‘Pleins_Textes’: IRD (Institut de Recherche pour le Développement) Electronic Library

Pier Luigi Rossi and Marcel Ngoma-Mouaya
Institut de Recherche pour le développement (IRD), France

Abstract: *In the early 1960s the Institut de recherche pour le développement (IRD; Research Institute for Development) constituted a reference holding by collecting the scientific production of its research staff. This provides access to its ‘scientific memory’ and is a reference for scientific knowledge about developing countries. In 1999 we started the systematic digitisation of this holding, thereby creating the IRD electronic library. This article presents the study which enabled the implementation of the digitisation procedures. We focus on an optimised chain of production providing quality result images in 150 dpi for colour thematic maps. Also discussed is the chronology of the digitisation process, which was established by a document’s typological analysis according to the publishing practices, stock availability, condition of binding, document age, and paper version availability jointly with electronic medium. A study concerning the production costs and an analysis concerning the size of the produced files is presented. We analyse our search system that links full text with bibliographical fields. Field information is automatically included in PDF files and in the index generated by the Verity search engine ‘Information Server’. We underline the transition from a system representing objects such as articles, chapters and monographs, to a system where objects have fine granularity (pages, paragraphs, sentences answering requests) which is generated by access to full-text documents. Finally, there is a description of the products that can be generated from Pleins_Textes, especially thematic CD-ROMs. These CD-ROMs are useful particularly in the technical environment of libraries in developing countries. Pleins_Textes, the IRD electronic library, is online at www.bondy.ird.fr/pleins_textes, with 400,000 pages at present.*

Keywords:

1. Introduction

IRD is a French state-owned science and technology research agency under the joint authority of the French Research and Overseas Development ministries. This institute carries out research in Africa, the Indian Ocean, Latin America and the Pacific in the following disciplines: earth and environment, living resources, social science and health. The IRD has 36 facilities in all, 31 of which are primarily located in the tropical zone.

Starting at the beginning of the 1960s, IRD chose the Bondy site in France to set up a reference library pooling all the scientific production of its research staff. This collection may be considered the ‘scientific memory’ of the institute.

The Horizon database was started in 1986 with the objective of analysing and classifying every documentary element in the collection (off-prints, articles, chapters, books). This information access system is based on the online posting of biographical references (Horizon database) and by a network of Institute libraries featuring all, or part, of the IRD’s reference collection. Most documents are consulted on-site, with inter-library loans accounting for a small percentage of total consultations.

From the mid-1990s, changes in both information storage and information routing technologies made it possible to migrate an increasing number of documents to electronic media (internet, CD-ROM). This meant that the end user no longer needed to physically go the library in question: the internet and CD-ROMs had made virtual collection consulting possible. These technological developments naturally resulted in the decision to digitise the IRD’s collection and to create an elibrary.

2. Digitisation

Given the current information technology trend enabling the end user to consult documents over the internet or from a CD-ROM, the digitisation of IRD’s collection will make it easier for a greater number of people to find the information they are looking for by facilitating direct access to documents and textual content, and by making it easier (and quicker) to run document information searches.

Our initial monitoring and testing phases, completed in the course of 1996 and 1997, enabled us to define a general digitisation procedure based on the following criteria:

- Documents to be stored in PDF format
- Optical character recognition (OCR) in 'image mode plus hidden text' for pages scanned in black and white
- Digitisation of bitonal documents (black and white) at 300 dpi
- Digitisation in 'bitonal image' mode for colour pages
- Colour digitisation at 150 dpi for thematic maps and colour covers.

The PDF format was chosen for a number of reasons: multi-platform compatibility, optimisation for internet routing, widespread use among scientific communities worldwide, and lastly, the high level of OCR as regards 'image + hidden text' mode. We considered this mode essential for our mass digitisation project. The edocument is a true copy of the original and the recognised text is used for indexing and searching.

2.1. Colour inside text, colour outside text

One of the issues to be resolved when digitising a scientific collection was the scanning of documents containing colour illustrations.

We decided to define the digitisation context by taking into account the state of the art of the material being scanned and available technologies, the operating modes for digitisation, and lastly the limits imposed by the OCR.

Improvements in digitisation technologies have been mainly to the benefit of desktop scanners, which have become faster with high-resolution and quality colour handling modes. This can be contrasted with industrial scanners, which have remained fairly 'rustic' machines with emphasis on high-speed bitonal output. For a major scanning operation such as ours, this meant that special attention had to be given to pages with colour illustrations: manual extraction, separate scanning and re-insertion of pages and files back in their positions.

We noted that files increased rapidly in size as we moved from bitonal to greyscale to colour mode. We also noted that compression is done with no loss of information (TIFF G4 compression) in bitonal mode, but generally results in information drop-off when compressing colour files.

