

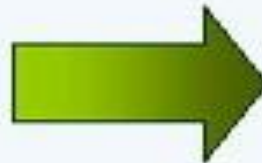
# STANDARDS REQUIRED FOR THE DIGITIZATION OF DOCUMENTS OR OTHER OBJECTS FOR FILE UPLOAD TO DIGITAL REPOSITORIES

Daryl L. Superio  
Southeast Asian Fisheries Development Center  
Aquaculture Department  
Tigbauan, Iloilo, Philippines  
[dlsuperio@seafdec.org.ph](mailto:dlsuperio@seafdec.org.ph)

# The Digitization Process



EXISTING  
HARDCOPIES

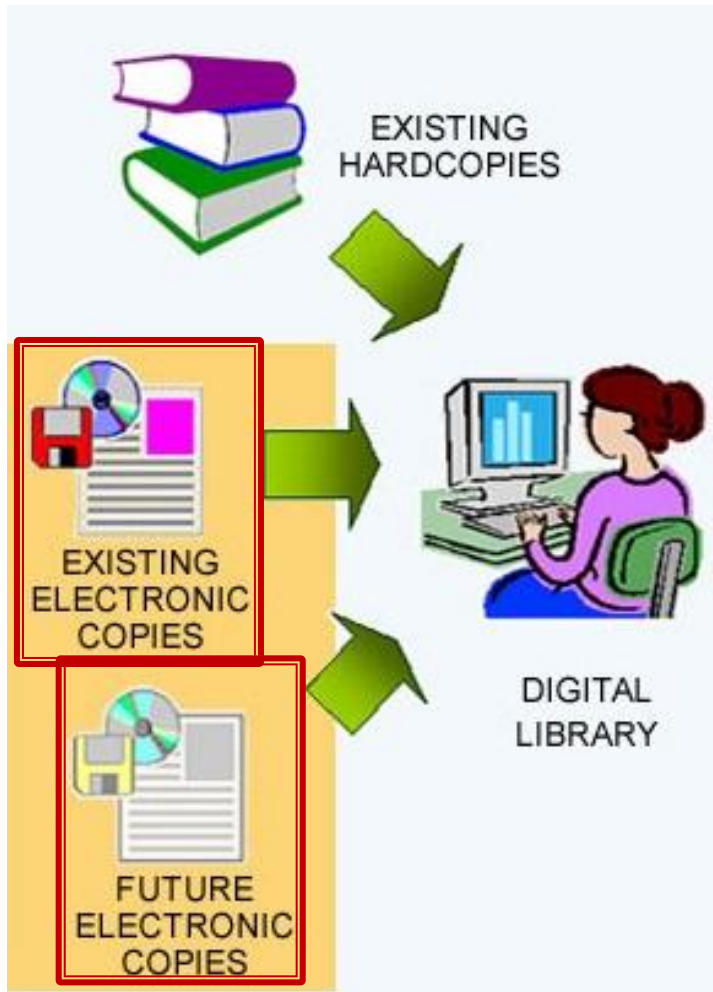


DIGITAL LIBRARIES

# The Digitization Process

- The digitization process will go through the following phases:
  - **scanning** documents and **converting** them to a format a word processor can read;
  - **proofreading** and **reformatting** them so they conform to your requirements; and
  - **adding metadata** (information used to catalogue the documents)
  - At the end of this process the hardcopy documents will have been transformed into electronic documents that can be included in a digital library

# Handling Electronic Documents



- **existing electronic copies** (electronic documents that have already been prepared); and
- **future electronic copies** (documents that are still in preparation, or that will be produced in the future).

# How Easy it is to Digitize Documents?

- The physical characteristics of a document determines how easy or how difficult it is to be digitized

EASY	DIFFICULT
White, clean opaque paper	Coloured, damaged or thin paper
Simple layout, single columns	Complex layout, multiple columns
Single sheets	Fragile, heavy bindings
Straight text with headings, few pictures	Many pictures, equations and tables
Standard computer typefaces	Unusual typefaces, poor quality printing, typewriting, handwriting
Unaccented Roman scripts	Accented and non-Roman scripts

# Basic Facilities and Requirements for Digitization

- **Equipment:**

scanner



storage devices



computer



# Basic Facilities and Requirements for Digitization

## ■ Software:

Software type	Purpose	Examples
Scanning and OCR	To convert the hardcopy image to a digital one, and then into text that a word processor can understand. A 'lite' version of scanning and OCR software is normally provided when you buy a scanner.	Ocropus (open source/multilingual) Readiris Abby Finereader Omnipage
Word processor and spellchecker	To correct text errors and to optimize page layout.	AbiWord (open source/multilingual) Openoffice (open source/multilingual) Microsoft Office Word
File conversion	To convert files from one format to another.	Zamzar (free) YouConvertIt (free)
Image management	To view, modify and manage images.	Picasa (open source/multilingual) Adobe Photoshop Iview
Image editing	To edit images.	Adobe PhotoShop Gimp (free)
PDF creation	Needed if you choose to create PDF documents.	PrimoPdf (free) Adobe Acrobat
PDF viewing	Needed if you choose to read PDF documents.	Adobe Reader (free) PDF-XChange Viewer (free)

# Basic Facilities and Requirements for Digitization

## ■ Language:

- When dealing publications with languages that use Roman scripts with a lot of accented characters (such as á, å etc.) and non-Roman scripts (Arabic, Chinese, Cyrillic, etc.). It is advised to do the following to solve the problem:
  - seek OCR software that is specific for your language,
  - set up a language-specific dictionary in your spellchecking or word processing program



# Basic Facilities and Requirements for Digitization

## ■ Personnel:

The following types of staff are needed for the digitization process:



A **manager** to coordinate the team and the digitization workflow.



