traitement de l'information. En effet, la situation est la suivante :

- le Fisheries Laboratory, équipé de matériel ICL, fournit des données mises en forme sur bande magnétique;

- l'IOS-Bidston, équipé de matériel IBM, fournit également des données sur bande magnétique;

- le KNMI ne peut fournir des données que sur ruban perforé;

- Math. Modelsea, ayant accès au Centre de Calcul de l'Université de Liège, (matériel IBM) fournit des données sur bande magnétique.

Il se fait que si les bandes produites à Liège et à Bidston sont parfaitement lisibles par les ordinateurs réciproques (c'est le même matériel), Bidston et Lowestoft avaient jusqu'alors tenté en vain d'échanger leurs données, leurs systèmes respectifs pour l'écriture/lecture des bandes magnétiques étant incompatibles pour des raisons qui seront détaillées plus loin; d'autre part, aucun de ces deux derniers centres n'est équipé pour traiter des rubans perforés. Après de multiples essais infructueux, il a été décidé d'adopter la procédure suivante : toutes les données de quelqu'origine qu'elles soient, sont envoyées à Liège, où le Groupe Math. Modelsea tente de les décoder et de les mettre à la disposition des autres centres. La base de données courantométriques issues de JONSDAP 73 est donc centralisée à Liège; encore faut-il déchiffrer les données issues des divers centres, en faire une base de données exploitable, et les redistribuer aux institutions coopérantes sous une forme telle que disparaissent les incompatibilités qui avaient rendu impossible l'échange direct. Le choix de Liège comme centre d'échange est fondé sur deux raisons :

i) le système d'ordinateur y est le plus complet et accepte tous les types de support;

ii) l'équipe scientifique qui s'y trouve compte notamment des scientifiques expérimentés dans les problèmes de traitement de l'information et est plus apte que toute autre à surmonter les difficiles problèmes rencontrés.

3.1.3.- Mise au point des techniques d'échange

En principe, il suffit de résoudre trois problèmes, étant donné que les systèmes de Liège et de Bidston sont parfaitement compatibles :

- décoder les données fournies par le K.N.M.I. sur ruban perforé et les faire entrer dans la base de données centralisée à Liège; en effet, à l'heure actuelle, l'ordinateur du K.N.M.I. est capable de *lire* sans problème des bandes magnétiques créées à Liège;

- décoder les données fournies par le Fisheries Laboratory sur bandes magnétiques produites par un système ICL;

- encoder les données de la base centrale de manière à ce qu'elles puissent être utilisées directement par le système de traitement de Lowestoft.

Le problème du décodage des données courantométriques originaires du K.N.M.I. est similaire, dans les difficultés qu'il a fallu contourner, au problème posé par les données météorologiques qui est exposé dans le paragraphe suivant; on ne le détaillera donc pas ici; il faut cependant rappeler que le système d'exploitation du Centre de Calcul de l'Université de Liège (OS/360-VS fonctionnant sous ASP) n'est absolument pas conçu pour faciliter le traitement des bandes perforées; il a donc été nécessaire de créer une procédure complète de décodage, comprenant notamment :

- un programme de décodage proprement dit (passage du code-ruban perforé au code EBCDIC);

- un programme de translation (traduction des chaînes de caractères EBCDIC en valeurs significatives à introduire dans la base de données) pour "assimiler" ces données.

Les difficultés rencontrées lors de l'échange entre Lowestoft et Bidston ou Liège sont d'une nature nature : elles proviennent d'une incompatibilité de structure fondamentale entre bandes magnétiques produites par des ordinateurs de conception différente. En effet,

- les ordinateurs IBM utilisent des caractères de 8 bits, qui groupés par 4, peuvent former des mots de 32 bits tandis que les ordinateurs ICL utilisent des mots de 24 bits (unités fondamentales de la mémoire centrale) sur lesquelles sont éventuellement définis des groupes de 4 caractères de 6 bits; dans le premier cas, l'unité de base de la mémoire centrale est le caractère (ou byte) de 8 bits, tandis que dans le second cas, c'est le mot de 24 bits;

- le système d'exploitation d'IBM permet d'écrire des bandes magnétiques sans étiquette d'identification spéciale; le code-caractère (permettant un échange relativement aisé avec d'autres systèmes) d'IBM est le code EBCDIC largement utilisé; le système d'exploitation d'ICL exige la présence en début de bande, d'une étiquette d'identification de format déterminé; de même, le début et la fin des enregistrements de données sont délimités par des étiquettes (sentinelles) de format spécial; toutes ces étiquettes doivent être écrites en code interne (octal); le code-caractère d'ICL est un code spécial, basé sur des caractères de 6 bits (code ICL);

- chez IBM, les enregistrements logiques sont définis par le système d'exploitation sur des enregistrements physiques; chez ICL, les enregistrements logiques compris dans les enregistrements physiques sont séparés, sur la bande magnétique elle-même, par des identificateurs qui en définissent la longueur.

On conçoit donc la difficulté de transférer des données d'un système à l'autre; le passage ne peut se faire qu'en analysant, au niveau des chaînes de bits écrites sur les bandes, les structures des enregistrements physiques.

Pour mettre des données provenant d'une bande ICL sous une forme directement lisible par le système d'exploitation OS/360, il faut :

- ignorer les diverses étiquettes et sentinelles,
- découper chaque chaîne de bits correspondant à 1 enregistrement physique en groupes de 6 , et faire correspondre à chacun de ces groupes un caractère EBCDIC;
- éliminer les identificateurs de longueur, en tenant compte du format d'écriture des données et reformer des chaînes de caractères continues; ces chaînes sont lisibles alors par des ordres de lecture standard du système IBM.

Pour mettre des données provenant d'une bande IBM convenablement structurée, sous une forme directement lisible par le système d'exploitation ICL, il faut :

- recréer les étiquettes et sentinelles sous la forme imposée;

- déterminer les identificateurs de longueur et les insérer (sous forme de chaînes de caractères EBCDIC) entre les enregistrements logiques;

- faire correspondre à chaque groupe de 8 bits un caractère du code ICL et reformer des chaînes continues; ces chaînes sont alors lisibles par des ordres de lecture standard du système ICL.

Il a fallu mettre au point toute une panoplie de programmes pour réaliser le transfert dans les deux sens, en passant par l'écriture des données sous un format intermédiaire.

