SPC COASTAL AND OCEANIC FISHERIES DIGITAL LIBRARY

Franck Magron
Reef Fisheries
Information Manager
SPC, Noumea, New Caledonia
FranckM@spc.int

The Secretariat of the Pacific Community (SPC) has been involved in marine resource management since its creation, 60 years ago. Since then, thousands of papers and reports have been produced by, for, or in collaboration with SPC, and have been archived over the years by SPC's library.

These documents are invaluable for retracing past records of marine resource exploitation in the Pacific Islands region, providing information that is no longer available anywhere else; for example, commodity export figures for the 1920s and 1930s. While digging through 60-year-old reports can be a thrilling experience, it requires considerable time and physical access to documents. In practice, old reports and meeting papers are rarely accessed and are not easily searchable.

Recent documents have been made available electronically on SPC's website, but they are scattered in different places and not searchable through a common interface. Moreover, not all fisheries officers in the Pacific have access to the Internet; and those that do sometimes have difficulty downloading documents because of limited bandwidth.

The challenge of this project was to resurrect historical documents and make them, as well as more recent papers, available to the largest audience possible — whether they have access to the Internet or not — and provide search and retrieval tools for thousands of documents collated into a digital library.

A strong collaboration between SPC's library, the Fisheries Information Section and the Reef Fisheries Observatory (with funding from the European Union through the Pacific Regional Oceanic and Coastal Fisheries Project) made it possible to create both a DVD (with annual releases) and an online version of the fisheries library that can be accessed from the Internet

http://www.spc.int/mrd/fishlib.php

The first part of this article describes the general process of creating a digital library, while the second part reviews its practical uses.

Building a digital library

Building a digital library is a process that involves scanning and performing optical character recognition (OCR) on historical documents so that they are searchable in full text. This process also attaches additional information (metadata) such as title, author or year of publication to each electronic document to improve searches and provide relevant results. Archived documents must be indexed using full text and metadata so that they can be searched and retrieved.

Figure 1 depicts this process, and shows that entering metadata and proofreading documents are tasks that require considerable staff time, whereas optical character recognition is a computer intensive operation that can be automated and requires little human intervention.

Optical character recognition of scanned documents

Optical character recognition (OCR) is the process of analysing pages of a scanned document and recognising the text, possibly introducing typos and errors that cannot be corrected automatically. We chose not to proofread the recognised text and to save scanned documents as "PDF/Image over text", a file format in which the original page is displayed as an image, while the underlying text is indexed and searchable.

Performances and the possibility of running automated tasks were evaluated for several OCR applications. We finally selected FineReader™, which performs well on most documents.

Performance, however, is poor for older documents where ink

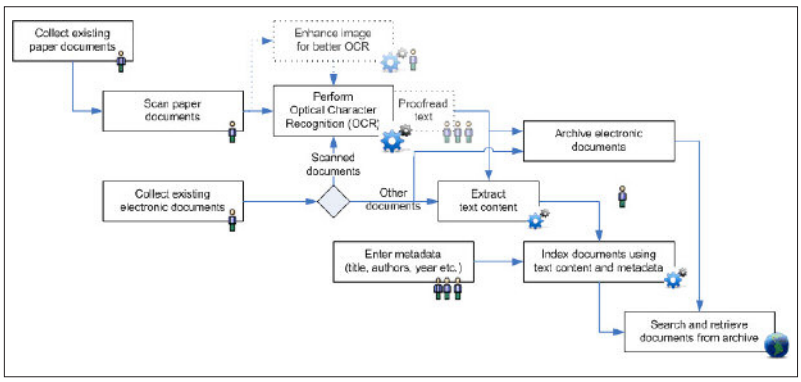


Figure 1: Task flow for building the digital library.

may have faded, bled or diffused through the paper. A typical black and white scan of a text page, for example, uses a threshold to separate text from background, which produces highly contrasted text but wipes out faded characters. The same page scanned as a "greyscale" image can, on the other hand, be enhanced using image processing as shown in Figure 2.

Because grayscale image enhancement takes much more human and computer time than the standard process, it is only used when a document cannot be otherwise scanned or text recognised.

Finally, the quality of text recognition by OCR software is better if the language is correctly set before the recognition is performed. The reason is that the software uses intrinsic properties of each language, such as the probability that a given character follows another, and word lists.

For example, a document in French recognised with English settings will be poorly recognised, which is an issue for SPC's conference papers that are available in both French and English, and for other bilingual documents. Figure 3 shows the impact of language settings on recognition.