Staff to do **scanning, OCR, proofreading and layout.**



Staff assigning **metadata.**



**Logistical and secretarial staff.**



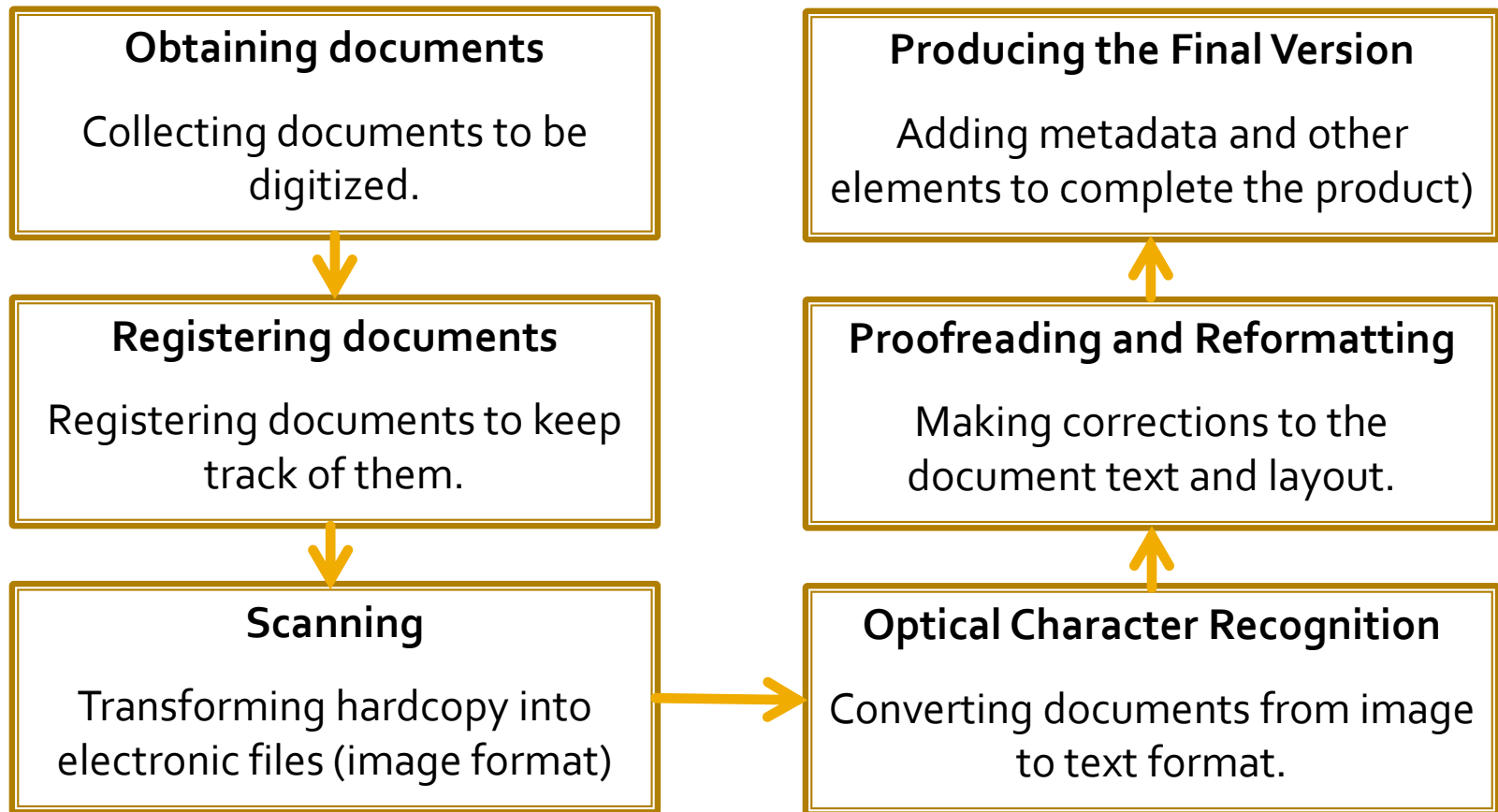
# Basic Facilities and Requirements for Digitization

## ■ Cost:

Equipment and software	Scanner, computers, software, office furniture.
Document acquisition	Registration, categorization, mailing and transport costs, staff time.
Scanning	Staff time, photocopying (if you photocopy documents before scanning them).
OCR, proofreading and layout	Staff time, consumables (disks, paper).
Metadata assignment	Staff time (depends on the number of documents, the difficulty of the subject, and the salaries of the specialists).
Management and overhead	Management, overhead, staff training.
Contingency	Additional, unanticipated expenses.

# From Hardcopy to Electronic Documents: Workflow

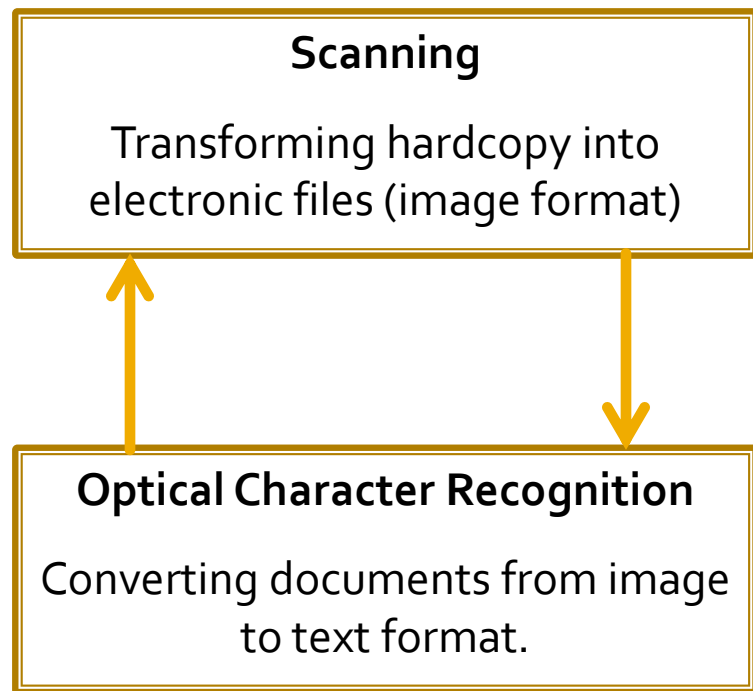
## ■ The Process



# From Hardcopy to Electronic Documents: Workflow

## ■ The Process:

- Before starting, consider the following options:



**It is possible to scan and OCR in a single operation, but it may be better to do these tasks separately: scan using the software that came with your scanner, then OCR the resulting files in a dedicated OCR program.**

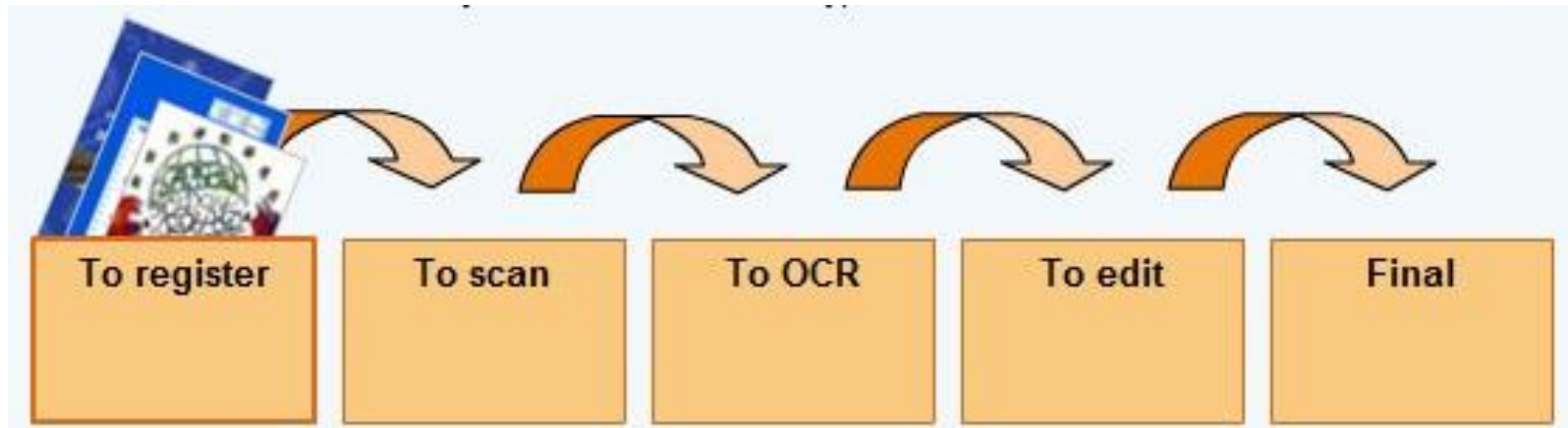
# From Hardcopy to Electronic Documents: Workflow

## ■ Managing Documents:

- Cataloging and establishing a filing system is advised when scanning a large number of documents
- Keep the hardcopies of documents at each stage of the process separate from those at earlier and later stages. As each document is processed, take it out of one folder, process it, and put it in the next folder.