Le format intermédiaire est celui sous lequel sont écrits les fichiers composant la base de données courantométriques; c'est un format transitoire entre le format d'échange et le format standard de la banque de données générale, qui sera décrit plus loin; le format intermédiaire est d'ailleurs provisoire et les données sont écrites sous format standard au fur et à mesure que les programmes d'exploitation de la banque de données sont mis au point. Quant au format d'échange, il a actuellement la structure suivante : chaque fichier correspondant à une série de mesures issues d'un même instrument comporte :

a) un enregistrement (en-tête) définissant les données, il définit :

- la zone marine où les mesures furent prises (ici, la mer du Nord),
- l'époque approximative des mesures (septembre-octobre),
- le nom de l'expérience en cours (ici, JONSDAP 73),
- la position du mouillage (mot code définissant une station de prise de mesures),
- la profondeur à laquelle se trouve l'instrument (fond, milieu, surface),
- le numéro de série de l'instrument,
- l'heure du début des mesures (jour, mois, année, heure, minute),
- le nombre de cycles de données,
- la variation magnétique,
- l'indication du traitement déjà effectué (par exemple, si une moyenne a été calculée),
- l'intervalle de temps entre cycles de données;

b) des enregistrements contenant les cycles de données;

c) des enregistrements contenant les moyennes horaires des mesures; ceci a été introduit pour permettre des comparaisons directes entre mesures fournies par des instruments dont l'intervalle d'échantillonnage n'est pas le même.

A l'heure actuelle, le problème de l'échange des données de JONSDAP 73 peut être considéré comme résolu. Il faut rappeler que le format actuel d'échange est exclusivement conçu pour la manipulation de *données courantométriques*.

3.2.- Données de l'expérience JONSMOD

L'expérience JONSMOD est encore actuellement en cours. Elle consiste essentiellement à comparer différents modèles pour la prédiction des marées et des ondes de tempête en mer du Nord, nourris par un ensemble de données communes, en l'occurrence des élévations connues du niveau de la mer et les conditions météorologiques qui existaient pendant la période simulée (automne 73); ces données communes, ainsi que des résultats permettant la validation des modèles, constituent l'ensemble à échanger. Les mêmes raisons qui imposèrent le choix de Liège comme centre d'échange pour JONSDAP 73 ont conduit les membres du groupe JONSMOD à y centraliser l'ensemble commun de données. Les problèmes posés sont en effet fort similaires, puisque le groupe JONSMOD comprend, outre les quatre institutions qui coopérèrent pour JONSDAP 73, des chercheurs des centres suivants :

- Institut Royal Météorologique Danois (Copenhague),
- Institut für Meereskunde (Hambourg),
- Service Hydrographique et Océanographique de la Marine (Paris).

Les données sont de nature moins sophistiquée que dans le cas précédent, mais le nombre de partenaires (donc de systèmes) est plus élévé. La complexité des échanges est donc fort similaire. On va dans les sous-paragraphes qui suivent, détailler les problèmes qui ont surgi à chaque étape de l'échange.

3.2.1.- Détermination des restrictions sur les supports de données

Le premier souci est d'essayer d'obtenir de chacun une liste exhaustive des possibilités et contraintes concernant leur matériel informatique. Après de nombreux rappels et recherches, on est en mesure de connaître au moins un type de données standard pour chaque centre concerné :

- Lowestoft	bandes magnétiques	code ICL
- Bidston	bandes magnétiques	code EBCDIC
- K.N.M.I.	rubans perforés	MC flexowriter code
- Paris	bandes magnétiques	code EBCDIC
- Hambourg	rubans perforés	code ISO (ASCII)
- Copenhague	rubans perforés	code ISO (ASCII) .

3.2.2.- Les formats

Si la gamme des machines et des codes de représentation est très étendue, les formats eux aussi sont très variables, chaque centre de recherche construisant le format propre à ses besoins et ses applications sans tenir compte d'un éventuel échange de données. De plus, certaines limitations du matériel informatique ne laissent aux utilisateurs que très peu de liberté quant au choix du support et de la disposition des données sur ce support. C'est ainsi que l'on rencontre des enregistrements sur bandes magnétiques et rubans perforés avec ou sans format, de longueurs fixe ou variable. Dans de semblables conditions, on peut s'attendre à recevoir n'importe quoi, sous n'importe quelle forme.

3.2.3.- Traitement du support de données

1) Lecture

Pour effectuer le traitement d'un support quelconque de données, il est indispensable de connaître un certain nombre de renseignements techniques. Lors des premiers travaux, il fut souvent nécessaire de réclamer un complément d'informations ou d'effectuer soi-même certains tests pour vérifier la validité des renseignements donnés. Parfois,

il est indispensable de déterminer soi-même des détails manquants. Les informations nécessaires pour lire les fichiers concernent la méthode d'enregistrement, le code de représentation accompagné éventuellement d'une table d'équivalence et enfin quelques détails techniques tels que : longueur de l'enregistrement physique, densité d'enregistrement, parité, etc. Toutes ces caractéristiques permettent d'effectuer une première lecture du fichier déterminant son contenu. Chaque institution employant des formats différents, il faut concevoir à l'entrée et à la sortie, des chaînes de programmes particulières à chaque application.

2) Traitement de rubans perforés

Les données reçues sur rubans perforés et restaurées sous cette même forme nécessitent un travail supplémentaire. En effet, le Centre de Calcul de l'université de Liège ne dispose pas d'un lecteur de rubans perforés directement relié à l'ordinateur. Il faut donc au préalable passer par un système d'encodage sur bande magnétique. Un encodeur lit l'information et la recopie sur une bande magnétique selon un format fixe comprenant 100 caractères. A la sortie, la procédure s'effectue en sens inverse : il faut créer un fichier sur bande magnétique avec des enregistrements de 100 caractères afin de perforer le ruban par la suite.

Cette opération d'encodage est très délicate. Elle nécessite beaucoup de manipulations de la part des opérateurs et du programmeur qui prépare le ruban perforé. En effet, pour être traité, ce ruban doit répondre à certaines exigences :

- il doit comporter des amorces au début et à la fin,
- il doit être enroulé dans un certain sens avec indication de début et fin. Il est de plus conseillé de ne pas utiliser des rubans d'une longueur supérieure à 300 mètres.

