Information Paper No. 8

Original: English

Data holdings and repository

BACKGROUND

1. One of the outcomes of the 10th Heads of Fisheries meeting (in relation to points 54 to 56), is that Heads of Fisheries (HoF) expressed their concern that historical data is no longer available to them.

2. It was proposed that SPC should look for historical data on their behalf and act as a repository for all forms of coastal fisheries and aquaculture data, and fisheries information in general as a back-up of national systems. Finally, SPC should convert data into a usable format for PICTs in the future.

3. SPC already holds data on behalf of countries, mainly for projects where SPC staff members were involved in collection and/or analysis, but also keeps database snapshots and files that are occasionally provided by partners and counterparts (see Annex A, below). The HoF outcomes recommend that SPC continues this activity on a larger scale, along with ensuring that data is reusable.

4. Point 55 in particular refers implicitly to archiving and preservation, but also to digital curation and sharing of datasets, and possibly of integration of datasets into a data warehouse. These tasks are very different in their purpose, technical solutions, costs and required staff time.

Digital archiving and preservation

5. Digital archiving consists of keeping files on a system where they can be retrieved at a later time by authorised people, which mitigates the risk of data loss due to hardware failure, human error, crime or disasters, through hardware and storage location redundancy.

6. Some risks can be mitigated with hardware redundancy and a back-up policy. For example, the microservers installed in-country use a rotating back-up scheme for SQL databases, mirrored hard discs and a separate extra hard drive for incremental system back-ups. Such a setup prevents data loss in the case of a single hard drive crash, and reduces the data loss in the case of data deletion or corruption.

7. Increased security generally requires off-site back-ups, provided that the required bandwidth is available and storage service is paid for. The cost of storing data online varies depending on the location of the servers, whether storage is locally or regionally redundant, and the required availability and the access speed. For example, data that is rarely retrieved – such as back-ups – can be stored in ‘cold’ storage systems for a relatively cheap price. Yet storage is a recurring cost that needs to be provisioned for until the final disposal of the data.

8. Archived files should be stored as read only to avoid accidental modification or corruption. If data is added or modified at a later time then a new version is created, which increases the storage space and cost.

9. In addition to storage, there is a requirement for refreshing and migration at times to ensure that data is transferred to new media or a system if there is a risk of hardware obsolescence or physical media deterioration (e.g. tapes, floppy discs, zip drive, burned CDs and DVDs all

have a limited shelf live). File format conversion might also be needed so that the data remains readable.

10. Digital archiving ensures integrity and readability of files, but not the quality or reusability of the data. For example, an excel spreadsheet with fish length/weight data is not usable in the long-term without knowing units, sample location and date, survey context, sampling method, etc. A database or file can be archived on behalf of its owner, but may not be readily usable by anybody else without front-end application, documentation and intrinsic knowledge of the data content.

11. Sharing and reuse of data often requires the creation of standalone datasets and associated metadata that is disseminated from available data repositories.

Databases vs datasets and digital curation

12. A relational database is a collection of schemas, tables and views that structure data in a database-management system. A dataset is a collection of data, from a single table or closely related tables for a specific survey type and period of time – for example, invertebrate survey for a site or catch data for an island and year. A database is likely to contain several datasets for different periods, locations and survey types.

13. A dataset is generally extracted from a database with a query that combines data from linked tables. It contains all the required information so that it can be analysed independently of the database. The metadata associated with the dataset must provide the information about the survey context and methodology, the meaning of columns, and all other information required for the understanding of the data, as well as information about ownership and end-user licensing.

14. Digital curation is the active management of datasets to improve data access and quality, and to encourage data sharing and reuse. Once a dataset has been appraised and selected for long-term curation and preservation, it goes through data cleansing and integrity checks, as well as the gathering of metadata, documentation, reports and all the information that is needed for the reuse of the dataset outside its original context.

