Department of Statistics MALAYSIA

For More Info : www.dosm.gov.my www.mycensus.gov.my

@StatsMalaysia @MyCensus2020

DOSM/BIPD/4.2020/Series 89

NEWSLETTER

StatsDigital

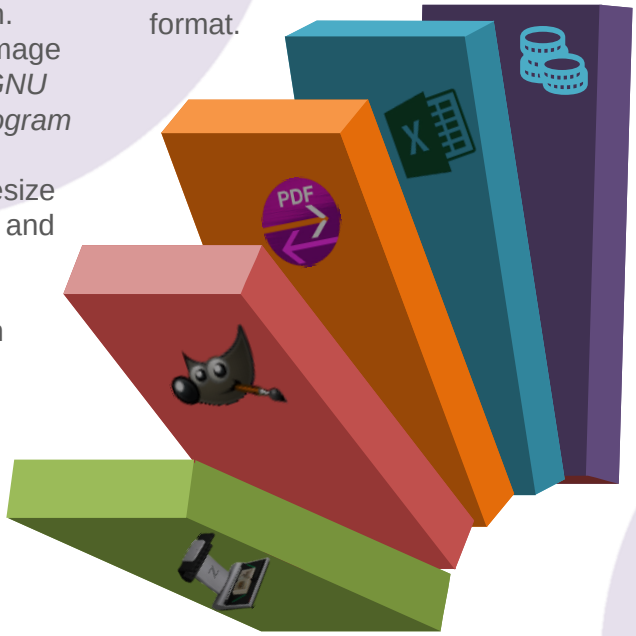
PREPARING SOURCES OF STATSDIGITAL

SCANNING

- Earliest step in the digitalization operation of the department's document.
- The scanner, *Zeutschel OmniScan 12002 Advanced Plus* (A2) is used to scan the publications.
- The scanned files are saved in TIFF format.

EDITING

- Editing is conducted to enhance the quality of the image. Image quality is very significant in the process of publication digitalization.
- Software used for the image editing purpose is the *GNU Image Manipulation Program* (GIMP).
- GIMP can be used to resize images, change colour, and cleaning unnecessary images/dirt.
- Image files are saved in TIFF format.



PROFILING

- Determination of the publication's confidentiality level .
- The confidentiality level consists of :
  - Confidential
  - General
  - Internal Distribution
  - Restricted
  - Secret

EXTRACTING DATASET

- Tables in PDF will be converted to Excel form.
- The converted table in Exce will be extracted to the CSV format using Excel Power Query.

OCR AND *e-BOOK*

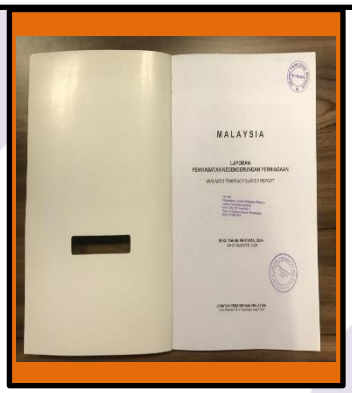
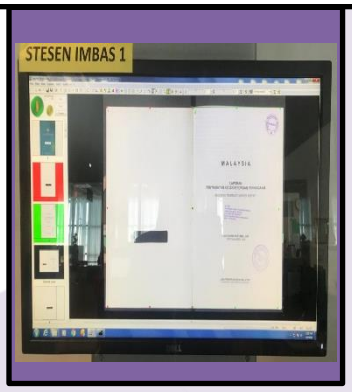
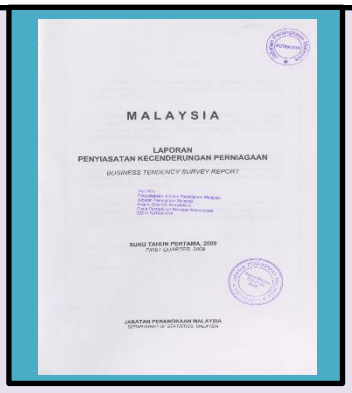
- *Optical Character Recognition* (OCR) is used to:
  - Create PDF file
  - Export PDF file to variety of formats
  - Reduce PDF size
  - Convert *non-searchable* PDF to *searchable* PDF

StatsDigital Portal contain

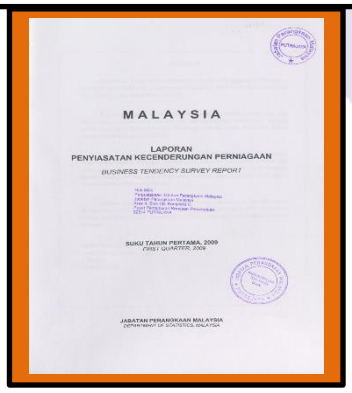
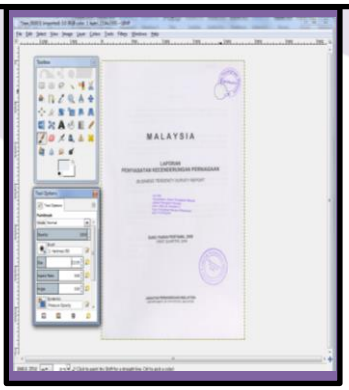
- 3 Categories
- 1,753 Publications*
- 17 Indicators
- 943 Datasets*
- 82 Sub-Indicators

* As at 2nd September 2020

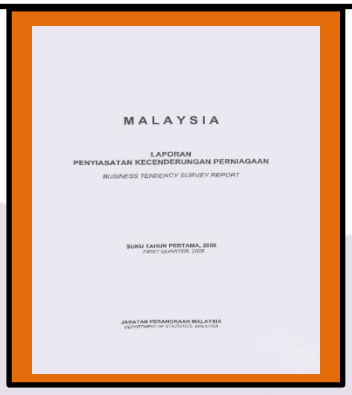
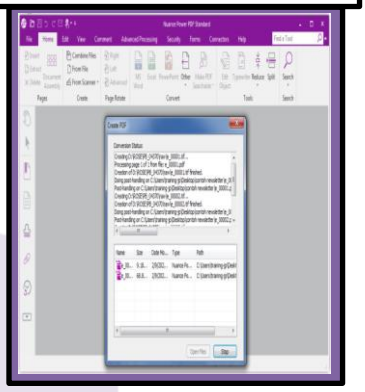
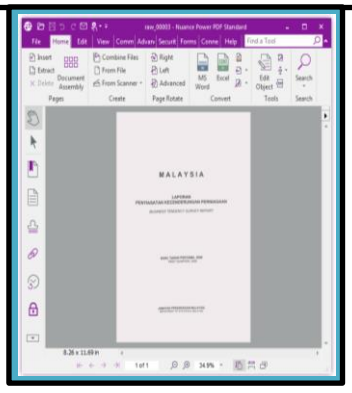
SCANNING PROCESS

INPUT	SCANNING	OUTPUT
		

EDITING PROCESS

INPUT	EDITING	OUTPUT
		

OCR & *e-BOOK* PROCESS

INPUT	OCR & *e-BOOK*	OUTPUT
		

EXTRACTION DATASET PROCESS

INPUT	DATA EXTRACTION	OUTPUT