OCR suffers when scanning colour pages and becomes unacceptable if scan resolution has to be reduced to keep the file size down. However, differences in quality of this nature inevitably mean disparities in the scanned collection, both as regards data searching as well as data exporting. For these reasons, pages with colour illustrations are scanned in bitonal format.

A different solution was adopted for pages with colour elements providing very specific information or defining visual identifiers for the document. In this case, the bitonal solution given above would not have given sufficiently good results (e.g. thematic maps, covers) as the level of optical recognition would have been too low. Moreover, such elements requiring special treatment are often easy to locate, take out and re-insert again (at the start or the end of document, thickness etc.).

These elements were scanned at 300 dpi in colour and sent to us in TIFF format. The scanned results are integrated into the PDF file by means of semi-automatic Adobe PhotoShop and Adobe Distiller macros. To give an idea of size, an original A2 page yields an output file of approximately less than 1 MB with a final resolution of 150 dpi.

3. Collection digitisation phases and methodology

The IRD collection consists of about 51,350 documentary units, corresponding to about 2,044,346 pages. In order to facilitate the organisation and scheduling of the digitisation process, we attempted to sort the documents on the basis of the following criteria: publisher type, availability of document(s) within the collection, condition of the binding and age of the document.

- Publisher type refers to documents published by the Editions de l'IRD publishing house, as well as documents from other public and private publishers and lastly documents considered grey literature.
- Availability of documents is defined as the physical availability of the documents within the stocks (published IRD books stock, collection, libraries). The higher the availability factor, the greater the number of documents 'available' for destruction or requiring only a slight degree of maintenance after scanning. A number of possibilities exist: a document published by the Editions de l'IRD might be freely available in the published books stock; it might be available in some IRD site (grey literature); it might exist in several copies in the collection; or the document might be on microfilm.
- The condition and type of the binding can play a role in deciding how the documents are treated. About 39 per cent of the IRD's documents consist of separate off-prints stored in clearly indicated archive boxes. This leads us to differentiate documents into those with hard binding and those with the light binding (generally stapled).
- The age of the document is defined as the period when a document exists only in paper version and the

period when an electronic version also exists. Comparing the electronic version of a document with its paper version to check for total fidelity is a difficult process. The procedures whereby a paper version would be included in the collection while, at the same time making the electronic version available online (eversion created from electronic sources) remain to be defined. Thus, we decided to construct our elibrary uniquely from original paper and/or microfilm sources.

The creation of these various categories formed the basis of the digitisation decision-making process and the methodology of the Institute. The digitisation process consists of the following steps:

1. Digitisation of all the documents available in the IRD published books stock with subsequent destruction of the source document. This step has the advantage of rapidly creating a voluminous 'sample' of scanned documents representative of the Institute's holding while, at the same time, keeping the costs of the operation very reasonable.
2. Digitisation of those journals and other books in the IRD holding in order to make up for incomplete series elsewhere (copies out of print in the published books stock).
3. Digitisation of documents with light binding (off-print).
4. Digitisation of grey literature of which there are additional copies on the Bondy site (documentary holding, published books stock, digitisation units) and in the stock of the Bondy library (extraction and sifting of holdings).
5. Global analysis of the digitisation operation to determine which documents remain to be scanned; this step is carried out after automatic indication of the digitised files in the bibliographic database.
6. Digitisation of copies two or three in the Institute's collection (hard binding) with low maintenance after scanning.
7. Second global analysis to determine which documents remain to be scanned.
8. Digitisation of the remaining documents from microfilm or from copy one followed by document restoration.

4. Means and costs

When we began the testing and monitoring phase, we set up a small-scale digitisation room at Bondy. This was an opportunity for us to familiarise ourselves with the procedures, phases and technology involved in digitisation work. Based on this experience, we drew up project specifications, which we submitted to contractors with a view to determining the costs of a major digitisation operation. We subsequently examined the resulting tender proposals and selected one subcontractor on the combined basis of expertise and price.

Table 1 summarises the main unit costs of the digitisation process. We estimate the average cost per page of digitisation at 1.20 French francs (1 French franc = 0.15245 euro). This figure takes into account related expenses (preparation, supplies, transport), the processing of thematic maps as well as the margin of error in quantifying the exact size of the volume of documents to be scanned. The overall estimated cost for digitisation of the entire collection is about 2,450,000 F (373,500 euro). We must also factor in the salary of one person who will be assigned to the project for two years with the task of preparing the documents to be digitised, and then bringing the electronic files into order: validation, division, naming, transferring of elibrary files, online posting and so on.

As regards size, we estimate the average size of a scanned page at about 80 KB, which, considering the size of the collection, yields a total elibrary of 210 GB.