# From Hardcopy to Electronic Documents: Workflow

- Managing Documents:



# From Hardcopy to Electronic Documents: Workflow

---

- **Managing Documents:**
  - keep track of the electronic versions of the documents you have scanned by keeping separate versions of each file in different subdirectories

# From Hardcopy to Electronic Documents: Workflow

- **Managing Documents:**
  - keep track of the electronic versions of the documents you have scanned by keeping separate versions of each file in different subdirectories
  - keep versions of the documents that you have scanned until you are finished, just in case the file become corrupt then you can go back to the previous version
  - make sure to make back-ups of each document



# From Hardcopy to Electronic Documents: Workflow

- **Scanning the Documents.** Before scanning make sure that:
  - the documents are clean, in complete pages, in proper order and in good condition
  - if the document is in poor condition try to find a new copy
  - if a sheet-fed scanner will be used, it is advised that a copy of a document/book be cut open to be able to feed individual sheets through the scanner
  - or, photocopy each page and feed the photocopy through the scanner

# From Hardcopy to Electronic Documents: Workflow

- **Scanning the Documents.** Remember that:
  - the better the quality, the more disk space the image takes up (*and the slower it downloads over the internet*)
  - for a textual document choose a setting with low resolution (72 dpi) and black and white
  - document that contains both text and graphics, may be scanned twice: once to scan the text in black and white, and again to scan the pictures in colour
  - save the text and graphic as separate file, then incorporate them later

# From Hardcopy to Electronic Documents: Workflow

- **Scanning the Documents.** Remember that:
  - the better the quality, the more disk space the image takes up (*and the slower it downloads over the internet*)
  - for a textual document choose a setting with low resolution (72 dpi) and black and white
  - document that contains both text and graphics, may be scanned twice: once to scan the text in black and white, and again to scan the pictures in colour
  - save the text and graphic as separate file, then incorporate them later
  - the software may produce a separate image file for each page of the document, in TIF format or in its own proprietary format that can be converted to PDF later

# From Hardcopy to Electronic Documents: Workflow

- **Optical Character Recognition**
  - OCR software converts a scanned image into a text file that a word processor can read
  - the software then breaks the text blocks down into lines and individual characters. It tries to match the image of each letter against patterns it recognizes as an 'a', 'b', etc.

# From Hardcopy to Electronic Documents: Workflow

- **Optical Character Recognition. How about those special characters** (Latin scripts with a lot of accented characters (such as á, å, æ, â, etc.), and non-Latin scripts (Amharic, Arabic, Burmese, Chinese, Cyrillic, Hindi, Japanese, Khmer, Korean, Thai...))?
  - use a language-specific dictionary in spellchecking or word processing program
  - use Unicode to represent characters
  - if the OCR software fails to recognize a large number of characters, it may be better to retype all or parts of the document, or to scan the text as an image

# From Hardcopy to Electronic Documents: Workflow

- **Optical Character Recognition**
  - **save your file, choose the format :**
    - **DOC, RTF or SXW if you want to produce PDF documents – or,**
    - **HTML - if you want to produce HTML documents**
  - **name the file following your file-naming convention**

# From Hardcopy to Electronic Documents: Workflow

- **Proofreading**
  - comparing the scanned text on screen with the hardcopy, and entering the corrections directly into the computer
  - printing out the scanned text and comparing it with the original copy
  - proofread tables and graphics carefully

# From Hardcopy to Electronic Documents: Workflow

## ■ Layout

- OCR software may produce a document that consists of straight text: no columns, no pictures, no headers and footers
  - reinsert these by hand, or correct where they appear on the page
- typeface, heading styles, and so on, can be changed to make the document more attractive and readable
- to avoid proofreading errors it is better to correct layout errors after the proofreading



# From Hardcopy to Electronic Documents: Workflow

- Layout
  - for HTML documents, use a simple layout: a single column of text, and so on
  - for documents destined to become PDFs, use a word processor to create a suitable layout
  - to create both HTML and PDF versions of the document, do all the proofreading and layout in a word processor, then convert the finished result into PDF and HTML formats
  - **Do not try to recreate the original layout exactly:** it can be very difficult and time-consuming

# From Hardcopy to Electronic Documents: Workflow

- **Producing the Final Version**
  - for many documents, some additional information can be added to the text so that readers can identify it easily
    - for a book, book title, author or editor, publisher and publication date must be included
    - for chapters in a book, title and author of that chapter and the original page numbers in the printed version of the book must be included
    - for journal articles, journal title, date, volume and issue number, the article title and authors, and the page numbers in the original printed journal must be included
    - this information may be added on the first page or in a footnote

# From Hardcopy to Electronic Documents: Workflow

- **Producing the Final Version**
  - in HTML and PDF files, "bookmarks" and hyperlinks to a document can be added.
  - a "live" table of contents for that document, can also be made so the user can click on a chapter title in the Table of Contents, and jump directly to that chapter in the text
  - When the final layout is finished the documents can be put in the "Final" folder: they are now ready to be included in a digital library.

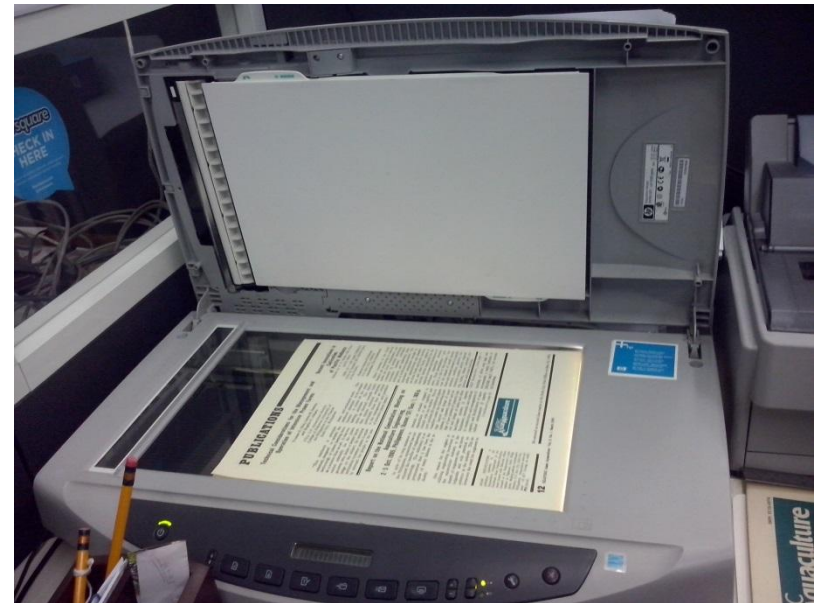
# Reference:

- *Digital libraries, repositories and documents: Information Management Resource Kit* (Eng. version 1.0). (2010). Rome, Italy: FAO.\*

\*Note: Majority, if not all of the contents of this presentation was taken from the above mentioned resource.

