**SK**네트웍스 **Family AI**과정 **20**기

# 데이터 수집 및 수집 데이터 보고서

| | |
|---|---|
| 산출물 단계 | 데이터 수집 및 저장 |
| 평가 산출물 | 수집 데이터 보고서 |
| 제출 일자 | 2026-01-26 |
| 깃허브 경로 | https://github.com/SKNETWORKS-FAMILY-AICAMP/SKN20-FINAL-6TEAM/tree/main |
| 작성 팀원 | 정소영 |

## 1. 수집 데이터 개요

| 데이터명 | 수집 대상 | 수집 목적 | 사용 예정 기능 | 출처/저작권 |
|---|---|---|---|---|
| 법령 | 현행법 (중소기업청,고용노동부, 국세청,공정거래위원회)<br><br>기관별 법령 해석 질의응답<br><br>판례 | VectorDB에 적재하여 RAG를 사용, LLM의 환각을 줄이고 사용자의 챗봇에 대한 신뢰도를 높인다. | RAG,<br><br>노무,재무, 창업 관련 상담 | 국가법령정보센터; 오픈 공공 데이터 |
| 지원 사업 | 스타트업 지원 사업 공고,<br>업종별 지원 사업 공고 | | RAG,<br><br>지원 사업 상담 | K-Startup,기업마당; 오픈 공공 데이터 |
| 재무 | 세무 캘린더,<br>중소기업 재무 지원 | | RAG,<br>재무 상담 | 국세청 |

| | | | | |
|---|---|---|---|---|
| | 제도 | | | |
| 창업-스타트업 | 업종별 스타트업 절차 | | RAG, 창업 상담 | 국세청, 법제처,LOCAL DATA |
| 인사-노무 | 근로 계약 관련 자료, 임금 및 징계, 퇴사, 4대보험 등의 정보 | | RAG, 인사/노무 상담 | 고용노동부 |

**2.** 수집 방법 및 자동화 절차

- 수집 방식 : API 호출, 웹 크롤링
- 수집 도구 또는 스크립트 설명:
    - 사용한 언어/라이브러리: Python, 국가법령정보센터 공공데이터 API,K-Startup,기업 마당 API
    - 자동화 여부 및 주기: 지원 사업의 경우에는 VectorDB 배치를 통해서, 신규 지원 사업 또는 기간이 만료된 지원 사업을 업데이트 한다

- 예시 스크립트 또는 흐름도 첨부:

```
"""
대한민국 전체 법령 수집 스크립트

국가법령정보센터 Open API를 사용하여 모든 현행 법령을 수집합니다.
https://www.law.go.kr/LSW/openApi.do

사용법:
    python collect_all_laws.py --api-key YOUR_API_KEY
    또는 .env 파일에 LAW_API_KEY 설정 후:
    python collect_all_laws.py
"""

...
```

```python
# ============================================================
# 설정
# ============================================================
BASE_URL = "https://www.law.go.kr"
SEARCH_URL = f"{BASE_URL}/DRF/lawSearch.do"
DETAIL_URL = f"{BASE_URL}/DRF/lawService.do"

OUTPUT_DIR = Path(__file__).parent.parent / "data" / "law"
OUTPUT_FILE = OUTPUT_DIR / "01_laws_full.json"
CHECKPOINT_FILE = OUTPUT_DIR / "collection_checkpoint.json"

# API 호출 간격 (초) - 서버 부하 방지
API_DELAY = 0.3


class LawCollectorAll:
    """전체 법령 수집기"""

    def __init__(self, api_key: str):
        self.api_key = api_key
        self.collected_laws = []
        self.failed_laws = []
        self.checkpoint = {"last_page": 0, "total_collected": 0}

    def _get_text(self, element, tag: str) -> Optional[str]:
        """XML 요소에서 텍스트 추출"""
        if element is None:
            return None
        found = element.find(tag)
        if found is not None and found.text:
            return found.text.strip()
        return None

    async def get_total_count(self) -> int:
        """전체 법령 수 조회"""
        params = {
            "OC": self.api_key,
            "target": "law",
            "type": "XML",
            "display": 1,
            "page": 1,
```

```
        }

        async with httpx.AsyncClient(timeout=30.0) as client:
            response = await client.get(SEARCH_URL, params=params)
            response.raise_for_status()

            root = ET.fromstring(response.text)
            total_count = self._get_text(root, ".//totalCnt")

            return int(total_count) if total_count else 0
```

