

## Document d'information 8

Original : anglais

## Catalogage et archivage de données



## CONTEXTE

1. Dans les conclusions de leur dixième Conférence (points 54 à 56), les directeurs des pêches se sont déclarés préoccupés par la question des données historiques auxquelles ils n'ont plus accès.
2. Il a été suggéré que la Communauté du Pacifique (CPS) recherche ces données historiques au nom des pays océaniques et joue le rôle de centre d'archivage pour toutes formes de données sur la pêche côtière et l'aquaculture, et pour les informations relatives aux pêcheries en général, afin que soit créée une sauvegarde des systèmes nationaux. Enfin, la CPS doit convertir les données dans un format exploitable par les pays océaniques à l'avenir.
3. La CPS conserve déjà des données au nom des pays océaniques, essentiellement pour des projets au cours desquels son personnel a participé à la collecte et/ou à l'analyse des données, mais elle détient également des captures et des fichiers de bases de données occasionnellement fournies par ses partenaires et interlocuteurs (voir annexe A ci-dessous). Dans leurs conclusions, les directeurs des pêches recommandent que la CPS poursuive cette activité à grande échelle, tout en s'assurant que les données sont exploitables.
4. Le point 55, notamment, traite de manière implicite de l'archivage et de la conservation, mais également de l'organisation de contenu et du partage des ensembles de données, voire de la constitution d'un entrepôt de données à partir de ces ensembles. Ces tâches visent des objectifs très différents, mettent en œuvre des solutions techniques diverses et impliquent des coûts et des ressources humaines variables.

## Archivage et conservation numériques

5. L'archivage numérique consiste à conserver des fichiers sur un système où ils peuvent être consultés ultérieurement par les personnes autorisées. En dupliquant le matériel et les lieux de stockage, cette méthode réduit les risques de perte de données due à une défaillance du matériel, à une erreur humaine, à une activité criminelle ou à une catastrophe.
6. Certains risques peuvent être réduits grâce à la redondance des matériels et à une stratégie de sauvegarde. Par exemple, les microserveurs installés localement utilisent une rotation de sauvegarde pour les bases de données SQL, une copie des disques durs ainsi qu'un disque dur externe distinct pour les sauvegardes de mises à jour du système. Cette installation permet d'éviter la perte de données en cas de panne d'un disque dur, et réduit la perte de données en cas de suppression ou de corruption de celles-ci.
7. Une sécurité accrue nécessite généralement des sauvegardes hors site, à condition que la bande passante nécessaire soit disponible et que le service de stockage soit payé. Le coût du stockage de données en ligne varie en fonction de l'emplacement des serveurs, de la situation locale ou régionale du stockage redondant, de la disponibilité requise et de la vitesse d'accès. Ainsi, les données auxquelles on accède rarement – telles que les sauvegardes – peuvent être conservées dans des systèmes de stockage « à froid » pour un coût relativement faible. Toutefois, le stockage représente un coût récurrent qui doit être provisionné jusqu'à la suppression finale des données.

8. Les fichiers archivés doivent être stockés en mode « lecture seule » afin d'éviter toute modification ou corruption accidentelle. En effet, si des données sont ajoutées ou modifiées par la suite, une nouvelle version du fichier sera créée, ce qui augmente les coûts de stockage et l'espace nécessaire.
9. En outre, une actualisation et une migration régulières sont nécessaires afin de s'assurer que les données sont transférées vers de nouveaux supports ou un nouveau système en cas de risque d'obsolescence matérielle ou de détérioration du support physique (ainsi, les cassettes, les disquettes, les disques ZIP, les CD et les DVD gravés ont tous une durée de vie limitée). Le fichier doit parfois être converti dans un autre format afin que les données restent exploitables.
10. L'archivage numérique assure l'intégrité et l'accessibilité des fichiers, mais ne garantit pas la qualité et l'exploitabilité des données. Par exemple, un tableur Excel contenant des données sur la longueur/le poids de poissons n'est pas exploitable à long terme si les unités, le lieu et la date de prélèvement de l'échantillon, le contexte de l'enquête, ou encore la méthode d'échantillonnage ne sont pas précisés. Une base de données ou un fichier peuvent être archivés au nom de son propriétaire, sans être exploitables directement par une personne ne disposant pas de l'application permettant de les consulter, de la documentation ou des connaissances nécessaires concernant les données.
11. Le partage et l'exploitation des données nécessitent souvent la création d'ensembles de données isolés et de métadonnées associées extraites des systèmes d'archivage de données disponibles.