# Digitization: SEAFDEC/AQD Institutional Repository's Best Practices

- **Hardware: Sheetfeed Scanner**
  - Use the same basic technology as flatbeds, but maximize throughput, usually at the expense of quality
  - Generally designed for high-volume business environments, they typically scan in black and white or gray scale at relatively low resolutions
  - Documents are expected to be of uniform size and sturdy enough to endure fairly rough handling



# Digitization: SEAFDEC/AQD Institutional Repository's Best Practices

- Scanning Software: Irfan View



- a very fast, compact and innovative FREEWARE image format viewer/converter for Windows9x/ME/NT/2000/XP/2003/Vista/Windows 7/Windows 8

# Digitization: SEAFDEC/AQD Institutional Repository's Best Practices

## ■ OCR Software: Abby FineReader



- an optical character recognition (OCR) software that provides unmatched text recognition accuracy and conversion capabilities, virtually eliminating retyping and reformatting of documents
- converts scanned paper documents, digital images of texts and image-only PDFs into actionable formats such as Microsoft® Word, Excel® or searchable PDFs
- support up to 190 languages are supported for text recognition

# Digitization: SEAFDEC/AQD Institutional Repository's Best Practices

## ■ The Digitization Process

- scanning original institutional publications using sheetfed/ordinary flatbed scanners
- Original copy was “sacrificed”, unbound and scanned
- Master file in TIFF format
  - Resolution of 300dpi or higher
  - Stored in hard disk and a back up drive
  - Kept separately from the derivative file



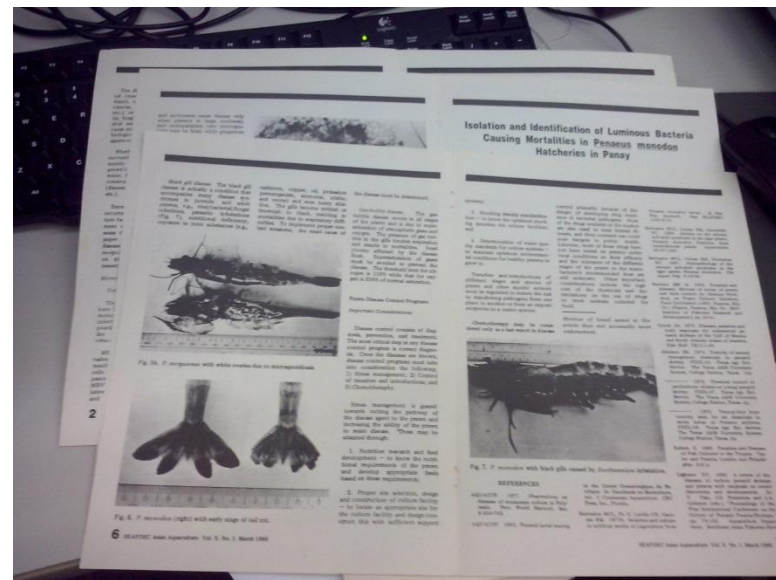
# Digitization: SEAFDEC/AQD Institutional Repository's Best Practices

## ■ The Digitization Process

1

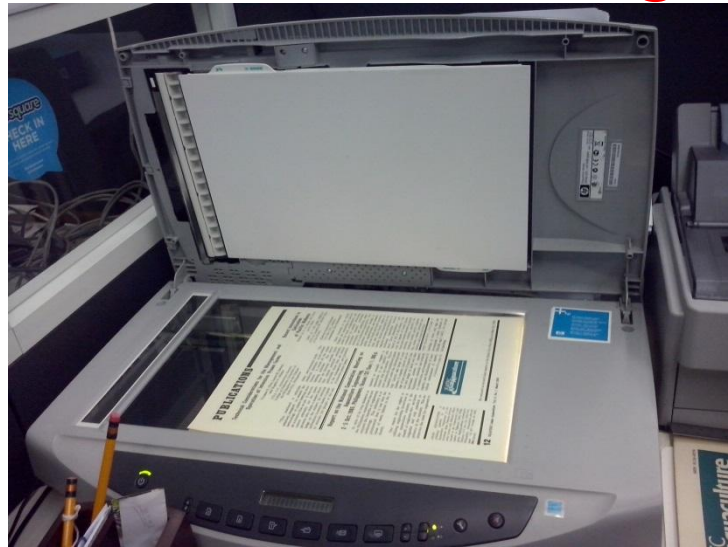


2



# Digitization: SEAFDEC/AQD Institutional Repository's Best Practices

- The Digitization Process



# Digitization: SEAFDEC/AQD Institutional Repository's Best Practices

## ■ The Digitization Process

- Thumbnail copy (JPG)
  - Derived for all material formats from TIFF master
  - Resolution: 72-100 DPI
- View / Service copy (JPEG/PDF)
  - Derived for all material formats from TIFF master
  - Resolution: 72-100 DPI
- Print Copy (PDF)
  - Derived for all material formats from TIFF master
  - Compressed
  - Resolution: 100-150 DPI

# Digitization: SEAFDEC/AQD Institutional Repository's Best Practices

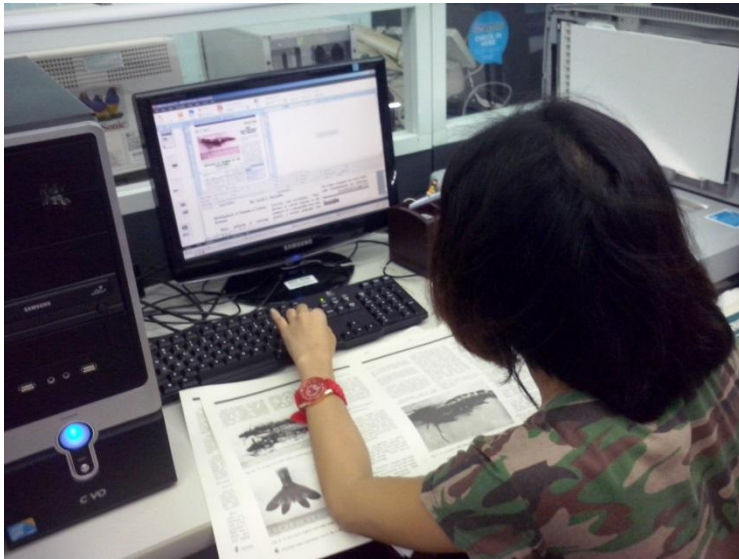
## ■ Optical Character Recognition (OCR)

