WORKING WEEK 2025
AND Locate25
THE NATIONAL GEOSPATIAL CONFERENCE
FIG Geospatial Council of Australia
Collaboration, Innovation and Resilience: Championing a Digital Generation
Brisbane, Australia 6–10 April

Large-scale extraction of spatial data from PDF documents

Matthew Walker
Principal Visualisation Consultant
– matthew.walker@umwelt.com.au

Samuel Rundle
GIS Consultant
– samuel.rundle@umwelt.com.au

Presented at the FIG Working Week 2025 in Brisbane, Australia, 6-10 April 2025

umwelt

ORGANISED BY FIG Geospatial Council of Australia
PLATINUM SPONSORS Australian Government CHCNAV esri THE SCIENCE OF WHERE Leica Geosystems HydroMeter Surveyors Australia

Turning unstructured data into structured data

OCR

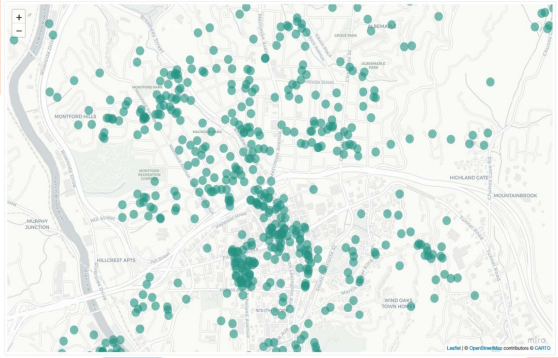
FME

PyMuPDF

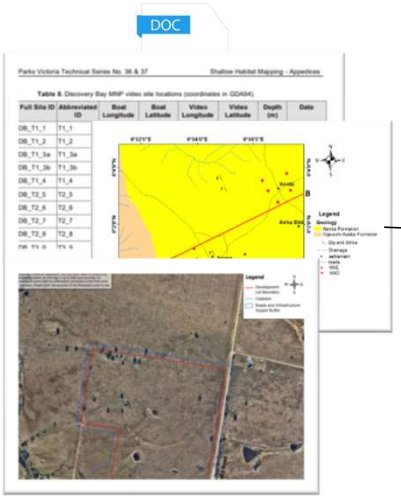
Python

Excel

ArcGIS Online



Unstructured data



Conversion of docs to per page PDFs

Consistent PDF data

OCR

PyMuPDF

Manual classification of image content

Maps

Tables

Tech diagrams

Auto-extraction of tables and coordinates

Python

Auto content extraction and plotting

Automated conversion to CSV

FME consistent conversion and metadata assembly

FME

Manual plotting

Manual Georeferencing

Spatial collation

Mapable data

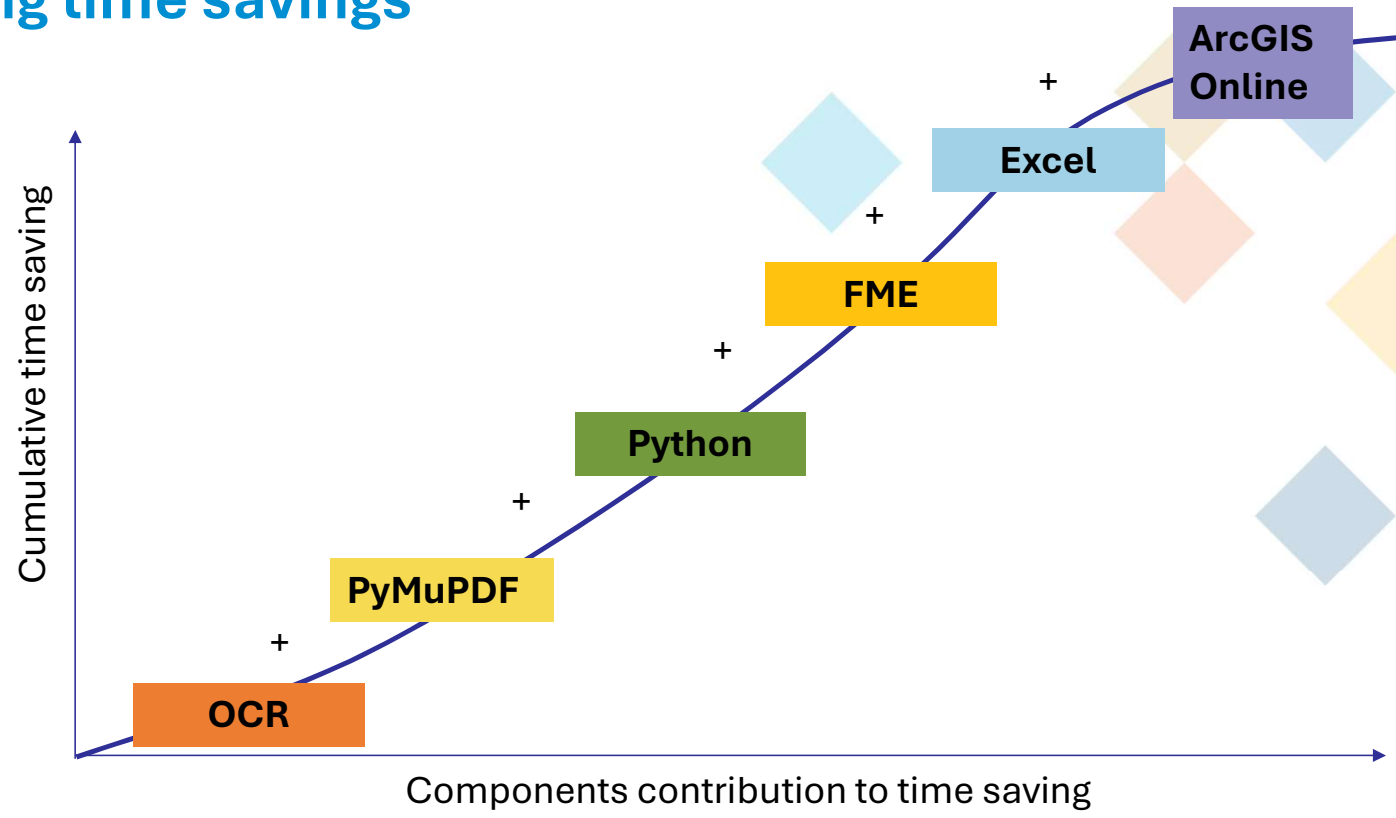
Publishing

Publish to online portal

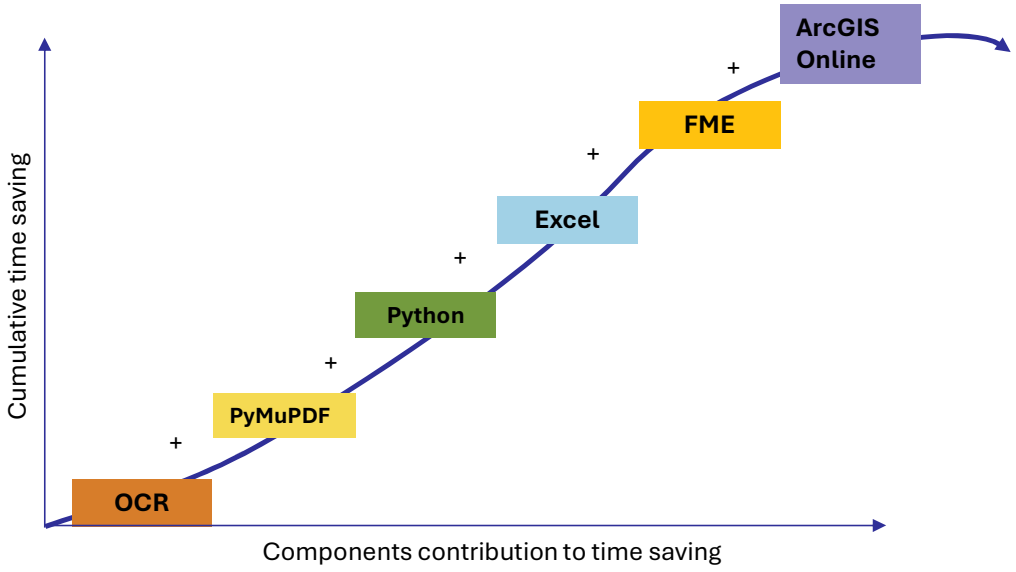
ArcGIS Online

Excel

Resulting time savings



Project Results



5500 documents containing over 500,000 pages

1500 spatial tables extracted, over 10,000 rasters extracted and georeferenced.
40,000 locations points derived from rasters

Learnings and areas for improvement