La plupart des rubans reçus ne correspondent pas à ces normes. Des vérifications minutieuses s'imposent afin de remédier aux imperfections. Les erreurs demandent de longues corrections manuelles et alourdissent considérablement le travail d'échange de données.

3) Décodage de l'information

L'information transférée des rubans encodés sur bande magnétique ou reçue directement sur bande magnétique est maintenant traitable par l'ordinateur. Les données sont prises en charge par un programme de décodage. Ce programme crée un fichier reprenant en EBCDIC tous les caractères bons ou mauvais. A chaque appel du sous-programme décodeur un enregistrement est transcodé. On obtient une chaîne de caractères EBCDIC correspondant aux caractères perforés sur le ruban. L'enregistrement peut avoir une longueur variable limitée toutefois à 100 caractères. Ce transcodage est effectué au moyen d'une table d'équivalence à deux entrées. Chaque caractère y est situé à une place correspondant à sa valeur numérique dans le premier code. On calcule donc automatiquement la valeur décimale des perforations de tous les caractères à décoder, et on obtient chaque fois un numéro de case de la table. On lit le contenu de cette case de la table, c'est le caractère EBCDIC correspondant. Lorsque toute l'information est traitée de cette façon, on obtient un fichier brut décodé.

4) Toilette des informations

Si les fichiers reçus sur bandes magnétiques correspondent généralement à la description fournie, il n'en est pas de même pour les fichiers sur rubans perforés. Ils contiennent parfois des caractères totalement étrangers aux données. Par exemple, il arrive de constater la présence d'une multitude de signes négatifs devant une donnée ou encore un décalage de l'information par rapport au format annoncé. Il s'est avéré nécessaire de mettre au point des programmes de nettoyage de données créant un fichier correct. Les enregistrements de celui-ci correspondent alors tous aux caractéristiques annoncées par l'expéditeur. Ce travail de nettoyage des données est très fastidieux car chaque cas doit être traité séparément, les erreurs étant distribuées de façon tout à fait aléatoire. Il consiste principalement à :

- éliminer tous les enregistrements composés de caractères spéciaux ne concernant pas les données;

- éliminer tous les caractères erronés se situant à l'intérieur d'enregistrements normaux;
- insérer ou supprimer des blancs en cas de décalage;
- corriger des inversions de chiffres.

Le fichier correct peut alors être traité et traduit dans les formats des divers organismes auxquels il est destiné.

5) Fourniture des données

Cette recodification s'effectue suivant un processus inverse. On constitue une table de l'alphabet avec les équivalents binaires dans le code désiré. On crée alors un fichier où chaque caractère est remplacé par son équivalent binaire dans le code de sortie, tout en respectant les normes de présentation de chacun des organismes. La fourniture des données peut, elle aussi, réclamer la copie du fichier sur ruban perforé.

6) Les aléas de l'échange des données

Le transcodage peut se compliquer légèrement lorsque la longueur des mots-machine varie d'un ordinateur à un autre. C'est le cas lorsqu'on reçoit des données provenant d'un ordinateur ICL, ou que l'on doit fournir des informations à un centre équipé de ce matériel. Cette difficulté a déjà été traitée précédemment.

Certaines difficultés supplémentaires peuvent surgir lorsque, d'un envoi à un autre, la structure de l'information d'un même organisme est modifiée. L'ensemble des programmes constitués pour transcoder les données de ce centre doit être alors totalement repensé. Le cas se présente lors du passage d'une information structurée à une suite de données séparées les unes des autres par un ou plusieurs blancs. Aucune spécification de longueur des blocs d'information n'est indiquée. A ce moment, de nouveaux programmes doivent isoler les données et reconstituer un fichier structuré.

D'autres problèmes peuvent encore se présenter si certaines données sont imprimées ou fournies sur graphique. Le traitement de telles informations entraîne toujours un surcroît de travail fastidieux.

3.2.4.- Acquis actuel de l'échange des données JONSMOD

Les premiers échanges de données effectués au cours de la campagne JONSMOD ont déjà permis à plusieurs centres d'utiliser, dans le cadre de leurs recherches, une masse de données jusqu'alors inaccessibles. Cette distribution de données est destinée à s'accentuer et à sortir du cadre de JONSMOD. Toutefois, cette opération JONSMOD a permis de tirer quelques enseignements précieux.

Le support le plus maniable reste la bande magnétique pour autant que la description de l'information enregistrée soit suffisante pour en permettre la lecture. Presque tous les autres supports sont acceptables mais nécessitent trop souvent des traitements supplémentaires, longs et coûteux.

De plus, devant la grande variété des formats employés par tous les centres concernés, il est nécessaire de discipliner et standardiser le format des données échangées. Cette standardisation permet à chacun de pouvoir utiliser directement les données reçues dans son logiciel de traitement. Plusieurs structures de format autodescriptif sont proposées. L'une de celles-ci prend de plus en plus d'extension et est déjà développée dans certains grands centres océanologiques. Il s'agit du format GF2 (anciennement GATE) dont il sera question plus loin.

Il est en tous cas indispensable de poursuivre un échange de données plus intensif afin de fournir aux chercheurs une banque de données plus complète, et de continuer le développement d'un format généralisé afin d'accélérer et faciliter cet échange de données.

Chapitre III

Exploitation des données

1.- Construction de la banque de données et du logiciel d'exploitation

Une base de données devient une banque de données à partir du moment où les données sont structurées de manière telle qu'on puisse écrire un logiciel capable de les rechercher, les analyser, les traiter; ce logiciel doit être de maniement aisé et commun à tous les types de données. La structure proposée ici n'a pas la prétention d'être accessible de manière transparente pour les utilisateurs, tel que certains systèmes utilisés dans le domaine commercial ou bancaire. Le logiciel ne sera pas étendu aussi loin pour deux raisons essentielles :

1) les données scientifiques (telles que les données océanographiques et météorologiques) sont de types et de structures beaucoup plus variées que les données commerciales, et nécessitent une exploitation beaucoup plus complexe. Il est absurde d'établir un logiciel très sophistiqué, qui augmenterait considérablement le temps de calcul alors que l'on est limité en moyens matériels et humains.

2) la banque de données et le logiciel d'exploitation sont créés dans le but d'être utilisés par des programmeurs qualifiés et par des scientifiques qui ont un minimum d'expérience en programmation et qui sont supposés connaître le type d'information qui leur est nécessaire et la manière dont ils veulent traiter les données. Une petite équipe de spécialistes du logiciel est cependant nécessaire pour aider les scientifiques à assembler les chaînes de programmes et pour les conseiller sur le choix des techniques de calcul.