- OCR using ABBYY Finereader version 11
  - did not settle for automatic conversion of scanned document to PDF nor on automatic OCR included or bundled with the scanners or with Adobe Reader/Acrobat.
- Each document was proofread and re-encoded
  - for publications with simple lay-out, the texts were re-encoded
  - for more complicated ones, text under image final output was chosen
  - some basic HMTL programming was required for tables pictures were scanned as is

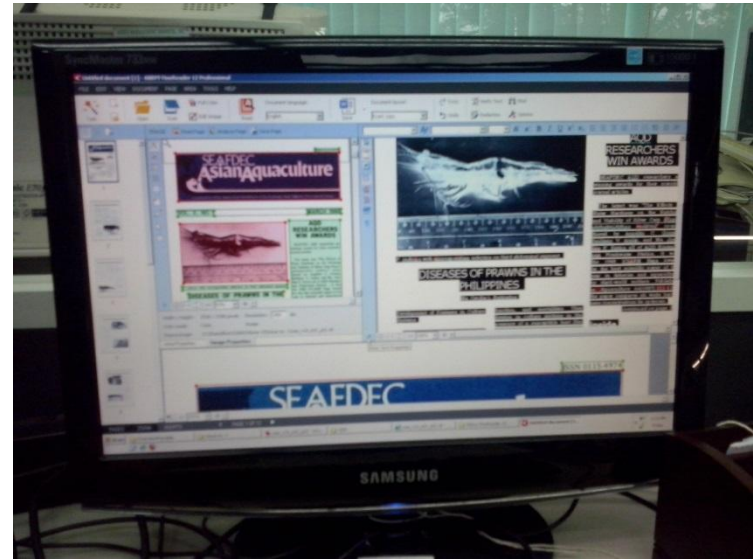
# Digitization: SEAFDEC/AQD Institutional Repository's Best Practices