…

## 3. 데이터 설명 및 구성

### 3.1 파일 및 필드 설명

#### (1) 법령 데이터 (Law)

| 파일명 또는 테이블명 | 필드명 | 데이터 타입 | 설명 | 예시 |
|---|---|---|---|---|
| **01_laws_full.json** | type | string | 법 타입 | “현행법령” |
| | keywords | string | 관련 핵심 키워드 | “중소기업”,”소상공인”,”벤처”,”창업” |
| | total_count | int | 가져온 관련 법의 개수가 몇 개인지 | 53 |
| laws: | law_id | string | 법 번호 | “010165” |
| | name | string | 법 한글명 | “대·중소기업 상생협력 촉진에 관한 법률” |
| | ministry | string | 담당 기관 | “중소벤처기업부” |

| | | | | |
|---|---|---|---|---|
| | enforcement_date | string | 시행일 | “20260102” |
| articles: | number | string | 카테고리 숫자 | “2” |
| | title | string | 큰 제목 | “정의” |
| | content | string | 핵심 내용 | “제2조(정의) 이 법에서 사용하는 용어의 뜻은 다음과 같다 ...” |
| clauses: | | | 조항/조목 | |
| content: | number | string | 조목 번호 | "①" |
| | content | string | 조목 내용 | "① 중소벤처기업부장관은 관계..." |
| items: | number | string | 조항 번호 | “1.” |
| | content | string | 조항 내용 | “1. \"중소기업\"이란 「중소기업기본법」 …” |

**(2)** 법령 해석 데이터 **(Law)**

| 파일명 또는 테이블명 | 필드명 | 데이터 타입 | 설명 | 예시 |
|---|---|---|---|---|
| **02_smba_expc_full ~ 05_ftc_expc_full.json** | type | string | 법 타입 | “법령해석례” |
| | org | string | 법 담당 기관 | “중소벤처기업부” |
| | total_available | string | 전체 이용 가능 개수 | “8599” |

| | | | | |
|---|---|---|---|---|
| | collected_count | string | 담당 부서 선택 숫자 | “500” |
| items: | id | string | 고유 식별 코드 | “313107” |
| | title | stirng | 제목 | “1959년 12월 31일 이전에 퇴직한 군인의 퇴직급여금 지급에 관한특별법 시행령 제4조제2항 및 3항” |
| | case_no | string | 케이스 넘버 | “05-0096” |
| | answer_date | string | 대답한 날짜 | “20051223” |
| | answer_org | string | 대답한 기관 | “법제처” |
| | question_org | string | 질문한 기관 | “국방부” |
| | question_summary | string | 질문 내용 | “... 재직기간을 계산함에 있어 현역병 복무기간을 우선 공제하고 남은 기간 중 전투근무기간이 있는 경우 이를 3배로 하여 산정하여야 하는지 …” |
| | answer | string | 대답 내용 | “... 재직기간을 산정함에 있어서 우선 현역병 복무연한을 공제한 후 나머지 기간 중에 …” |
| | reason | string | 대답 이유 | “「1959년 12월 31일 이전에 퇴직한 군인의 |

| | | | | |
|---|---|---|---|---|
| | | | | 퇴직급여금 지급에 관한 특별법(이하 “특별법”이라 한다) 제2조」의 규정에 의하면 …” |