Cette solution a déjà été expérimentée dans plusieurs centres de recherche océanographique, comme la WHOI, l'université de Southampton et le BNDO au Centre Océanologique de Bretagne.

Le but de ce projet est de construire un système suffisamment flexible pour permettre le stockage, l'archivage, la sélection et l'analyse des jeux de données présents et futurs, quelles que soient leur origine et leur structure. Le système est destiné :

- à faciliter le travail des scientifiques en leur donnant des outils standards
  1) pour l'analyse de leur matériel,
  2) pour la validation et la comparaison des modèles mathématiques;
- à être étendu facilement par des programmes aux fonctions spécifiques.

Finalement, le stockage des données et le logiciel doivent être portables (c'est-à-dire capables d'être utilisables avec peu de modifications sur la plupart des ordinateurs) afin que le plus grand nombre de centres de recherche bénéficie du travail d'élaboration. Il est donc essentiel que la structure des programmes et des enregistrements de données soit compatible avec la grande majorité des ordinateurs utilisés dans les centres.

2.- Conditions de portabilité du format de stockage et du logiciel

Les spécialistes en programmation savent parfaitement que toute information écrite par un ordinateur peut être lue par un autre, pourvu que le logiciel de translation existe.

La portabilité du logiciel est moins évidente. Tous les langages ne sont pas utilisables sur tous les ordinateurs, et la translation d'un langage à un autre demande souvent beaucoup de temps et d'argent. En général, les centres de recherche océanographique ne disposent pas des moyens nécessaires pour une telle conversion. C'est pourquoi le logiciel développé dans ce projet de recherche est destiné à être utilisé avec seulement quelques simples modifications relatives aux

opérations d'entrée/sortie, sur la plupart des ordinateurs scientifiques.

En raison de la structure interne des programmes, de la flexibilité du format de stockage et de la totale impossibilité d'écrire des programmes d'exploitation de données absolument indépendants du matériel (*hardware*), il est nécessaire que le matériel et le logiciel présentent les caractéristiques suivantes :

1) un accès direct entre la mémoire centrale et les bandes magnétiques et disques est possible,

2) un mot réel en simple précision est stocké en deux mots entiers,

3) un compilateur pour ANSI FORTRAN IV existe,

4) des enregistrements de différentes longueurs peuvent être lus et écrits sur bande magnétique (éventuellement, au moyen d'un ensemble restreint de programmes en langage spécial d'assemblage).

ANSI FORTRAN IV est le langage de programmation le plus répandu pour des applications scientifiques et techniques. C'est donc le meilleur langage pour écrire une bibliothèque scientifique.

La seconde condition ci-dessus, signifie qu'un nombre en virgule flottante en simple précision est représenté en mémoire sous la forme de deux mots entiers. Ceci se fait automatiquement sur des ordinateurs dont la structure est basée sur l'utilisation des mots de 16 bits ou 24 bits et ceci peut être obtenu au moyen de déclarations (instructions) spéciales sur des ordinateurs dont la structure est basée sur l'utilisation de caractères de 8 bits ou de mots de 60 bits.

Le format de la bande magnétique est portable en ce sens que la structure des enregistrements sur bande magnétique est la même pour tous les ordinateurs. (Les principes énoncés ici sont valables pour le traitement des données et le format de stockage et non pour le format d'échange. Ce dernier sera développé plus loin.) Ainsi, on utilise le mode d'enregistrement binaire pour la concision et la performance des entrées/sorties, c'est-à-dire que les données sur bande magnétique sont une copie exacte de leur représentation en mémoire centrale. Un format d'échange de données sera écrit dans un code caractère international, comme EBCDIC ou ASCII. Un tel code ne convient pas pour une analyse

rapide des données et est beaucoup moins compact que le code binaire. Mais en dépit du mode d'enregistrement binaire, les bandes écrites par un ordinateur peuvent être lues facilement par un autre si les conditions suivantes sont satisfaites :

1) les spécifications des bandes magnétiques sont compatibles sur les deux ordinateurs (c'est-à-dire, même nombre de pistes, même densité, même technique d'encodage, même marque de fin d'enregistrement et de fin de fichier).

2) la structure des deux ordinateurs est basée sur l'utilisation de mots de 16 bits ou de caractères de 8 bits; la structure des deux ordinateurs est basée sur l'utilisation de mots de 24 bits; la structure des deux ordinateurs est basée sur l'utilisation de mots de 60 bits (les ordinateurs sont alors appelés ordinateurs de structure compatible au niveau du mot ou du caractère).

3) des programmes existent pour la translation des codes-caractère et des codes internes.

Des dispositions sont prises au niveau du format de stockage pour définir le code interne utilisé dans l'écriture des nombres réels et des caractères, de sorte que les programmes appropriés (condition 3) puissent être sélectionnés. Les clefs définissant ces codes peuvent être décodées par tout ordinateur de structure compatible avec l'ordinateur source, parce que les nombres entiers positifs ont le même format dans tous les systèmes basés sur la même longueur du mot entier.

3.- Description générale du logiciel

Le logiciel peut être défini comme une extension d'un sous-ensemble de "WHOI Standard buoy format package" [Maltais (1969)] plus simple sous divers aspects. Le logiciel original est modifié dans le but d'être plus général (pour traiter une grande variété de données) et plus portable. Il est étendu pour répondre aux exigences du modèle mathématique. Le logiciel ainsi défini consiste en :

1) Un ensemble de programmes généraux pour lire et écrire les enregistrements de données, pour étiqueter et retrouver les fichiers de

données (sur bandes magnétiques et sur disques) et pour transférer les données entre les mémoires de masse et la mémoire centrale (dans des tableaux accessibles aux utilisateurs) et vice-versa.

2) Un ensemble de programmes de traitement de données (utilisant les programmes généraux) pour sélectionner, éditer, imprimer et tracer les données de façons différentes selon leur structure.

3) Un ensemble de programmes mathématiques :

- pour analyser les données,
- pour comparer et adapter les modèles aux résultats expérimentaux.

