0. 문서 메타

- 문서명: 업로드 적재 후 OCR/이미지 추출 및 결과 보기 포탈 요구사항 정의서
- 버전: v1.2
- 작성일: 2026-01-22
- 권한: USER / ADMIN
- 구성: Web(Nginx) + Front(React) + API(FastAPI) + Job Runner(별도 프로세스) + MinIO + PostgreSQL
- 핵심 시나리오: 업로드 시 적재만 수행 → 사용자 버튼으로 OCR/이미지 추출 실행 → 결과 보기/다운로드

---

1. 목적 및 범위

1.1 목적

사용자가 포탈에 로그인하여 파일을 업로드하면 시스템은 원본 파일을 MinIO 에 저장하고, 파일 메타데이터 및 MinIO 참조를 PostgreSQL 에 적재한다.
이후 사용자가 특정 파일에 대해 OCR 텍스트 추출 및 이미지 추출/생성을 버튼으로 실행하면, 시스템은 비동기 Job 으로 처리하여 결과를 저장하고,
사용자는 포탈에서 추출 결과를 조회(보기) 및 다운로드한다.

1.2 범위(포함)

- 인증/권한(USER/ADMIN)
- 업로드 적재(원본 MinIO 저장 + PostgreSQL 등록)
- 사용자 트리거 기반 OCR 텍스트 추출
- 사용자 트리거 기반 이미지 추출(embedded/렌더링/둘다)
- Job 상태/이력/재시도/복구(큐 없이 DB 기반)
- 결과 보기 기능(텍스트 뷰어, 이미지 갤러리/뷰어)
- 원본/이미지 다운로드(Presigned URL)
- 관리자 콘솔(사용자/정책/Job/감사로그/정합성 점검)
- 비기능(보안/성능/가용성/관측성/백업)

1.3 범위(확장/옵션)

- 텍스트全文 검색(PostgreSQL FTS/ES 연동)
- 악성파일 AV 스캔/DLP 연계
- 결과 이미지 ZIP 다운로드(서버 부하 고려)

---

2. 사용자 유형 및 권한

2.1 권한 정의

- USER
  - 파일 업로드/조회/다운로드
  - OCR 텍스트 추출 실행
  - 이미지 추출 실행
  - 추출 결과 보기/다운로드
  - 실패 재처리(정책)
  - 본인 파일 삭제(정책)
- ADMIN
  - 사용자 관리(USER/ADMIN)
  - 시스템 정책 관리(업로드/OCR/이미지/Presign/재시도/보존)
  - 전체 Job 모니터링/실패 분석/재시도/취소(옵션)
  - 감사로그/정합성 점검

2.2 접근 통제 원칙

- USER 는 본인 소유 파일 및 결과만 접근 가능
- ADMIN 은 전 범위 접근 가능
- Presigned URL 발급 전 권한 검증 필수

---

3. 시스템 구성(아키텍처) 요구사항(SR)

3.1 시스템 구성 요소 및 역할

- SR-001 Web 서버(Nginx)
  - 정적 프론트엔드(React) 파일을 제공해야 한다.
  - /api/* 요청을 API 서버로 Reverse Proxy 해야 한다.
  - TLS 종료(HTTPS), 정적 캐시, 기본 요청 제한 정책을 적용할 수 있어야 한다.
- SR-002 API 서버(FastAPI)
  - 인증/권한, 파일 메타 관리, Presigned URL 발급, Job 생성/조회, 결과 조회 API 를 제공해야 한다.
  - 바이너리(원본/이미지)를 직접 중계하지 않고 Presigned URL 방식을 기본으로 해야 한다.
- SR-003 Job Runner(큐 미사용)
  - API 와 분리된 프로세스/서비스로 운영해야 한다.
  - PostgreSQL 의 jobs 를 폴링하여 작업을 수행해야 한다.
  - 작업 할당은 원자적 락(중복 처리 방지)을 사용해야 한다.

- SR-004 Storage/DB
  - MinIO: 원본 및 이미지 저장
  - PostgreSQL: 텍스트 본문 + 메타데이터 + 참조 + Job + 감사로그 저장

※ 본 문서의 files/jobs/extracted_* 등 데이터 모델은 논리 모델이며, 구현 시 PostgreSQL 테이블/인덱스로 매핑하여 적용한다.