- The Digitization Process

5



6

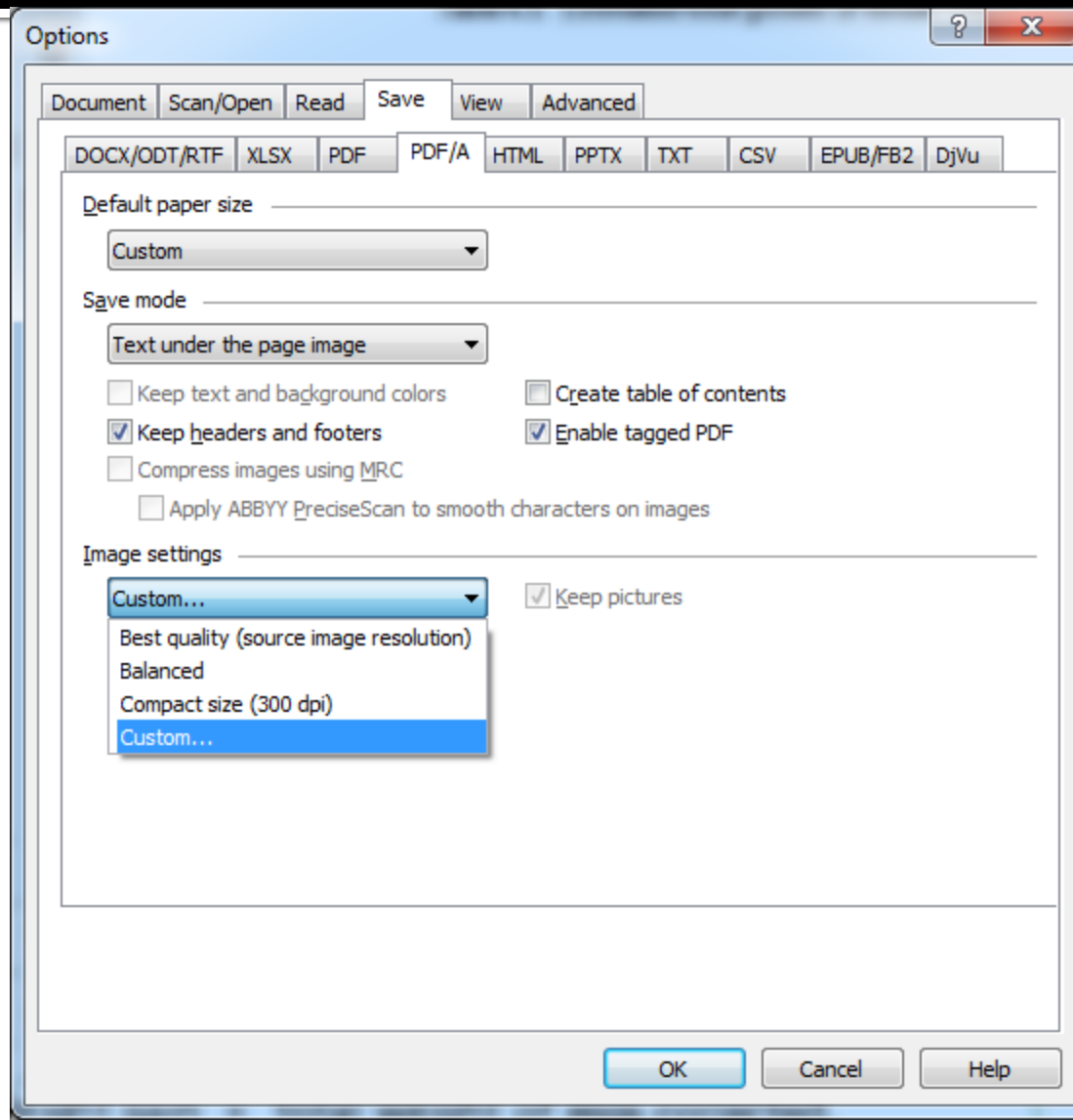


# Digitization: SEAFDEC/AQD Institutional Repository's Best Practices

## ■ Delivery Format

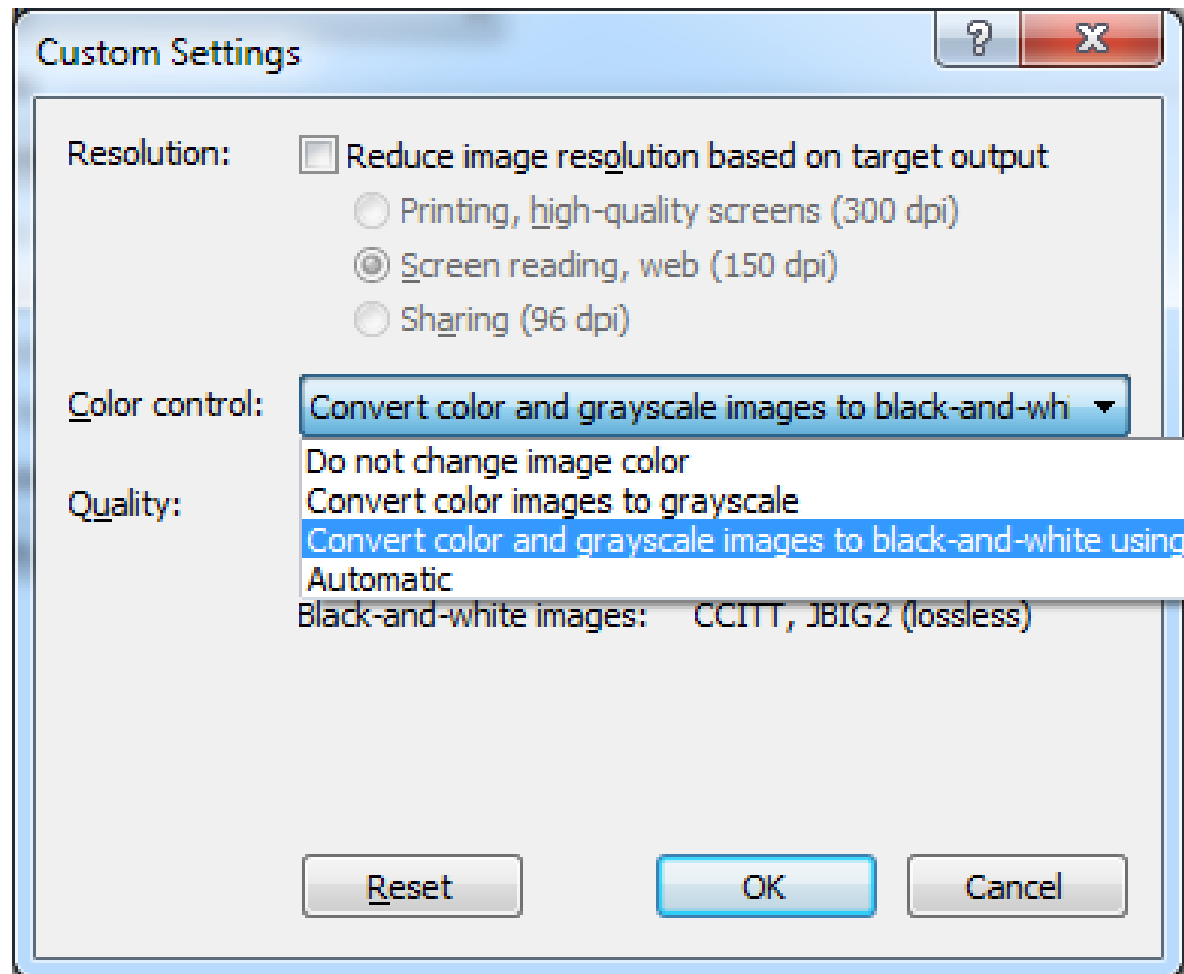
- Born-digital publications and scanned publications were converted to PDF/A as a standard delivery format
- PDF/A is the archival version of the PDF file format, an open source tool managed by International Standards Organization (ISO).
  - It enables long-term digital preservation of electronic documents that are already self-contained.