4.- Stockage des données en format standard

Le format se décrit par lui-même : chaque fichier de données contient toutes les informations nécessaires pour être lu correctement quelque soit le nombre de variables, le nombre d'échantillons, leur nature, leur structure (fonction continue d'une autre variable ou non), leur mode d'enregistrement, et le nombre d'enregistrements sur le fichier. Les bandes magnétiques écrites en format standard contiennent également, au début des données, un fichier d'identification et, après le dernier fichier de données, un fichier de fin de données. Chaque fichier de données consiste en :

1) deux enregistrements d'étiquette,

2) des enregistrements de données (en nombre quelconque).

L'enregistrement de la première étiquette a une longueur fixée et définit :

- les clefs pour la conversion des données binaires (tableau 2),
- la date de création du fichier,
- le nom du fichier,
- le nombre de variables,
- le nombre d'échantillons,
- la date des mesures,
- le type de la seconde étiquette.

Tableau 2

Etiquette 1

Zone	Contenu de la zone
1	mot clé (entier = 1)
2	Interchange code; convention : 1 → EBCDIC 2 → ASCII (entier) 3 → BCD
3	Internal code; convention : 1 → IBM (entier) 2 → HP
4-11	nom de données (16 bytes) | 4 b. (code mnémonique) 4 b. # croisière ICES 2 b. # code ICES pays 2 b. # code ICES bateau ou bouée 4 b. Histoire des données
12 - 16	époque de la création du fichier : année - jour - heure - minute - seconde (5 entiers)
17 - 18	source des données (4 b.)
19	# variables (entier)
20	# échantillons (entier)
21	Format de la 2e étiquette (entier)
22 - 26	époque initiale des données : année - jour - heure - minute - seconde (5 entiers)
27 - 98	144 b. de commentaires

L'enregistrement de la seconde étiquette définit (tableau 3) :

- la position géographique où les mesures ont été effectuées (latitude, longitude ou autre information);
- les paramètres d'échantillonnage;
- la valeur initiale de la variable continue;
- pour chaque variable :
  - le nom,
  - les unités,
  - le code de l'appareil de mesure,
  - le code de l'instrument,
  - le numéro de série,
  - le type des données à l'enregistrement (entier, réel, synoptique),
  - la profondeur ou la pression,
  - trois attributs.

Tableau 3

Etiquette 2

Zone	Contenu de la zone	
1	mot clé (entier = 2)	
2	# échantillons/intervalle (entier)	
3 - 4	intervalle d'échantillonage (réel)	
5 - 6	intervalle entre échantillons (réel)	
7 - 8 9-10	jour julien } ou { valeur initiale secondes } ou { - 1 } (2 réels)	
11 - 14	8 b. : latitude degrés - minutes - secondes - N , S (4 x 2 b.)	} 24 bytes pour indiquer un autre format
15 - 18	8 b. : longitude idem	
19 - 22	8 b. : variation magnétique	
23 - 28	12 b. : nom de variable	jusqu'à 64 fois cette zone
29 - 34	12 b. : unités dimensionnelles	
35 - 36	4 b. : code mesure	
37 - 38	4 b. : code instrument	
39 - 40	4 b. : série	
41	2 b. : type des données { 'I' entier ; 'R' réel (T,V,R) ; 'S' synoptique }	
42	profondeur (en m) ou pression (mb) ou 0 (entier)	
43 - 44	1er attribut (réel) : valeur de biais	
45 - 46 47 - 48	2ème attribut (réel) } non attribué 3ème attribut (réel)	

Tableau 4

Enregistrement des données

Zone	
1	mot clé (entier = 3)
2 - 9	nom des données (16 b.)
10	# échantillons précédents (entier)
11	# échantillons dans l'enregistrement (entier)
12	pas utilisé
13 - 14	jour julien ou valeur initiale
15 - 1166	données proprement dites

Un enregistrement de données consiste en (tableau 4) :

- le nom du fichier de données (pour contrôler si les données lues sont celles décrites dans l'enregistrement de la première étiquette);
- le nombre d'échantillons dans les enregistrements précédents;
- le nombre d'échantillons dans l'enregistrement courant;
- la valeur initiale de la variable continue;
- les données.

La structure détaillée de chaque type d'enregistrement est décrite dans les tableaux 1 à 3, uniquement pour les ordinateurs dont la structure est basée sur l'utilisation de mots de 16 bits ou 32 bits.

Tous les fichiers existant actuellement dans la banque de données sont en voie d'être mis en forme suivant les spécifications décrites ci-dessus.

5.- Etat actuel du logiciel d'exploitation

5.1.- Mémorisation d'ensembles de données sous une forme standardisée

Vu la nécessité de manipuler de grands ensembles de données, il s'est avéré utile de déterminer une méthode d'archivage. Un groupe de programmes qui permettent de conserver n'importe quel ensemble de données sous une forme standardisée ont été mis au point. Ces programmes rangent les données sur bande magnétique car c'est ce support qui offre le plus d'avantages lorsque l'on traite de grands volumes d'informations. Les données sont répertoriées en une série de fichiers, chaque fichier étant constitué par des données qui ont un caractère commun. La structure d'un fichier est élaborée de manière à être autodescriptive. La définition des données ainsi que tous les renseignements signalétiques propres à chaque variable sont eux-mêmes enregistrés sur le fichier. Ce procédé permet de conserver dans un même fichier un nombre quelconque de variables.

5.2.- Structure des fichiers

Dans le format que l'on a adopté, trois types de fichiers existent.

1) Fichier d'identification de la bande

Il est constitué d'une suite d'enregistrements identiques qui identifient la bande par un numéro de série. Ce fichier est unique sur une bande.

2) Fichier de données

Il contient toutes les informations qui décrivent les données à enregistrer, plus les données proprement dites.

3) Fichier de fin des données

Ce fichier est unique sur la bande et contient un enregistrement spécial qui détermine la fin de la bande.

5.3.- Description synthétique des programmes

1) Programme d'écriture du fichier d'identification de la bande (TAPID)

Ce programme initialise la bande avec des paramètres tels que le numéro d'immatriculation, le jour et l'année de création. A la suite du fichier d'initialisation, il crée un fichier de fin de données.

2) Programme d'écriture du fichier de fin de données (EØD)

Le fichier de fin de données contient la juxtaposition de 24 fois la chaîne de caractères 'EØD'.