3.2 라우팅/도메인 규칙(권장)

- 동일 도메인 운영 권장:
  - https://portal.domain/ (Nginx 정적)
  - https://portal.domain/api/* (Nginx → FastAPI)
- 별도 도메인 운영은 옵션: api.domain (CORS/쿠키 정책 추가 요구)

---

4. 저장/적재 원칙

4.1 저장 원칙

- 업로드 시점: **적재만 수행(추출 수행하지 않음 기본)**
  - 원본: MinIO portal-raw
  - 메타/참조: PostgreSQL files
- 추출 시점(사용자 트리거):
  - OCR 텍스트: PostgreSQL extracted_texts.text 에 본문 저장
  - 이미지: MinIO portal-images 에 저장 + PostgreSQL extracted_images 참조 저장

4.2 MinIO 버킷/오브젝트키 규칙(필수)

- 버킷
  - portal-raw : 원본
  - portal-images : 이미지
- objectKey(권장)
  - 원본: raw/{userId}/{yyyy}/{mm}/{fileId}/{originalFileName}
  - 이미지: images/{userId}/{yyyy}/{mm}/{fileId}/{jobId}/{imageType}/p{pageNo}_{idx}.{ext}
- 버킷은 private 이며, 접근은 Presigned URL 을 기본으로 한다.

4.3 Presigned URL 정책(필수)

- 업로드/다운로드/미리보기는 Presigned URL 로 제공한다.
- URL 만료 TTL 은 정책으로 설정한다(예: 5~30 분).
- 발급 전 권한 검증을 수행해야 한다.

---

5. 시나리오(업무 프로세스)

5.1 업로드 적재 시나리오(추출 없음)

1. USER 로그인
2. 업로드 요청 → API 가 **업로드 Presigned URL** 발급
3. 브라우저가 MinIO 로 원본 직접 업로드
4. 업로드 완료 등록 호출 → API 가 `files` 생성(상태: STORED)

5.2 OCR 실행 및 결과 보기

1. USER 가 파일 상세에서 “OCR 텍스트 추출” 클릭
2. API 가 jobs 생성(jobType=OCR_TEXT, QUEUED)
3. Job Runner 가 Job 을 RUNNING 으로 할당 후 처리
4. 결과 텍스트를 extracted_texts.text 에 저장
5. USER 가 텍스트 뷰어에서 조회/다운로드

5.3 이미지 추출 실행 및 결과 보기

1. USER 가 “이미지 추출” 클릭
2. API 가 jobs 생성(jobType=IMAGE_EXTRACT, QUEUED)
3. Job Runner 가 embedded 추출 또는 렌더링 수행
4. 이미지를 MinIO 에 저장 + PostgreSQL 참조 저장
5. USER 가 갤러리/뷰어에서 조회/다운로드

---

6. 기능 요구사항(FR)

6.1 인증/세션

- **FR-001 로그인**: 로그인 후 서비스 이용
- **FR-002 세션 정책**: 만료/비활동 만료 설정
- **FR-003 접근 차단**: 미인증/무권한 요청 401/403

6.2 파일 업로드 적재(추출 없음)

- FR-010 업로드 UI
  - 드래그앤드롭/파일선택, 진행률 표시
- FR-011 업로드 제한 정책
  - 허용 확장자/최대 용량/업로드 개수 정책 설정
- FR-012 메타데이터 입력
  - 필수: 문서명(기본 파일명)
  - 선택: 태그/설명/분류
- FR-013 업로드 Presigned URL 발급
  - API 는 업로드용 Presigned URL 을 발급해야 한다.
- FR-014 업로드 완료 등록
  - 클라이언트는 업로드 완료 후 API 에 완료 등록을 호출해야 한다.
  - (옵션) API 는 MinIO 오브젝트 존재/ETag 확인 후 등록한다.
- FR-015 원본 참조 및 해시 저장
  - files.minio(bucket, objectKey) 저장
  - SHA-256 저장
- FR-016 파일 상태
  - STORED/DELETED 등 상태 관리

6.3 추출 실행(사용자 트리거)

6.3.1 OCR 텍스트 추출

- FR-020 OCR 실행
  - 파일 상세에서 “OCR 텍스트 추출” 버튼 제공
- FR-021 OCR 정책(ADMIN)
  - 언어(ko/en), OCR 조건, 품질 옵션(DPI 등) 설정 가능
- FR-022 OCR Job 생성
  - jobs 에 jobType=OCR_TEXT 생성(QUEUED)
- FR-023 OCR 결과 저장
  - extracted_texts.text 에 본문 저장
  - 저장 단위 정책: DOCUMENT 또는 PAGE
  - preview/textLength 저장
- FR-024 OCR 버전 관리
  - 동일 파일 다회 실행 시 결과 버전 이력 제공(최신 표시 정책)

6.3.2 이미지 추출/생성

- FR-030 이미지 추출 실행
  - 파일 상세에서 “이미지 추출” 버튼 제공

- FR-031 이미지 정책(ADMIN)
  - embedded/렌더링/둘다, 렌더 DPI 설정
- FR-032 이미지 Job 생성
  - jobType=IMAGE_EXTRACT 생성(QUEUED)
- FR-033 이미지 저장(MinIO)
  - portal-images 저장
- FR-034 이미지 참조 저장(PostgreSQL)
  - extracted_images.minio(bucket, objectKey) + 메타 저장