### **Bases de données, ensembles de données et organisation de contenu**

12. Une base de données relationnelle est un ensemble de schémas, de tableaux et de vues qui structure les données dans un système de gestion des données. Un ensemble de données est une série de valeurs contenues dans un seul tableau ou plusieurs tableaux étroitement liés pour un type d'enquête et une période spécifiques, par exemple les données d'une enquête concernant les invertébrés pour un site précis ou les données concernant les captures pour une île et une année précises. Une base de données contient généralement plusieurs ensembles de données pour différentes périodes, différents endroits et différents types d'enquête.
13. En général, on extrait un ensemble de données d'une base de données grâce à une requête qui associe des données issues de tableaux liés. Cet ensemble contient toutes les informations nécessaires qui peuvent ensuite être analysées hors de la base de données. Les métadonnées associées à l'ensemble de données doivent fournir toutes les informations concernant le contexte et la méthodologie, la signification des colonnes, et toute autre indication nécessaire à la compréhension des données, ainsi que tous les renseignements concernant le propriétaire et le contrat de licence de l'utilisateur final.
14. Le terme « organisation de contenu » désigne la gestion active des bases de données. Elle vise à améliorer la qualité des données et l'accès à celles-ci, et à encourager le partage et l'exploitation des données. Lorsqu'un ensemble de données a été évalué et sélectionné pour faire l'objet d'une organisation et d'une conservation à long terme, on procède au nettoyage des données et à la vérification de leur intégrité. On rassemble ensuite les métadonnées, la

documentation, les rapports et toutes les informations indispensables à l'exploitation de l'ensemble de données en dehors de son contexte d'origine.

15. L'un des objectifs de l'organisation de contenu est d'assurer l'uniformité entre les bases de données. Ainsi, si le nom d'une espèce change, ou si elle est divisée en plusieurs sous-espèces d'après la génétique des populations, l'ensemble de données d'origine doit être modifié pour tenir compte de ces évolutions. Autre exemple : l'évolution des frontières politiques, la modification des zones de recensement ou des toponymes.

## Entrepôt de données

16. Un entrepôt de données est un système qui rassemble des données, parfois hétérogènes, issues de plusieurs sources afin que les utilisateurs puissent les analyser, les intégrer à des rapports et les explorer. Cela nécessite une couche d'intégration des données capable d'extraire, de transformer et de charger les données dans un format adapté à l'analyse croisée des ensembles de données. Un entrepôt de données ne remplace pas les bases de données existantes. Il vient s'y ajouter pour faciliter l'exploration des données. Souvent, des outils, des compétences spécifiques et une connaissance intrinsèque de tous les ensembles de données et de leurs limites sont requis pour pouvoir les utiliser.
17. En phase exploratoire, les sources de données sont souvent agrégées et combinées manuellement ou grâce à un programme aux fins d'analyse statistique, ce qui offre davantage de souplesse qu'un tableau multidimensionnel (cube) dans un entrepôt de données.
18. Un entrepôt de données se révèle utile lorsque l'on dispose de nombreuses données issues de différentes sources et qu'il faut les analyser ensemble régulièrement pour les mêmes unités de temps et d'espace. Ainsi, les données d'exportation des pêcheries pourraient être croisées avec les données des douanes (par année, destination, pays et produit), les informations fournies par les enquêtes sur les revenus et les dépenses des ménages pourraient être comparées aux données sur les captures à l'échelle des villages, etc.