3) Programme d'acquisition des paramètres décrivant les données (GLABE)

Au début d'un fichier contenant les données, se trouvent deux enregistrements décrivant les données qui suivent. Le premier enregistrement contient des informations qui concernent le fichier globalement (par exemple le nom du fichier, le nombre de variables, ...). Cet enregistrement est de longueur fixe. Le second enregistrement contient les informations qui définissent les variables (par exemple, le nom de la variable, le type, le code mesure, ...). Cet enregistrement est de longueur variable.

Le programme GLABE initialise les deux enregistrements au moyen de paramètres lus sur des cartes perforées. A l'issue de l'exécution du programme, les deux enregistrements ne sont pas encore écrits sur la bande, mais leurs images sont conservées en mémoire de l'ordinateur dans deux vecteurs.

4) Programme d'écriture des deux enregistrements de tête d'un fichier de données (PLABE)

L'image des deux enregistrements initialisés par GLABE sont effectivement copiés sur la bande magnétique.

5) Programme de recherche d'un fichier déterminé sur la bande (RLABE)

Un nom de fichier et un nombre signifiant la quantité de fichiers à explorer sont fournis au sous-programme. Celui-ci va passer en revue une suite de fichiers jusqu'à ce qu'il ait découvert celui qui est recherché, ou jusqu'à ce que le nombre de fichiers examinés soit égal à celui qui lui a été spécifié. Un code retourné par le programme indique s'il a trouvé le fichier ou si une anomalie existe.

6) Programme d'écriture des données (OUTDA)

L'utilisateur doit ranger ses données par lots de taille arbitraire dans un tableau qui sera pris en charge par le programme. Le programme OUTDA puise les données dans ce tableau et les transfère dans une mémoire tampon interne. Le nombre de cycles ainsi transférés à chaque appel est fourni à OUTDA par un paramètre. Pour mémoriser une variable temps, il faut avoir initialisé l'une des séries avec les secondes dans la journée et le paramètre 'JULIAN' avec le jour julien. Le contenu de la mémoire tampon est écrit sur la bande lorsqu'elle est remplie à son maximum ou lorsque le paramètre 'JULIAN' est modifié.

En effet, avec la structure adoptée, des données réparties chronologiquement sur deux jours ne peuvent figurer dans le même enregistrement.

7) Programme de lecture des données sur la bande (INDA)

L'appel de ce sous-programme provoque la lecture d'un enregistrement de la bande et le transfert des informations qui y sont contenues

dans un vecteur. Les informations y sont structurées d'une manière qui permet leur exploitation directe par l'utilisateur ou par les programmes de sélection.

8) Programme évolué permettant de sélectionner sur la bande en format standard des variables particulières (SHELL)

Il a pour fonctions :

a) de retrouver un fichier déterminé sur une bande,

b) de sélectionner des variables particulières hors du fichier,

c) de n'extraire que les données comprises entre deux limites (limites incluses),

d) de modifier les données à l'aide d'opérations algébriques élémentaires,

e) de transférer les données sélectionnées sur un autre support. Chaque variable est transférée dans un fichier séparé sur disque.

A chaque appel de SHELL, un groupe de données est déplacé dans chaque fichier disque. Un paramètre détermine la manière dont il faut échantillonner les données :

- HOW=0 : Toutes les données sont transférées par un seul appel à SHELL.

- HOW=1 : Les données sont transférées par une suite de blocs correspondant à un intervalle de temps d'échantillonnage. Il faut que les données soient équidistantes dans le temps. L'intervalle d'échantillonnage est enregistré dans l'étiquette # 2.

- HOW=2 : Les données sont transmises par blocs de longueur arbitraire. Si la variable continue est le temps, le paramètre SAMP fixe la durée d'échantillonnage sur la bande. Si la variable continue n'est pas le temps, SAMP fixe le nombre de données qu'il faut regrouper à chaque appel.

- HOW=3 : Les données sont transmises en une suite de blocs qui correspondent à la taille du bloc sur la bande.

- HOW=4 : Les données sont transmises par blocs correspondant à l'intervalle de temps d'échantillonnage enregistré dans l'étiquette # 2 du

fichier. Dans ce cas-ci, les données enregistrées ne sont pas équidistantes dans le temps.

9) Programme de lecture des paramètres sur la carte standard (SICC)

Le programme SHELL a besoin pour son fonctionnement d'un ensemble de paramètres. Certains sont obligatoires, d'autres sont facultatifs. Les paramètres obligatoires sont obtenus par le programme SICC. Ces paramètres sont :

- le nom du fichier à traiter,
- le numéro de la variable continue; si la variable continue est le temps, il faut mettre 0,
- la période de temps exprimée en jours juliens et secondes qui détermine l'intervalle sur lequel les données seront échantillonnées sur la bande; ou bien le nombre de données à regrouper pour former un bloc de données sur les nouveaux fichiers disques,
- l'endroit du fichier à partir duquel il faut extraire les données, exprimé soit sous la forme de jours juliens et secondes, soit sous la forme d'un numéro de séquence,
- la limite supérieure des données à extraire, au-delà de laquelle il ne faut plus échantillonner. Le programme signale qu'elle est atteinte au moyen d'un code.

5.4.- Travaux en cours

On procède actuellement à la transformation de la base de données (données météorologiques, données de la bouée, données des courantomètres, données des expériences communes) en une banque de données exploitable à l'aide du logiciel qui vient d'être décrit. Ce logiciel est par ailleurs en développement constant.

6.- Référence

MALTAIS, J. A nine channel digital magnetic tape format for storing oceanographic data, Woods Hole Oceanographic Institution, Ref. 69-55. Unpublished manuscript.

Chapitre IV

Echange des données

1.- Introduction

Une autre fonction du centre d'exploitation des données est de fournir ces données à des institutions extérieures qui coopèrent avec les chercheurs du Programme dans le cadre d'expériences communes. Ces expériences sont destinées à nourrir les modèles mathématiques développés par différents groupes de recherche; le volume de données qu'elles fournissent impose que celles-ci soient échangées à l'aide de supports d'information directement traitables par ordinateur et portables d'ordinateur à ordinateur.

L'expérience de plusieurs années a montré les difficultés inhérentes à l'échange d'informations entre systèmes différents (voir le chapitre II). Pour éviter de semblables problèmes lors d'expériences futures, il est nécessaire de définir un format flexible et standardisé pour l'échange international de données entre différents systèmes de traitement. Ce format n'est pas destiné à traiter ou analyser des informations, mais à les transférer facilement d'un ordinateur à l'autre. Le but principal est donc de développer *un format standard pour l'échange international des données* d'un centre à un autre, tel qu'il soit utilisable par chacun des chercheurs dans tous les centres traitant les données.

