Work Plan – Draft for the First Submission (≤ 4 pages)

Provisional title: File Search using Elasticsearch  
Student: Rodrigo Allende Rial  
Supervisor: Víctor Rodríguez Doncel  
School / Degree: ETSIINF – UPM / Computer Engineering  
Date: 02-10-2025  
Version: v0.1

This document is tailored for the end‑of‑week‑3 submission. It must fit in four pages and include four sections: (1) General description + objectives, (2) Task list, (3) Gantt chart, (4) Copy of the supervisor’s proposal. Language: English.

1. General description and objectives

1.1 Executive summary

This project aims to develop a command-line tool for Windows that allows fast indexing and searching of documents in multiple formats (TXT, PDF, etc.), using Apache Tika for content extraction and Elasticsearch for indexing and retrieval. The tool will serve as a flexible and transparent alternative to Windows Search, enabling advanced queries and metadata filtering. The work plan includes learning the required technologies, designing the architecture, implementing indexing and search modules, and performing validation against Windows Search in terms of speed and accuracy. The final outcome will also include a proposal for integrating the tool with AI systems, thus extending its potential applications.

1.2 Scope

Included: local indexing on Windows (TXT, PDF), content + metadata extraction, text search with filters, filesystem change tracking (create/update/delete), CLI and setup scripts.  
Excluded (this phase): GUI, OCR for scanned images, full AI integration (design only).

1.3 Objectives

• O1. Design and implement a CLI to index and search local documents on Windows.

• O2. Evaluate performance and effectiveness against Windows Search with a reproducible protocol.

• O3. Design (not implement) an extension for integration with personal AI

systems.

1.4 Measurable targets (SMART)

• Indexing: ≥ 2,000 docs in < 5 min (p50) on a reference machine.

• Query latency: p50 < 80 ms, p95 < 200 ms (k=10).

• Quality: Precision@10 ≥ 0.70 on a curated test set.

• Freshness: reflect filesystem changes in the index in < 10 s.

• Reproducibility: scripts + dataset to repeat the benchmark end‑to‑end.

2. Task list (WBS)

Goal: break work into manageable blocks with deliverables and acceptance criteria. Total effort: 324 hours.

ID

Task

Deliverable

Acceptance criteria

Estimate (h)

Weeks

WBS‑1

Technical ramp‑up (IR, ES, Tika, CLI/PowerShell)

Notes + minimal demos

Local examples run successfully

40

1–3

WBS‑2

Design (architecture, mappings, commands)

Design document

Supervisor approves v1

30

4–5

WBS‑3

Implementation I – Indexing

CLI init/index + logs

Sample folder indexes end‑to‑end

90

6–8

WBS‑4

Implementation II – Search

CLI search + filters

Returns top‑k and export

50

9–10

WBS‑5

Testing & validation

Benchmark report

Metrics + Windows Search comparison

40

11–12

WBS‑6

Documentation

Thesis draft + user manual

Supervisor review completed

47

13–16

Buffer / adjustments

—

Risks controlled

—

17–21

Initial risks (brief): problematic PDF parsers (mitigate with retries), index disk usage (tune mappings/analysers), FS events burstiness (debounce + periodic rescans).

3. Gantt chart (timeline)

Timeframe: academic weeks up to the final submission. Highlight official milestones.

Week

Dates

Main block

Milestone / Deliverable

1

08–14 Sep

WBS‑1 (IR/ES)

Minimal local demo

2

15–21 Sep

WBS‑1 (Tika/CLI)

Tika→ES example

3

22–28 Sep

Draft Work Plan

Submission to supervisor

4

29 Sep–05 Oct

WBS‑2 (architecture/mappings)

Design v1

5

06–12 Oct

WBS‑2 (CLI/spec)

Official Work Plan (12 Oct)

6–8

13 Oct–02 Nov

WBS‑3 (indexing + watcher)

Stable prototype

9

03–09 Nov

WBS‑4 (search)

Progress report (9 Nov)

10–12

10–30 Nov

WBS‑4/5

Initial benchmark

13–16

01–28 Dec

WBS‑6

Thesis + manual

17–21

29 Dec–28 Jan

Wrap‑up + defence

Final submission (28 Jan)

4. Copy of the supervisor’s proposal

Offer ID 9145 – “File Search using ElasticSearch” (UPM)  
Summary:

The student will develop a command-line tool for Windows that can index and search documents in different formats (TXT, PDF, and possibly others), extracting both their textual content and relevant metadata. Apache Tika and Elasticsearch will be used to create and maintain an up-to-date index, efficiently detecting changes in the file system. The result will be a more flexible and controllable alternative to Windows Search, providing the foundation for a personal AI-powered system.

Official URL: https://fi.upm.es/ccfi/tfgm/tutores/detalle\_oferta.php?id\_oferta=9145

Attachment:

Presentation quality (10%) – quick checklist

• ≤ 4 pages, clean layout and consistent typography.

• Clear, formal language; no filler.

• Consistency across objectives ↔ tasks ↔ Gantt.

• Explicit risks and assumptions.

• Section titles match the required names exactly.

Internal notes (not part of the submission)

• Own contribution vs. external sources: mark explicitly from now on for the final thesis.

• Templates: keep a Word/LaTeX version aligned with ETSIINF for the final document.

• Repository: maintain a changelog and meeting minutes.