6.4 Job 관리(큐 없이 DB 기반)

- FR-040 Job 상태
  - QUEUED/RUNNING/SUCCEEDED/FAILED/PARTIAL(옵션)
- FR-041 원자적 작업 할당(중복 방지)
  - Job Runner는 findOneAndUpdate 등 원자적 락으로 QUEUED Job을 획득해야 한다.
- FR-042 단계 기록
  - OCR: LOAD_ORIGINAL → OCR → PERSIST_TEXT
  - 이미지: LOAD_ORIGINAL → EXTRACT/RENDER → PERSIST_IMAGES
- FR-043 재시도/백오프
  - retryCount/maxRetries/nextRetryAt 기반 재시도 정책 제공
- FR-044 워커 장애 복구
  - heartbeatAt 기반 RUNNING 장기 작업을 FAILED 또는 QUEUED로 복구(정책)
- FR-045 재처리
  - USER는 실패 작업 재처리 요청 가능(정책), ADMIN은 강제 재시도 가능

---

7. 결과 보기 기능(필수 추가)

7.1 텍스트 결과 보기(텍스트 뷰어)

- FR-050 텍스트 결과 조회
  - OCR 결과가 존재하면 사용자에게 조회 기능을 제공해야 한다.
  - 결과가 없으면 “미추출” 상태 및 OCR 실행 안내를 제공해야 한다.
- FR-051 뷰어 표시 방식
  - DOCUMENT: 전체 스크롤 보기
  - PAGE: 페이지 네비게이션(이전/다음/목록) 제공

- **FR-052 화면 내 검색**
  - 뷰어 내 문자열 검색(클라이언트 수준)을 제공해야 한다.
- **FR-053 텍스트 다운로드**
  - 텍스트를 txt 로 다운로드할 수 있어야 한다(서버 생성/스트리밍).
- **FR-054 버전 선택 조회**
  - OCR 결과 버전 목록을 제공하고 특정 버전을 선택해 볼 수 있어야 한다(정책).

7.2 이미지 결과 보기(갤러리/뷰어)

- **FR-060 이미지 결과 목록**
  - 이미지 결과가 존재하면 목록(썸네일) 조회 제공
  - 없으면 "미추출" 상태 및 이미지 추출 안내 제공
- **FR-061 갤러리 기능**
  - 썸네일 그리드, 페이지 번호/유형 표시
  - 필터: imageType(EMBEDDED/RENDERED)
  - 정렬: pageNo 기준
- **FR-062 이미지 뷰어**
  - 확대 보기, 이전/다음 이동
  - 메타 표시(pageNo, format, size, 생성시간)
- **FR-063 미리보기/다운로드 제공 방식**
  - Presigned URL 방식으로 제공(만료 시 재발급)
- **FR-064 이미지 다운로드**
  - 단일 이미지 다운로드 제공(Presigned URL)
  - (옵션) 전체 이미지 ZIP 다운로드

7.3 결과 보기 권한/감사

- **FR-070 권한**
  - 결과 보기/다운로드는 소유자(USER) 또는 ADMIN 만 가능
- **FR-071 감사로그**
  - 결과 조회/다운로드 이벤트를 감사로그에 기록(정책)
  - action 예: view_text, download_text, view_image, download_image

---

8. 조회/검색/상세

- **FR-080 파일 목록**
  - 검색(파일명/태그/기간/상태), 정렬, 페이징
- **FR-081 파일 상세**
  - 원본 정보 + 다운로드
  - OCR 탭(상태/버전/뷰어/다운로드)

    - 이미지 탭(갤러리/뷰어/다운로드)
    - Job 탭(이력/단계/오류/재처리)
- **FR-082 원본 다운로드**
    - Presigned URL 발급 후 MinIO 에서 직접 다운로드

---

9. 삭제/보존/정합성

- **FR-090 삭제 정책**
    - USER: 본인 파일 삭제 가능
    - 하드 삭제 시 MinIO(원본/이미지)와 PostgreSQL(텍스트/메타/Job) 동시 삭제
- **FR-091 보존 정책(ADMIN)**
    - 보존기간 설정 및 만료 처리(옵션)
- **FR-092 정합성 점검(ADMIN)**
    - PostgreSQL 참조 ↔ MinIO 오브젝트 존재 여부 점검 리포트
    - 고아 오브젝트 정리(옵션)

---

10. 관리자 기능(ADMIN)

