

Research Process Documentation

Author: Richard Francis

Date: 2026.05.18

Purpose: Documenting the Behaviour → Systems → Intelligence pipeline

Contents

- [1. Ingestion](#)
- [2. Meaning Extraction](#)
- [3. Sync to MongoDB](#)
- [4. Live Intelligence](#)
- [Key Metrics](#)
- [Process Flows](#)

1. Ingestion

Raw files enter the system through the inbox. They are cleaned, normalised, renamed, and archived into canonical form.

- **Input:** Raw HTML, PDFs, notes, exports
- **Output:** Canonical archive (000001.html → N)
- **Processor:** organise.py

2. Meaning Extraction

Each archived file is analysed for structure, topics, metadata, and semantic meaning. Results are written to index.json.

- **Input:** Canonical archive
- **Output:** Structured JSON (index.json)
- **Processor:** process.py

3. Sync to MongoDB

The JSON index is atomically upserted into the MongoDB core collection, ensuring the research graph stays consistent.

- **Input:** index.json
- **Output:** MongoDB documents
- **Processor:** syncMongo.py

4. Live Intelligence

The website and API read directly from MongoDB to surface evidence, projects, and results.

- **API:** /api/files, /api/files/:id
- **Frontend:** Buzztrap template
- **Store:** rf_core.files

Key Metrics

- **Files Processed:** —
- **JSON Records:** —
- **Mongo Documents:** —
- **Fresh Upserts:** —

Process Flows

Stage	Description
Behaviour	Files dropped into inbox
Systems	Python processors clean, extract, and structure
Intelligence	MongoDB + API + Buzztrap render live results

The complete process output

```
In [11]: ### organise.py
```

```
[Running] python -u "d:\Richard\DataCore\ResearchProcessor\organise.py"
Processed: aws-boto3.html -> 000011.html Processed: aws-sagemaker-get-
started-tutorial.html -> 000012.html Processed: bs4_perfect_soup_bbc.html ->
000013.html Processed: chatGPT_4.11.23.html -> 000014.html Processed: CN7030
Tutorials.html -> 000015.html Processed: cnx.html -> 000016.html Processed:
course-work.html -> 000017.html Processed: dataframes.html -> 000018.html
Processed: django-aws-rf-ai-integration.html -> 000019.html Processed: django-
solution.html -> 000020.html Processed: emma-setUp.html -> 000021.html
Processed: free-aws-courses.html -> 000022.html Processed: functions_w3.html -
> 000023.html Processed: gp-architecture2.html -> 000024.html Processed: gp-
vpc.html -> 000025.html Processed: gpArchitecture.html -> 000026.html
Processed: home-work.html -> 000027.html Processed:
kinesis_data_streams_cli.html -> 000028.html Processed: lambda_practice.html ->
000029.html Processed: LinearRegression.html -> 000030.html Processed:
MASTER_attach_send_email.html -> 000031.html Processed: match-matrix.html -
> 000032.html Processed: Multi-VPC-Architecture.html -> 000033.html
Processed: my_email.html -> 000034.html Processed: nba_geeksforgeeks.html ->
000035.html Processed: nba_real_python.html -> 000036.html Processed:
notebook-export-template.html -> 000037.html Processed: notebook-template-
export.html -> 000038.html Processed: notebook-template-export_v1.html ->
000039.html Processed: pandas_sql_scrape.html -> 000040.html Processed:
premier_league.html -> 000041.html Processed: pysparkpractice.html ->
000042.html Processed: read-s3-csv-jupyter.html -> 000043.html Processed: rf-
hardware.html -> 000044.html Processed: scrabeBBC_v3.3.html -> 000045.html
Processed: scrapeBBC_complete_ai_function.html -> 000046.html Processed:
scrapeBBC_complete_pldb.html -> 000047.html Processed: spark_describe.html -
> 000048.html Processed: sqlite_euro2020.html -> 000049.html Processed:
startup-guide.html -> 000050.html Processed: statistical_analysis.html ->
000051.html Processed: traffic_data_descriptive_analysis.html -> 000052.html
Processed: unsw_nb15_descriptive_analysis.html -> 000053.html Processed:
w3schools_ML.html -> 000054.html Archive updated.
```

[Done] exited with code=0 in 3.176 seconds

In [12]: `### process.py`

```
[Running] python -u "d:\Richard\DataCore\ResearchProcessor\process.py"
Meaning extracted -> JSON updated.
```

[Done] exited with code=0 in 0.419 seconds

In [13]: `### syncMongo.py`

[Running] python -u "d:\Richard\DataCore\ResearchProcesso

[RF-SYNC] Commencing JSON -> MongoDB Atomic Upload...

[JSON] Success: Loaded 54 records from index.json [MONGO] Initializing pipe connection to mongoddb://192.168.1.199:27017/... [MONGO] Blasting bulk updates (54 documents)...

[SUCCESS] MongoDB core collection successfully updated. -> Processed Items: 54 -> Fresh Upserts: 44 -> Modified Updates: 0

[Done] exited with code=0 in 2.724 seconds

 Results