L'apparition du format de données GATE pour l'échange international adopté par la COI, avec quelques modifications, sous le nom GF2 (*Generalized format 2*) a résolu une partie du problème. Cependant, peu

de centres océanographiques disposent du logiciel nécessaire pour échanger des données en GF2. Il apparaît donc du plus haut intérêt d'étudier un logiciel et ses possibilités d'adaptation aux différents systèmes existants. Le travail est considérable car le GF2 n'a jamais été utilisé pour l'échange de données océanographiques de nature chimique ou biologique, et la possibilité de l'adaptation de telles données au GF2 doit être étudiée de manière approfondie. Cependant une étude préliminaire a montré que la souplesse des structures préconisées est telle qu'il ne semble pas y avoir de difficultés de principe à l'incorporation de données biologiques, chimiques ou géologiques à un système d'échange basé sur GF2.

2.- Description générale du GF2

Les données peuvent être écrites ou retrouvées à partir du format au moyen de programmes très simples et assez rapides. Par ailleurs, la structure du format (enregistrements, fichiers, code et format d'écriture, ...) est destinée à être entièrement auto-descriptive et peut être entièrement automatisée si on le désire.

Les analogies entre le GF2 et le format de stockage précédemment décrit sont nombreuses. En fait, tous deux sont issus du même format, [Maltais (1969)] mais ont évolués différemment car ils ont été développés dans des buts différents. En effet, le format de stockage n'est utilisé que sur un seul ordinateur à la fois et est conçu pour être lu et écrit rapidement; ceci impose l'utilisation de code interne et l'écriture en enregistrements de longueur variable; le GF2 doit servir de lien entre ordinateurs différents; l'utilisation d'une représentation en code caractère et d'enregistrements de longueur fixe s'impose au prix d'une manipulation plus lente et plus lourde; tous deux ont en commun leur grande souplesse d'utilisation, leur capacité de s'adapter à n'importe quel type de données et leur caractère totalement auto-descriptif.

2.1.-

La structure de la bande consiste en six types d'enregistrements physiques séparés par des marques d'enregistrement et bloqués en fichiers (usuellement séparés par une marque de fin de fichier 'EOF', quelque fois appelée marque de bande), comme suit :

"fichier test"
EOF
enregistrement de début de bande
EOF

enregistrement "en-tête de fichier" type 1	un enregistrement en-tête de fichier type 2 est nécessaire pour chaque groupe de 12 paramètres dans un cycle de données, c'est-à-dire 1 pour moins de 12 paramètres, 2 pour moins de 24, 3 pour moins de 36, etc.
enregistrement "en-tête de fichier" type 2	
enregistrement "en-tête de fichier" type 2	

enregistrement de données
⋮
enregistrement de données
EOF

enregistrement "en-tête de fichier" type 1	un enregistrement en-tête de fichier type 2 est nécessaire pour chaque groupe de 12 paramètres dans un cycle de données, c'est-à-dire 1 pour moins de 12 paramètres, 2 pour moins de 24, 3 pour moins de 36, etc.
enregistrement "en-tête de fichier" type 2	
enregistrement "en-tête de fichier" type 2	

enregistrement de données ⋮ enregistrement de données	autant que nécessaire

EOF
enregistrement de fin de bande

EOF EOF	le double EOF signifie la fin des données sur cette bande

Deux marques de fin de fichier 'EOF' à la fin des informations sur chaque bande, même dans les fichiers à bandes multiples, sont obligatoires. D'autres usages de 'EOF', excepté pour la séparation des fichiers et après l'enregistrement début de bande, sont interdits.

2.2.-

Tous les enregistrements physiques doivent être de longueur constante avec exactement 1920 caractères (24 groupes de 80 caractères) complétés par des blancs ou des '9' si nécessaire pour obtenir la longueur imposée. Les enregistrements de début de bande et d'en-têtes sont bloqués par groupes de 80 caractères de sorte que l'information nécessaire à la description de la bande et des en-têtes peut être facilement copiée à partir de (ou sur) cartes perforées si l'on le désire.

2.3.-

Le type d'enregistrement de la bande magnétique doit être identifié par son premier caractère :

fichier test [1],
(0) enregistrement de début de bande,
(1) enregistrement "en-tête de fichier" type 1,
(2) enregistrement "en-tête de fichier" type 2,
(3) enregistrement de données,
(4) enregistrement de fin de bande.

Le sommaire du contenu de chacun de ces enregistrements est donné ci-dessous.

---

1. Le fichier test se reconnaît directement à son contenu, par son premier caractère, octal 77 (BCD), hexadécimal FF (EBCDIC) ou 11111111 (binaire).

Type	Nom	Contenu
"Fichier test"		Un enregistrement test consiste en 1920 caractères - test. Un caractère test est défini comme une suite de "1", l'octal "77" en BCD, l'hexadécimal "FF" en EBCDIC. C'est pourquoi le fichier test consiste en un nombre suffisant d'enregistrements test, composé chacun de 1920 caractères, pour remplir environ 20 mètres de la bande (protection contre une mise en place défectueuse).
0	En-tête de la bande	- Identification de la bande, nom du pays et de l'Institution écrivant la bande et information de la succession des bandes. - Date/heure de création de la bande et type d'ordinateur utilisé. - Tables de translation (codes des caractères) utilisées par l'ordinateur écrivant la bande. - Zone non attribuée, pouvant être utilisée pour des remarques ou explications en langage clair.
1	"En-tête de fichier" type 1	- Détails de la provenance des données (pays et institution) dans le fichier et nom du fichier. - Information descriptive au sujet de la plateforme primaire qui a observé des données se trouvant dans ce fichier. - Informations spéciales pour le cas où une plateforme secondaire supporte une plateforme primaire (pour un système de bouée, la bouée est la plateforme primaire tandis que le bateau effectuant le mouillage est la plateforme secondaire). - Position, époque et valeur de la variation magnétique au début de la période d'observation ou de la station. - Mêmes informations mais à la fin de la période d'observation ou de la station (incluant un code pour résumer la validité des résultats pour tout le fichier).