- **FR-100 사용자 관리**
    - 사용자 생성/비활성화, 권한(USER/ADMIN) 부여
- **FR-101 정책 관리**
    - 업로드 제한, Presign TTL
    - OCR 언어/옵션
    - 이미지 추출 방식/DPI
    - 재시도/타임아웃/복구 정책
- **FR-102 Job 모니터링**
    - 전체 Job 조회/실패 원인/재시도/취소(옵션)
- **FR-103 감사로그 조회**
    - 로그인/업로드/추출 실행/결과 조회/다운로드/삭제/정책 변경

---

11. 비기능 요구사항(NFR)

11.1 보안

- NFR-001 HTTPS/TLS

- NFR-002 Presigned URL 만료 정책
- NFR-003 Presign 발급 전 권한 검증
- NFR-004 MinIO private 버킷 운영
- NFR-005 감사로그 접근 통제 및 보관 정책

11.2 성능/안정성

- NFR-010 업로드/다운로드는 MinIO 직통
- NFR-011 OCR/이미지 추출은 Job Runner 비동기
- NFR-012 대용량 처리 시 타임아웃/재시도 정책
- NFR-013 이미지 썸네일 제공 최적화(옵션): 캐시/프리페치

11.3 운영/관측성

- NFR-020 구조화 로그(JSON)
- NFR-021 모니터링 지표
  - Job 처리시간, 성공/실패율, MinIO/PostgreSQL 오류율, 재시도 횟수
- NFR-022 백업
  - PostgreSQL 백업
  - MinIO 백업/버저닝(옵션)
- NFR-023 정합성 점검 배치
  - 주기적 점검 및 리포트 제공

---

12. 화면 요구사항(UI)

- UI-001 로그인
- UI-002 업로드(적재 전용)
- UI-003 파일 목록
  - 상태 표기: STORED / OCR_DONE / IMG_DONE / OCR_FAIL / IMG_FAIL 등
- UI-004 파일 상세
  - 탭 1: 원본 정보/다운로드
  - 탭 2: OCR 결과(상태/버전/텍스트 뷰어/다운로드/OCR 실행 버튼)
  - 탭 3: 이미지 결과(상태/갤러리/뷰어/다운로드/이미지 추출 버튼)
  - 탭 4: Job 이력(단계/오류/재처리)
- UI-005 관리자 콘솔
  - 사용자/정책/Job/감사/정합성

---

13. 데이터 모델(요약)

13.1 files

- _id, ownerId, status(STORED/DELETED)
- originalFileName, contentType, sizeBytes, sha256
- minio: { bucket:'portal-raw', objectKey, etag? }
- metadata: { title, tags[], description }
- latestOcrJobId?, latestImageJobId?
- createdAt, updatedAt

13.2 jobs

- _id, fileId, ownerId
- jobType: OCR_TEXT | IMAGE_EXTRACT
- status: QUEUED | RUNNING | SUCCEEDED | FAILED | PARTIAL(옵션)
- steps[]: { name, status, startedAt, endedAt, errorCode?, errorMessage? }
- retryCount, maxRetries, nextRetryAt
- lockedBy, heartbeatAt
- createdAt, updatedAt

13.3 extracted_texts

- _id, fileId, jobId, version
- granularity: DOCUMENT | PAGE,pageNo?
- text, preview, textLength
- createdAt

13.4 extracted_images

- _id, fileId, jobId
- imageType: EMBEDDED | RENDERED
- pageNo?, indexInPage?
- minio: { bucket:'portal-images', objectKey }
- width, height, format, sizeBytes, sha256
- createdAt

13.5 audit_logs

- _id, actorId, role
- action(upload, run_ocr, run_image_extract, view_text, download_text, view_image, download_image, delete_file, change_policy 등)
- targetType(file/job/image/text), targetId, result, createdAt, ip, userAgent

14. 수용(검수) 기준

- 업로드 시 원본이 MinIO 에 저장되고 PostgreSQL files 에 참조가 생성된다.
- 업로드만으로 OCR/이미지 추출이 자동 실행되지 않는다(기본 정책).
- USER 가 OCR 실행 시 jobs(OCR_TEXT)가 생성되고 완료 후 extracted_texts.text 에 저장된다.
- USER 가 이미지 추출 실행 시 jobs(IMAGE_EXTRACT)가 생성되고 이미지가 MinIO 에 저장되며 extracted_images 에 참조가 저장된다.
- 텍스트 뷰어에서 결과를 조회하고 txt 로 다운로드할 수 있다.
- 이미지 갤러리/뷰어에서 결과를 조회하고 Presigned URL 로 다운로드할 수 있다.
- USER 는 본인 데이터만 접근 가능, ADMIN 은 전체 접근 가능하다.
- 결과 조회/다운로드/정책 변경 등 주요 이벤트가 감사로그에 기록된다(정책).