15. Another task of data curation is to maintain consistency between datasets – for example, if a species name changes or is split into several sub-species according to genetic distribution, then the original datasets would need to be altered to reflect the changes. Another example would be changes in political boundaries, enumeration areas or place names.

Data warehouse

16. A data warehouse is a system that combines data from several sources, possibly heterogeneous, for analysis, reporting and mining by users. It requires a data integration layer that extracts, transforms and loads data in a format suited for the cross analysis of the datasets. A data warehouse does not replace existing databases; it comes in addition to them for the purpose of data mining. It often requires specific tools, skills and an intrinsic understanding of all datasets and their limitations.

17. In an exploratory phase, data sources are often aggregated and combined manually or programmatically for statistical analysis, which provides more flexibility than a predefined multi-dimensional array (cube) in a data warehouse.

18. A data warehouse becomes useful when a lot of data is available from various sources, and needs to be analysed jointly and regularly for the same temporal and spatial units – for example, fisheries exports data could be cross-analysed with customs data (by year, destination country and commodity), or HIES information could be combined with catch data at the village level, etc.

ISSUES AND CONCERNS

19. Not all fisheries departments and NGOs benefit from IT service and system administrators that are in charge of handling back-ups of discs and databases. As there is often not enough internet bandwidth available to transfer large amount of files or data on a daily basis, the back-up strategy needs to be tuned and strategised with a good balance between the risk of loss, value of the data (can it be re-acquired and how much would it cost?) and cost of storage and digital preservation.

20. Identifying and recovering legacy datasets will require involvement of people that have been involved in the surveys, data entry and analysis – not only to locate the data and reports, but also to appraise the dataset and produce metadata if the dataset is selected for long-term storage. Sometimes legacy data is perceived as being lost but still exists on a failed server hard drive, or an electronic or paper copy still exists somewhere.

POSSIBLE DISCUSSION POINTS

- Are the databases and data files sufficiently backed-up and how is this done? What are the issues and challenges faced and workarounds in place in a particular country?
- Are there existing datasets and legacy databases that contain data that would be valuable for reuse and are there constraints to share it to a large audience?
- What should be the criteria and who are the people involved for the initial appraisal and ulterior reappraisal of the datasets, especially if they are stored by a third party (such as SPC) on your behalf?

ANNEX A: TYPES OF DATASETS HELD BY SPC COASTAL FISHERIES PROGRAMME

Underwater invertebrate and fish survey data (SPC methodology)

Surveys conducted by countries and SPC staff members under PROCFish/C, SciCOFish, Aquarium trade and Climate Change monitoring projects. Additional data for surveys conducted by countries with the same methodology is backed-up occasionally when SPC assistance is sought. The data consists of fish observations, densities and estimated biomasses on 50 m transects, invertebrate counts and densities (various methods), as well as substrate and habitat information.

Socio-economic data (PROCFish/C methodology)

Surveys conducted by countries and SPC staff members under PROCFish/C. Additional data for surveys conducted with the same methodology is backed-up when provided to SPC. The data consists of household and fisher interviews.

Coral photoquadrats, coconut crab, seagrass and mangrove survey data (SPC methodology)

Surveys conducted by countries and SPC staff members under SciCOFish, Climate change monitoring and DFAT projects. The data is stored online on the web database.

Export data

Export permit requests and shipment data, mostly for aquarium trade related commodities (aquarium fishes, clams, corals), but similar data for other commodities is also stored occasionally as snapshots of country-specific databases.

Creel and Market data

Creel and Market surveys conducted by countries and SPC staff members. Data consists of landing catch and fish for sale at the market, with additional information on boat equipment, purpose of catch, costs, prices, etc.

Other market data that is also backed up from country-specific database.

Artisanal data

Landing data for small-scale fisheries as entered in TUFART, TUFMAN 2 and Tails. The data consists of landing catch per fishing event and trip. It also contains information about FAD deployments for some countries.

Country-specific databases

Occasional snapshots of licensing databases, water quality and other country-specific databases.