Data Quality is key and needs validation at every stage

OCR can introduce some interesting characters when given free reign and though they're not always visible, they can break downstream processes

Designing with the end in mind upfront is key

Over constraining schemes and attributes = data loss, Lack of structure = difficult to use data

Erroneous data in the original documents limits the entire remaining process (e.g. incorrect coordinates)

Future improvements – what can we improve on moving forward?

Utilising **Machine Learning Models** for better detection and segmented by content types

Automating **data validation** at each stage for best data throughput and reduction in QC

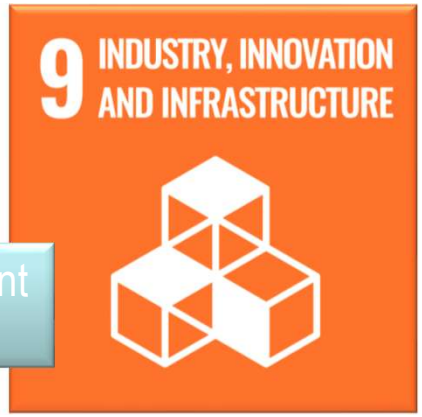
Automating **processing progress tracking** and reporting

Better methods of bundling **and batch uploading** resulting datasets

Automatic **Georeferencing** of extracted raster maps

The most relevant SDGs related to the presentation and theme of this session

1st relevant SDG



2nd relevant SDG

3rd relevant SDG

SUSTAINABLE DEVELOPMENT GOALS

International Federation of Surveyors supports the Sustainable Development Goals

Thank You!

Matthew Walker
Principal Visualisation Consultant
– matthew.walker@umwelt.com.au

Samuel Rundle
GIS Consultant
– samuel.rundle@umwelt.com.au

umwelt

The most relevant SDGs related to the presentation and theme of this session

1st relevant SDG

2nd relevant SDG

3rd relevant SDG

SUSTAINABLE DEVELOPMENT GOALS

International Federation of Surveyors supports the Sustainable Development Goals

STEP 1: SELECT HERE THE THREE MOST RELEVANT SDGs
STEP 2: COPY THE SDG INTO PREVIOUS SLIDE