Table 1:

Principal per unit costs of digitisation

Preparation of documents (analysis of the contents: photos, colour pages, annexes, removal of bindings, document restoration)	Hour	200.00 FF
Global service incorporating 300 dpi recto/verso digitisation, indexing of each work, generation of multipage files. OCR processing with Acrobat Capture (image + hidden text), systematic monitoring and document rotation, if required	Image	0.86 FF
Digitisation of large-format maps via 35 mm microfilm (drawings which are an integral part of the work. Files and documents returned to their place)	Drawing	30.00 FF
Digitisation of large-format maps via colour digitisation (plans which are an integral part of the work. Files and documents returned to their place)	Drawing	70.00 FF
Supply of CD-WORM	Unit	120.00 FF
Document transport	Run	120.00 FF

5. Information search modes and data storage formats

In order to define how the source contents of the elibrary might be accessed and searched, we studied the various types of relationships possible between the bibliographic description of a given document and the document itself. Jacqueson and Rivier (note 4) outline three ways of linking up the document description with the document itself (see Figure 1).

We studied two software packages, Texto-Web (Cincom; www.cincom.com/cindoc) and Information Server (Verity; www.verity.com) and, on this basis, established a model based on three objects (description, link, document) as well as a model based on an object (document containing its own description).

The three-object model can be used to search for information in the bibliographical description (field structure) and to access documents over a link. The single-object model can be used to search for information in the bibliographical descriptions (field structure) as well as in the edocument itself. We found out that the first solution lends itself to online consulting but cannot be adapted to the creation of CD-ROMs without the creation of specific software (which implies a certain financial investment). The second solution can be used for both online consulting (Information Server) as well as for the creation of CD-ROMs at no cost (Acrobat Reader + Search).

We used a system of unique identifiers to create the relations between bibliographical descriptions of a given document and the document itself. The identifier in question is the inventory number of the document, which is present in the bibliographical description and which is actually used as the name of the electronic file. We developed programs that can be used to manage the link to the edocument (in the case of the three-object model) and programs that allow the data to be transferred to electronic files (in the case of the one-object model). The program used to send the data into the PDF files is available at no cost.

As regards our elibrary, we opted for the single-object model which can search in both the documents themselves as well as in the structured descriptive fields. These fields are retrieved automatically from the library database and are then sent into the PDF files (authors, title, subject, keywords) and into the Information Server indexing system.

We find that the solution which consists of searching in the content of a document (combined, if need be, with the structure of biographical fields) is more in line with the idea of browsing the elibrary while, at the same time, locating the most relevant documents. For each electronic document in PDF format, the search leads the user into the page containing the most relevant elements (word, phrase). The required search element is highlighted and only the page(s) with the related elements is downloaded by the user. Thanks to this searching and display solution, the documentary unit in its entirety (article, chapter, books, often with a large number of pages) is no longer of relevance: instead, the exact pages, paragraphs and phrases corresponding to the user's query are emphasised. Focusing on the exact information searched for by the user enables a new document granularity to be defined, which is good enough to be used directly online by the user.

In order to make sure that the identification of the documentary units remains at a constantly high level while the scanning operation is in progress, we are limiting the online access to the bibliographical description until such a time as the document itself has been scanned. The two collections with a heterogeneous structure and file format (HTML, PDF) are accessible at the same time. We use a common denominator taken from the bibliographical database to carry out searches in the fields (title, authors, keywords). As regards text searches, the common denominator is the text resulting from the optical character recognition (PDF files) or, alternatively, the text taken from the complete biographical notice (HTML files).

These 'classic' search routines are complemented by the notion of thematic categories, by association of documents by groups of shared words (clusters), by geographical thesauri.

See Figure 2 for more details on the Pleins_Textes web search page.

Figure 1:

Relationship models between document description and document content

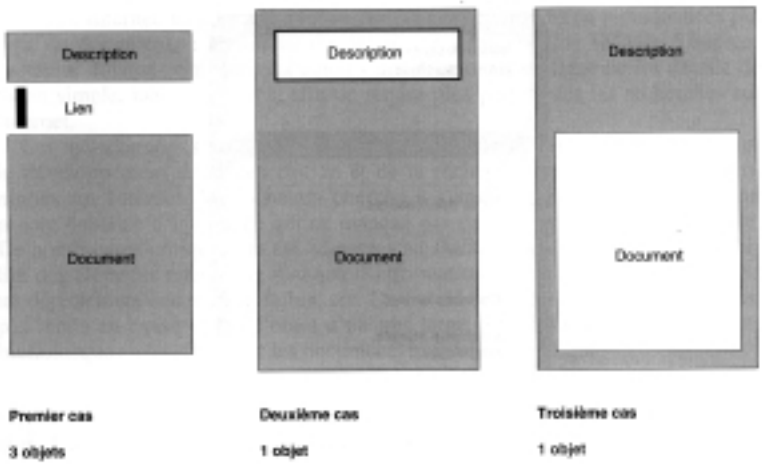
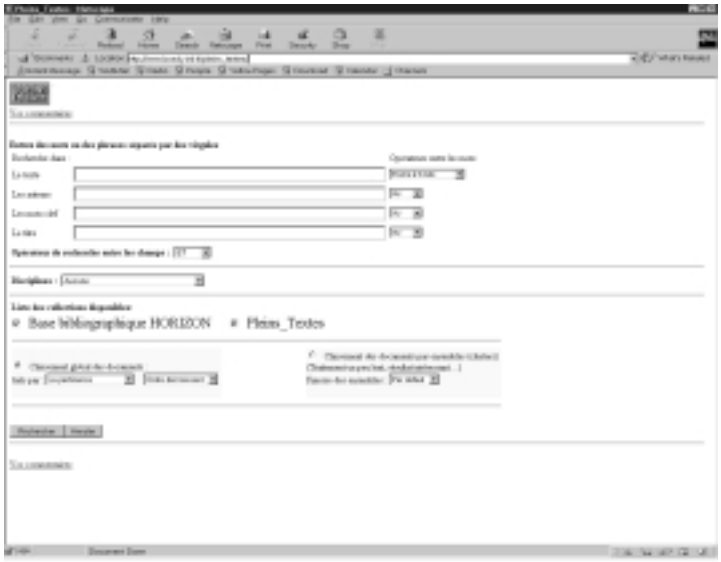


Figure 2:

Pleins_Textes web search page



The unique identifier in the bibliographical descriptions (used for the name of the PDF file) can be used to make precise searches of digital documents based on search results generated in the bibliographical database. Linking up document descriptions and document files in this matter can also be used to generate different types of web presentation or classification pages for which the links to the PDF files are generated automatically. We make full use of this possibility during the recording of CD-ROMs for Institute partners containing the scientific productions of Institute research staff and related to a given region or country. We consider this feature to be very interesting as regards the reconstitution of thematic documentary collections on developing countries.

6. Conclusions

The project for the digitisation of the documentary holdings constituting the scientific memory of the IRD over more than 50 years has resulted in the creation of Pleins_Textes, an online reference library concerning scientific research in the intertropical zone countries.

The work involved in the creation of this project has enabled us to assess and appreciate the potential of the PDF format. Within the context of the en-masse digitisation, we placed great importance on the 'image + hidden text' format, thanks to which the scanned document is a faithful copy of the original and the recognised text is used for indexing and searching as well as by the final user.

An analysis of the different document categories enabled us to define the digitisation phases. The resulting strategy led us to begin our digitisation with IRD-published documents. These documents may be considered a representative 'sample' of the Institute's holdings and are easily digitised at low cost since the document used for digitisation exists in several copies and is destroyed after the scanning.

The development of the digitisation procedure for colour covers and thematic maps means that the high levels of information and relevance required to characterise the contents of these scientific documents are ensured.

Our study of digitisation procedures and technologies enabled us to secure interesting costs with external contractors. Moreover, we now have the experience and expertise necessary to develop digitisation projects on behalf of Institute partners, in particular in developing countries.

The search mode solutions selected have considerably enhanced our experience of granularity-related issues of relevance to digital documents. Furthermore, they enable us to conserve the access modes specific to the bibliographical databases (field-structured information). This integration means that we are also in a position to record thematic libraries on CD-ROM for the benefit of our partners.

Pleins_Textes is online at www.bondy.ird.fr/pleins_textes and presently features 400,000 pages.

Marcel Ngoma-Mouaya
Institut de Recherche pour le développement (IRD)
32 Avenue Henri Varagnat
93143 Bondy cedex
France
Email: rossi@ird.fr, ngoma@bondy.ird.fr

References

[1] Borgman, C.L. (2000) *From Gutenberg to the Global Information Infrastructure. Access to Information in the Networked World*, London, MIT Press.

[2] De la Vega, J.F. (2000) *La communication scientifique à l'épreuve de l'internet. L'émergence d'un nouveau modèle*, Villeurbanne, Presses de l'ENSIB.

[3] Dupoirier, G. (ed.) (1999) Les bibliothèques numériques, *Document numérique*, **2**(3/4), Paris, Hermes Science Publications.

[4] Jacqueson, A. and Rivier, A. (1999) *Bibliothèques et documents numériques, Concepts, composantes, techniques et enjeux*, Paris, Editions du cercle de la librairie.

[5] Kraus, H. (1999) *Scan et retouche d'images*, Paris, Campus Press.

[6] Le Moal, J.C. and Hidoine, B. (eds) (2000) *Bibliothèques numériques*, Paris, ADBS.

[7] Lesk, M. (1997) *Practical Digital Libraries : Books, Bytes and Bucks*, San Francisco, CA, Morgan Kaufmann.