# Digitization: SEAFDEC/AQD Institutional Repository's Best Practices



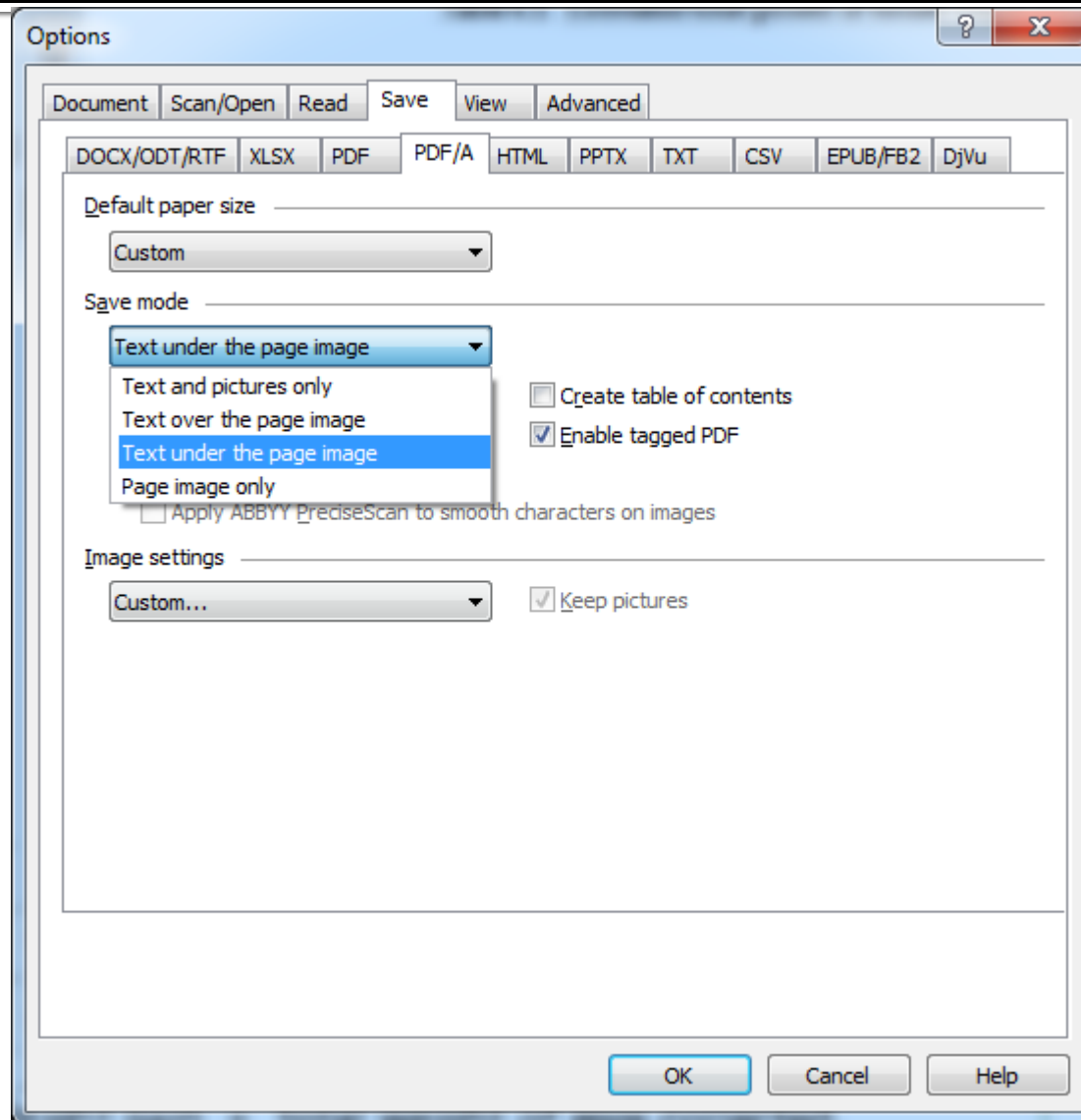


# Digitization: SEAFDEC/AQD Institutional Repository's Best Practices

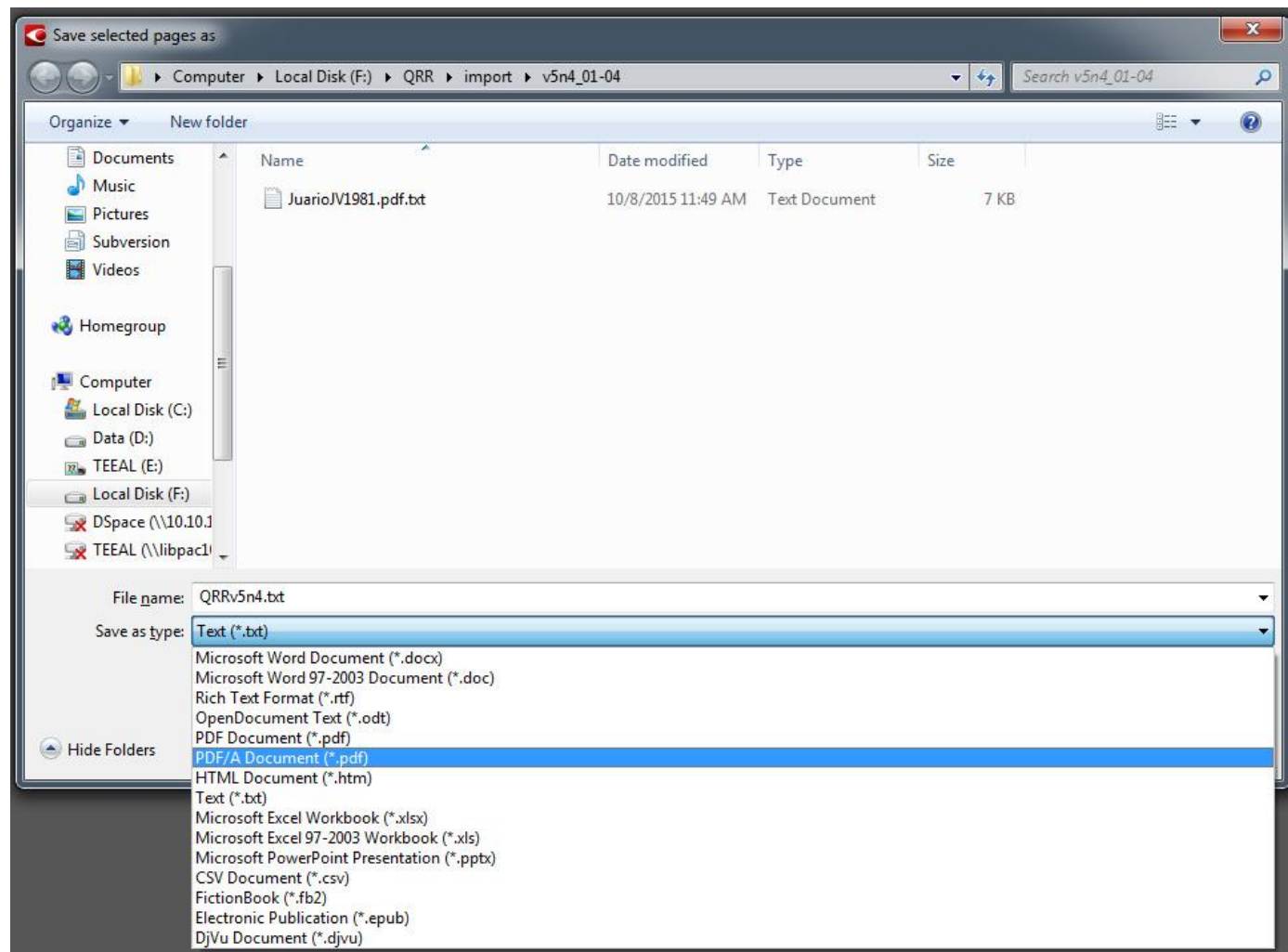




# Digitization: SEAFDEC/AQD Institutional Repository's Best Practices



# Digitization: SEAFDEC/AQD Institutional Repository's Best Practices



# Digitization: SEAFDEC/AQD Institutional Repository's Best Practices

Options

Document Scan/Open Read Save View Advanced

Document languages

English Edit Languages...

Document type

☒ Auto ☐ Fax ☐ Typewriter

Color mode

☒ Full color (preserves colors when adding pages)  
☐ Black and white (OCR becomes faster, but all colors are lost)

Document properties

☒ Save the document's metadata, including the following:

Title:

Authors:  ...

Subject:

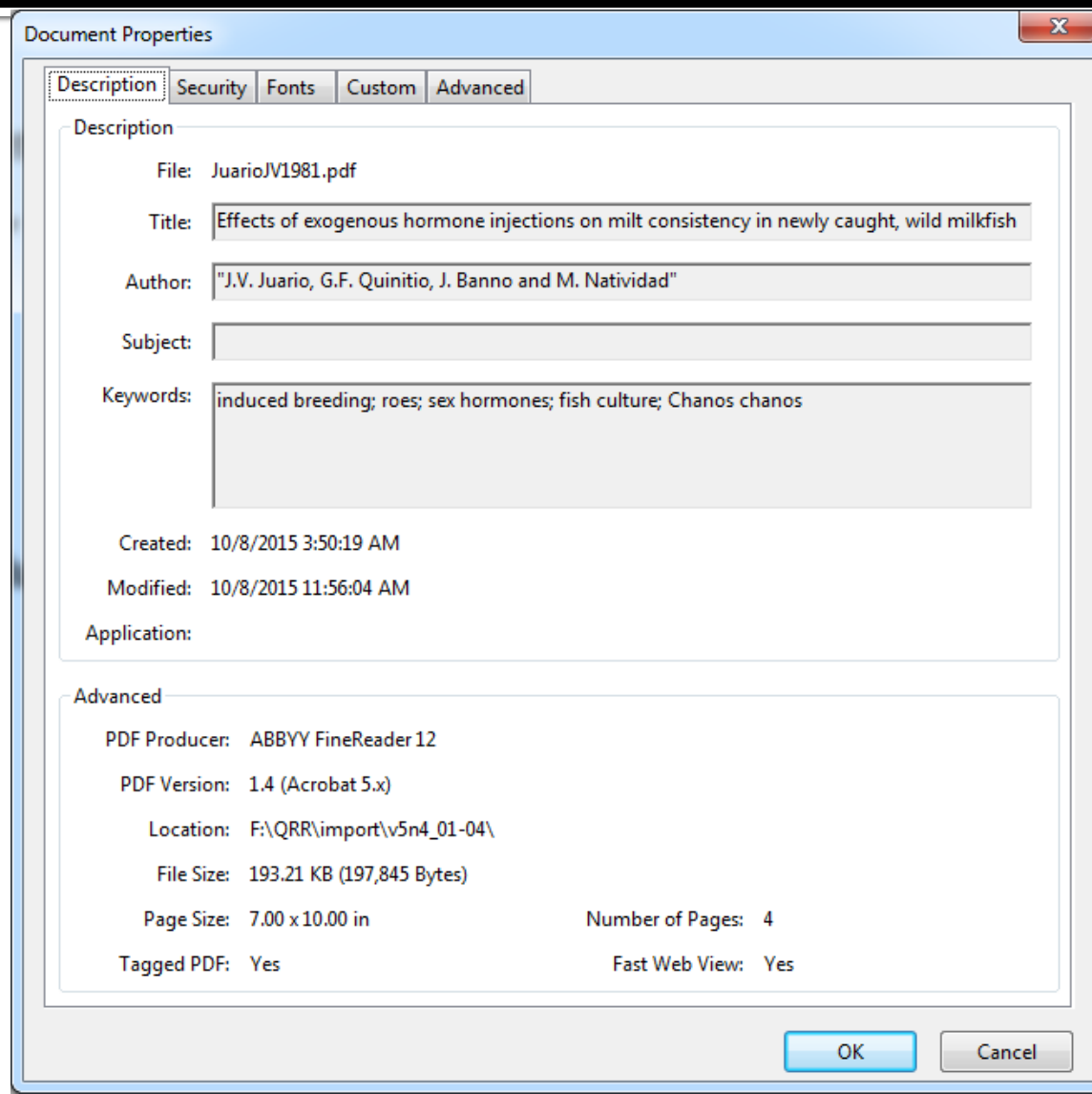
Keywords:

Document location

F:\QRR\v5\QRRv5n4

OK Cancel Help

# Digitization: SEAFDEC/AQD Institutional Repository's Best Practices



The image shows a 'Document Properties' dialog box with two tabs: 'Description' and 'Advanced'. The 'Description' tab is active, showing fields for File, Title, Author, Subject, and Keywords. The 'Advanced' tab is also visible, showing fields for PDF Producer, PDF Version, Location, File Size, Page Size, Number of Pages, Tagged PDF, and Fast Web View.

**Document Properties**

**Description** | Security | Fonts | Custom | Advanced

Description

File: JuarioJV1981.pdf

Title: Effects of exogenous hormone injections on milt consistency in newly caught, wild milkfish

Author: "J.V. Juario, G.F. Quinitio, J. Banno and M. Natividad"

Subject:

Keywords: induced breeding; roes; sex hormones; fish culture; Chanos chanos

Created: 10/8/2015 3:50:19 AM

Modified: 10/8/2015 11:56:04 AM

Application:

**Advanced**

PDF Producer: ABBYY FineReader 12

PDF Version: 1.4 (Acrobat 5.x)

Location: F:\QRR\import\v5n4\_01-04\

File Size: 193.21 KB (197,845 Bytes)

Page Size: 7.00 x 10.00 in

Number of Pages: 4

Tagged PDF: Yes

Fast Web View: Yes

OK Cancel

# Digitization: SEAFDEC/AQD Institutional Repository's Best Practices

## Mangrove-friendly Aquaculture Studies at the SEAFDEC Aquaculture Department

J.H. Primavera and A.T. Triño  
Southeast Asian Fisheries Development Center  
Aquaculture Department  
5021 Tigbauan, Iloilo, Philippines

### Abstract

The SEAFDEC Aquaculture Department studies on mangrove-friendly aquaculture (MFA) can be categorized under two models: a) mangrove filters where mangrove forests are used to absorb effluents from high-density culture ponds, and b) aquasilviculture or the low-density culture of crabs, shrimp and fish integrated with mangroves. In a study using the first model, shrimp pond effluents were retained in an enclosed mangrove area prior to release to receiving waters. Nutrients and other water quality parameters, and bacterial levels were monitored in the untreated effluents and post-mangrove water.

In the second MFA model, mangrove pens and ponds installed in old growth and newly regenerating mangrove sites in Aklan, central Philippines were stocked with mud crab *Scylla olivacea*/S. *tranquebarica* and shrimp *Penaeus monodon*. Investment costs, survival and production, and cost-return analysis for the pens and ponds are reported in the paper. Aside from the aquasilviculture trials in collaboration with local government units, other activities in the Aklan mangrove sites are the survey and mapping of the 75-ha area in Ibayay, construction of a treehouse, and the educational use as field site by Coastal Resource Management trainees of SEAFDEC Aquaculture Department and field biology students of the University of the Philippines in the Visayas.

### Introduction

Amidst the growing concern of international environmental non-government organizations (NGOs) over the ecological impacts of aquaculture, in early 1996 the SEAFDEC Council proactively mandated the Aquaculture Department to conduct studies on environment-friendly shrimp culture and to build up its expertise on mangroves. Under this initiative, a Mangrove-Friendly Aquaculture (MFA) Seminar-Workshop was organized in July 1999. The Workshop identified two MFA levels or models: a) mangrove as filters where the absorbing function of mangroves is used to process or treat effluents from high-density culture ponds, and b) aquasilviculture (or silvo-fisheries) where low-density culture ponds/pens are physically integrated with mangrove trees.

This paper reports on studies that fall under both MFA models.

# SEAFDEC/AQD Institutional Repository

[Login](#)[SAIR Home](#)

## DSpace @ SEAFDEC/AQD

Southeast Asian Fisheries Development Center, Aquaculture Department Institutional Repository (SAIR) is the official digital repository of scholarly and research information of the department. This is to enable the effective dissemination of AQD researchers' in-house and external publications for free and online. The repository uses DSpace, an open source software, developed at Massachusetts Institute of Technology (MIT) Libraries. It is an Open Archives Initiative (OAI)-compliant.

Initially, the repository shall contain preprints, full-texts or abstracts of journal articles, books and conference proceedings written by SEAFDEC/AQD scientists and researchers. The aim is to promote these publications especially those published in international peer-reviewed journals and generate higher citation through increased visibility.

It will also provide free access to all in-house publications of SEAFDEC/AQD. Full-text digitized copies of fishfarmer-friendly materials like books, handbooks, policy guidebooks, conference proceedings, extension manuals, institutional reports, annual reports (AQD Highlights), and newsletters (SEAFDEC Asian Aquaculture, Aqua Farm News, AquaDept News and AQD Matters) can be retrieved and downloaded.

In the future, SAIR will expand its collection to include images, presentations, audios, and videos among others.

The objectives of the repository are to: (1) to provide reliable means for SEAFDEC/AQD researchers to store, preserve and share their research outputs and (2) to provide easy access and increase the visibility of

### BROWSE

[All of SAIR](#)[Communities & Collections](#)[By Issue Date](#)[Authors](#)[Titles](#)[Subjects](#)

### MY ACCOUNT

[Login](#)[Register](#)

### DISCOVER

[Type](#)

**MARAMING  
SALAMAT!**