## PROBLÉMATIQUES ET PRÉOCCUPATIONS

19. Tous les services des pêches et toutes les ONG ne disposent pas d'un service informatique ou d'administrateurs de systèmes chargés d'effectuer les sauvegardes des disques et des bases de données. La bande passante étant souvent insuffisante pour transférer quotidiennement un grand nombre de fichiers ou un volume important de données, il convient d'adapter la stratégie de sauvegarde afin de trouver un bon équilibre entre le risque de perte, la valeur des données (peuvent-elles faire l'objet d'une nouvelle collecte et combien cela coûterait-il ?) ainsi que le coût du stockage et de la conservation numérique.
20. L'identification et la récupération des ensembles de données historiques nécessiteront la participation des personnes ayant réalisé les enquêtes, saisi les données et effectué les analyses, non seulement pour localiser les données et les rapports, mais également pour évaluer l'ensemble de données et pour produire les métadonnées le concernant dans le cas où on le sélectionnerait pour le stockage à long terme. Parfois, on pense que les données

historiques ont été perdues alors qu'elles sont encore présentes sur un serveur ou un disque dur défaillant, ou qu'il en existe une copie électronique ou papier quelque part.

## SUJETS DE DISCUSSION POTENTIELS

- La sauvegarde des bases et des fichiers de données est-elle suffisante et comment est-elle effectuée ? Quelles difficultés et quels problèmes le pays rencontre-t-il et quelles mesures ont été mises en place pour les contourner ?
- Existe-t-il des ensembles de données et des bases de données historiques contenant des informations dont l'exploitation pourrait se révéler précieuse, et le partage de ces données est-il soumis à des restrictions particulières ?
- Quels sont les critères pour l'évaluation initiale et l'évaluation ultérieure des ensembles de données et qui est chargé de ces évaluations ? Cette question est particulièrement importante lorsque les données sont stockées par un tiers (comme la CPS) en votre nom.

## ANNEXE A : TYPES D'ENSEMBLES DE DONNÉES GÉRÉS PAR LE PROGRAMME PÊCHE CÔTIÈRE DE LA CPS

### **Données des enquêtes sur les invertébrés sous-marins et les poissons (méthodologie de la CPS)**

Enquêtes réalisées par le personnel des pays et de la CPS dans le cadre de PROCFish/C, SciCOFish, et des projets de suivi du secteur de l'aquariophilie et du changement climatique. Lors d'enquêtes réalisées par les pays suivant la même méthodologie, des données supplémentaires sont parfois sauvegardées lorsque l'assistance de la CPS est sollicitée. Les données comprennent des observations de poissons, des densités et une estimation de la biomasse sur des transects de 50 mètres, des comptages d'invertébrés et des densités (diverses méthodes), ainsi que des informations sur le substrat et les habitats.

### **Données socio-économiques (méthodologie PROCFish/C)**

Enquêtes réalisées par le personnel des pays et de la CPS dans le cadre de PROCFish/C. Lors d'enquêtes réalisées suivant la même méthodologie, des données supplémentaires fournies à la CPS sont sauvegardées. Les données sont issues d'entretiens auprès des ménages et des pêcheurs.

### **Données issues des études sur le recouvrement corallien par photo-quadrats, et des enquêtes sur les crabes de cocotier, les herbiers et les mangroves (méthodologie de la CPS)**

Enquêtes réalisées par le personnel des pays et de la CPS dans le cadre de SciCOFish et des projets de suivi du secteur de l'aquariophilie et du changement climatique. Les données sont stockées sur une base de données accessible en ligne.

### **Données sur les exportations**

Données concernant les demandes de permis d'exportation et les expéditions, essentiellement pour les produits destinés à l'aquariophilie (poissons, bécards, coraux d'aquarium). Toutefois,

des données similaires concernant d'autres produits sont parfois stockées sous la forme d'instantanés de bases de données nationales.

### **Données d'enquêtes auprès des pêcheurs et sur les marchés**

Enquêtes réalisées auprès des pêcheurs et sur les marchés par le personnel de la CPS. L'ensemble de données contient des renseignements concernant les prises débarquées et le poisson en vente sur les marchés, ainsi que des informations supplémentaires sur les équipements du bateau, la destination des prises, les coûts, les prix, etc.

Autres données sur les marchés sauvegardées à partir de bases de données nationales.

### **Données sur la pêche artisanale**

Données sur les prises débarquées pour la pêche artisanale saisies dans TUFART, TUFMAN 2 et Tails. Les données concernent les prises débarquées par opération de pêche et sortie en mer. Cet ensemble de données contient également des informations sur les mouillages de DCP pour certains pays.

### **Bases de données nationales**

Instantanés occasionnels de bases de données sur les licences de pêche, sur la qualité de l'eau et d'autres bases de données nationales.