Note that it is possible to detect the language of a document a posteriori using recognised text statistical properties, and checking to ensure the language settings were correct, or redoing the recognition with corrected settings.

Bilingual documents are more difficult to process automatically. One solution could be to perform the recognition in both languages and then merge the two documents line by line after language detection; however, this has not yet been routinely implemented in our system.

Metadata, text processing and classification

Attaching metadata to each document can be time consuming, but it is essential both for displaying search results and for determining the relevance of "hits". For example, the presence of search terms in titles or other metadata fields of a document means it is probably more relevant than other documents where the search terms just occur within the text.

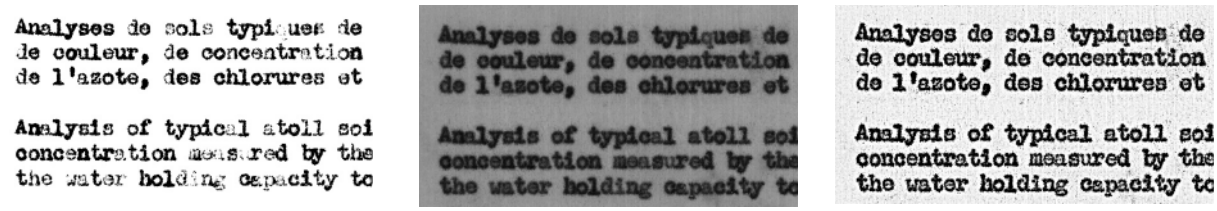
Restricting searches to a certain category of documents is also made possible by metadata. For example, displaying articles from a specific bulletin published during the last three years, or searching documents written by specific authors, is feasible if the information for year of publication, journal and authors has been previously recorded.

Metadata includes:

- Normalised bibliographic elements (MARC, Dublin Core) used to cite the document. These elements can generally be extracted directly from a library catalog such as Koha or bibliographic management programs (e.g. ProCite, Reference Manager, EndNote).

- Keywords and other information about the intrinsic content of the document. Keywords can either be provided by librarians and domain specialists, or they can be inferred automatically by text classification. The pertinence of keywords depends on the broader context of the digital library thematic. For example the keyword fisheriesbrings very little information in the context of the coastal and oceanic fisheries digital library, and could even interfere with a search. In the context of all SPC documents on the other hand, it can be used to restrict the search to fisheries-related documents.

- Information about the context of the document such as:



(1) Moreover, **our** informations **about** phyllosomas and pseudibacus of the genus Scyllarides are summed up, and an **attempt** of a key for **the** identification (…)
Dans un article **precedent**, paru dans les **manes** Cahiers (CROSNIER, 1972), nous avons **etudie** les larves phyllosomes et les post larves de Panulirus

(2) Moreover, **oui**- informations **aboul** phyllosomas and pseudibacus of the genus Scyllarides are summed up, and an **attempt** of a key for **ihe** identification (…)
Dans un article **précédent**, paru dans les **mêmes** Cahiers (CROSNIER, 1972), nous avons **étudié** les larves phyllosomes et les post larves de Panulirus

Figure 2 (top): Document scanned as black and white text, as a grayscale image, and the latter after image enhancement. Figure 3 (left): The same bilingual document recognised with English (1) and French (2) settings.

1. links to other articles and papers from the same bulletin issue or workshop
2. links to other publications from the same author(s) in the digital library
3. list of other publications where the document is cited
4. list of other publications cited by the document
5. list of closest documents in the digital library (measure based on text and metadata)

- Location and properties of electronic files. The same document can be distributed in several formats such as PDF and DjVu or even split into parts to ease Internet download, but all of these files share the same bibliographic and content-related information.

When doing a full text search, metadata is complementary to the text. Because a full text search will look for all occurrences of searched terms in the document and evaluate their importance based on their number of occurrences relative to the length of the document, there is little to gain by adding the very same terms to the keyword list. Yet the relevance of results can be improved by adding synonyms, broader concepts, and disambiguating terms when necessary.

For example, it is possible to search for documents relating to islands or countries that either go by different names or different spellings — Chuuk or Truck, PNG or Papua New Guinea, New Hebrides or Vanuatu — by injecting synonyms during the indexation of documents and the transformation of quoted expressions in queries.