- Information sur l'environnement (météorologique et océanographique) des stations d'observation (utilisation optionnelle).
- Information sur le nombre de paramètres répétés une seule fois par enregistrement, nombre de paramètres par cycle de données, intervalle entre les échantillons, nombre de cycles de données dans le fichier, et indication de l'utilisation de nombres entiers ou réels (ou les deux) et de caractères alpha-numériques comme valeurs de paramètres.
- Spécifications détaillées du format pour l'enregistrement des 1920 caractères des enregistrements de données. Zone libre mais utilisable pour des remarques ou explications en langage clair.

2 "En-tête de fichier" type 2

- Spécifications du paramètre # 1 dans le cycle de données incluant un code résumant la validité des résultats pour le paramètre # 1 (pour tout le fichier) et définition d'un des deux attributs pour le paramètre # 1.
- Définition du second attribut et documentation sur la méthode de mesure et de collection du paramètre # 1.
- Idem pour les paramètres 2-12.

(Pour les cycles de données de 13 → 24 paramètres, on a besoin d'un deuxième enregistrement d'"en-tête de fichier" type 2, de 25 → 36, un troisième, etc.)

3 Enregistrement de données

- Nombre de cycles de données dans cet enregistrement, nombre de cycles de données précédant cet enregistrement et compteur d'enregistrement de données.
  (Note : Ces nombres *ne sont pas* les paramètres définis au point 7 de la description de l'en-tête de fichier type 1)
  Le paramètre d'enregistrement appartient aux données observées, et pas aux enregistrements de données eux-mêmes. Les caractères 1-20 de cet

		enregistrement contiennent des explications pour la lecture des données sur la bande et pas pour l'identification des observations. - Cycles de données analysées.
4	Enregistrement fin de bande 'EOT'	- Nom ou numéro de la bande qui suit. - Caractères tous mis à '9' pour que cet enregistrement puisse être lu avec la même instruction de format que l'enregistrement en tête de fichier type 1. - Zone libre mais utilisable pour des commentaires.

2.4.-

Si un fichier est trop long pour une seule bande magnétique, il peut être continué sur d'autres bandes. L'enregistrement début de bande procure une information indiquant si le fichier est la suite d'un fichier se trouvant sur la bande précédente et l'enregistrement de fin de bande montre si le fichier se prolonge sur une nouvelle bande. L'enregistrement 'EOF' après l'enregistrement en début de bande se trouve sur toutes les bandes, mais les enregistrements "en-tête de fichier" ne sont pas répétés sur des bandes suites.

Chaque fichier de données est entièrement auto-décrit par des enregistrements "en-tête de fichier" placés au début de celui-ci. En particulier, l'enregistrement "en-tête de fichier" type 1 doit contenir des détails de format sur tous les enregistrements de données (en Fortran, ceci serait une instruction du format utilisé pour écrire un enregistrement de données). L'enregistrement en-tête de fichier type 2 contient des informations sur chacun des paramètres du jeu de données. Un fichier peut être défini comme étant toutes les observations individuelles qui constituent un ou des jeux de données.

Exemples

* Chaque profil STD, si l'information dans les enregistrements de tête change pour chaque profil.

* Tous les profils STD reçus d'un bateau pour une période de temps donnée si l'information contenue dans les enregistrements de tête est toujours la même, et si les enregistrements de données contiennent toutes les informations qui changent à chaque profil (c'est-à-dire, latitude, longitude, temps, etc.).

* Un bref aperçu de la température à la surface de la mer en provenance d'observations aériennes.

* Une série temporelle qui provient d'un courantomètre.

2.5.-

Chaque enregistrement de données consiste en un nombre entier de cycles de données (c'est-à-dire un cycle de données ne peut pas chevaucher deux enregistrements physiques). Par exemple, tous les instruments d'échantillonnage prennent des échantillons selon des séquences régulièrement répétées; la répétition d'une séquence de base constitue un cycle de données.

Exemples

* (profondeur ou pression et température) à partir d'un XBT ou MBT.

* (temps, pression, température, humidité, élévation, vents) à partir d'une sonde atmosphérique.

* (temps, direction, direction, direction, vitesse) à partir d'un instrument de mesure pour les courants océaniques (Ceci arrive quand l'instrument de mesure échantillonne un paramètre beaucoup plus souvent qu'un autre; ici, c'est le cas de la direction du courant).

* (latitude, longitude, temps, profondeur, température, salinité) permettant d'établir une carte horizontale de température et de salinité.

Les paramètres peuvent ou non être tous échantillonnés simultanément. Si l'échantillonnage est séquentiel, comme dans l'exemple du courantomètre, à chaque paramètre est associé un déphasage par rapport

au temps spécifié dans le cycle de données et ces déphasages doivent être spécifiés dans l'enregistrement en-tête du fichier type 2. En général, la pression ou quelqu'autre variable peut être la variable indépendante, et cette variable doit être le premier élément du cycle de données. Les déphasages spécifiés dans l'enregistrement en-tête du fichier type 2 ont alors les mêmes unités que la variable indépendante. Si les déphasages sont tous nuls, le premier paramètre dans un cycle de données ne doit pas être nécessairement une variable indépendante.

2.6.-

En Fortran, il est recommandé d'utiliser uniquement les formats A, I, F et X. Les formats E et D ne sont pas toujours compatibles entre les différentes installations. Les facteurs d'échelle des paramètres (voir enregistrement en-tête du fichier type 2) doivent être utilisés pour éviter les formats E et D et pour faciliter la conversion des unités d'enregistrement aux unités communes.

2.7.- Il faut bien veiller à ce que les longueurs de mots des ordinateurs utilisés soient suffisantes pour la résolution que nécessite chaque paramètre. De nouveau, l'usage des facteurs d'échelle peut résoudre ce problème.

Conclusion

L'équipe chargée de la compilation et du traitement des données a rassemblé une importante base de données, dont la plupart proviennent d'instruments automatiques. Elle développe le logiciel nécessaire pour exploiter les données et les échanger avec d'autres équipes de recherche dans le cadre d'expériences communes. A cet égard, cette équipe a été choisie par les responsables du projet INOUT (partie de JONSDAP 76), sur base de l'expérience et de la compétence qu'elle a acquise dans ce domaine, pour réaliser la circulation des informations de ce projet, dont le centre de données (INOUT Data Center) sera pour lors situé à Liège. Le format GF2 a été proposé pour simplifier les problèmes de l'échange, et le logiciel adéquat pour l'utilisation pratique de ce format sera également réalisé à Liège.