Classification and text analysis can also be used to automatically create additional keywords and give more weight to some terms when evaluating the relevance of results. An additional metadata field with country name for example, makes it possible to restrict a search to documents related to a specific country and improves the relevance of search results that include a country name. This field can be generated automatically based on the geographic names in the title and full text, for example Kiribati will be inserted if the title contains Gilbert Islands or Tarawa.

Finally, disambiguation is the process of analysing the context of the document to determine the meaning of an expression before it is used for classification or replaced by synonyms. For example *T. gigas* can refer to *Tridacna gigas* but also to *Tricornis gigas* or *Tetraedron gigas*. The presence of the term *Tridacna* indicates that *T. gigas* will be expanded to include *Tridacna gigas*.

The creation of synonyms and additional keywords is not mandatory to create a digital library and it complicates the indexing process, however, it greatly improves the results and user experience.

Producing a DVD and online version of the digital library

When metadata and searchable electronic documents have been produced, they must be imported into a system that indexes text and selected metadata and provides an interface to search and retrieve the documents.

For the current DVD version (Digital Library 2006), we used Greenstone 2.71 (available from www.greenstone.org), an open-source software that we customised for use of sub-collections indexes, hierarchical browsing of documents, and to provide an SPC fisheries look and feel. Greenstone is freely redistributable, has a simple search interface, and can be launched from a DVD without any software installation.

For the online version, we developed our own interface using Lucene, a powerful open-source indexer that we extended for multi-word synonyms and preprocessing of unquoted queries to transform list of terms such as American Samoa reef fisheries into quoted terms ("American Samoa" "reef fisheries") and give more emphasis to documents that contain the quoted expressions. The online version currently provides access to more than 5000 documents, compared with 2600 documents for the 2006 Digital Library on DVD.

Because the technologies used for the online and DVD versions are not the same, the results for similar queries can be different, especially the order in which the results are displayed. The on-line search engine is designed to give the best top 10 results as possible for your query whereas the DVD version has more features to browse leisurely the documents.

Using the fisheries digital library

Online digital library

The online digital library can be accessed from:

http://www.spc.int/mrd/fishlib.php

A standard search is done on both full text and metadata, but it is also possible to search only specific metadata fields, and to restrict the search to specific collections (e.g. bulletins, manuals) and languages (e.g. English, French). The help page provides more details about advanced search features.

Figure 4 displays the result of a search on giant clams for the years between 1992 and 2003, where Vanuatu (or a synonym) appears in the title, and where the search is restricted to SPC's *Fisheries Newsletter* and *Beche-de-*

mer bulletins in English. The search result shows the document title and author, its bibliographic reference, as well as the full text fragments where the searched terms appear.

About 50 documents are added every week by the SPC library with an expected completion of the scan of past SPC coastal and oceanic fisheries documents around mid 2008.

Digital library on DVD

The DVD version is made available to Pacific Island fisheries administrations that received several copies of the 2006 version, whereas the online version is accessible by the general public from anywhere in the world.

To launch the digital library, insert the DVD and press Enter Library. This starts the library, launches your favourite Internet browser, and opens the library home page. Once you have entered the library, you can either search the documents by title, authors, countries and text content, or browse the documents by collection (bulletins, meetings, manuals, posters, etc.), title, authors and countries. Finally, you can open the PDF document by clicking on its thumbnail or PDF icon.

Help on queries is provided when clicking on the Help button on the top right of the document.

A new DVD will be released in 2008 with an updated version of the digital library containing around 5000 electronic documents.

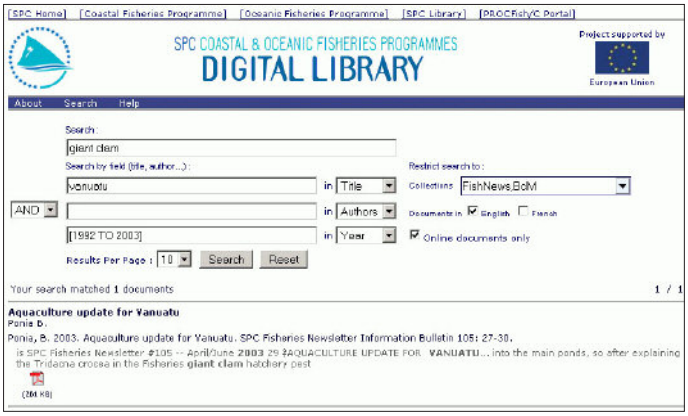


Figure 4: Web interface of the online digital library.