**(3)** 판례 데이터 **(Law)**

| 파일명 또는 테이블명 | 필드명 | 데이터 타입 | 설명 | 예시 |
|---|---|---|---|---|
| **prec_labor / tax_accouting.json** | type | string | 판례 | |
| | category | string | 종류 | 세무/회계 |
| | keywords | string | 필터 키워드 | [“법인세”,”부가가치세”...] |
| | total_count | int | 전체 개수 | 3068 |
| | collected_at | string | 취득 시간 | “2026-01-23T10:59:41.502923” |
| items | id | string | 고유 id | “613209” |
| | case_name | string | 판례 이름 | “법인세부과처분무효확인” |
| | case_no | string | 판례 번호 | “2025두34254” |
| | decision_date | string | 법원이 해당 사건에 대해 판결을 선고한 날짜(판결 선고일) | “20251204” |

| | | | | |
|---|---|---|---|---|
| | court_name | string | 판결 기관 | “대법원” |
| | court_type | string | 사건 분야 | “세무” |
| | decision_type | string | 판결 문서의 종류 | “판결” |
| | decision | string | 선고 결과 상태 | “선고” |
| | summary | string | 요약 | “[1] 과세예고통지 및 과전적부심사 제도의 … “ |
| | decision_summary | string | 판결 요약 | “[1] 과세예고통지는 과세관청이 조사한… “ |
| | reference | string | 참조한 법 | “[1] 헌법 제 12조 제1항, 국세기본법 제81조의15…” |
| | reference_cases | string | 해당 판례가 인용(참조)한 다른 판례들 | “[1] 대법원 2016.4.15 선고 2015두52326 판결 … “ |
| | full_text | string | 판례문 전문 | “[원고, 피상고인] ooo주식회사 (소송…” |

**(4)** 지원사업 데이터 **(Funding)**
**+** 부족한 컬럼 정보는 **HWP -> PDF** 변환 혹은 **PDF**를 통해 추출**(**지원 대상**/**제외 대상**)**

| 파일명 또는 테이블명 | 필드명 | 데이터 타입 | 설명 | 예시 |
|---|---|---|---|---|
| **extracted_sections.json** | | | | |
| metadata | input_directory | string | pdf 경로 | "D:\\f.pp\\kstartup-pdf" |

| | | | | |
|---|---|---|---|---|
| | total_files | int | 전체 파일 수 | 265 |
| | processed_files | int | 처리된 파일 수 | 265 |
| | files_with_keyword | int | 키워드 발견 파일 | 147 |
| | positive_keywords | list[string] | 긍정적인 키워드 | [“지원대상”,”지원대상” … ] |
| | negative_keywords | list[string] | 부정적인 키워드 | [“참여제한”,”참여제한” … ] |
| statistics | total | int | 전체 파일 | 265 |
| | processed | int | 처리된 파일 | 265 |
| | with_keywords | int | 키워드 발견 파일 | 147 |
| | with_positive | int | Positive 키워드 발견 | 139 |
| | with_negative | int | Negative 키워드 발견 | 86 |
| | skipped | int | 스킵된 파일 | 118 |
| | failed | int | 실패한 파일 | 0 |
| results | filename | string | 파일명.pdf | "(2026.01.07.) 2026년 협성대학교 창업보육센터 상반기 신규 입주기업 모집공고.pdf" |

| | | | | |
|---|---|---|---|---|
| | filepath | string | 전체 경로 | "D:\\f.pp\\kstartup-pdf\\(2026.01.07.) 2026년 협성대학교 창업보육센터 상반기 신규 입주기업 모집공고.pdf" |
| | total_pages | int | 페이지 수 | 5 |
| | total_chars | int | 총 글자 수 | 2890 |
| sections | keyword | string | 발견된 키워드 | "입주자격" |
| | keyword_type | string | positive/negative | "positive" |
| | content | string | 추출된 본문 내용 | "○ 입주자격 : 아래 [표1. 입주자격] 참조\n\n--- PAGE 2 ---\n\n- 2 -\n[표1. 입주자격]\n구분\n주 요 내 용\n공통\n자격\n1) 창업 후 7년 이내 중소기업\n- 사업개시일: 법인등기일(법인), 사업자등록일(개인…" |
| | page_start | int | 시작 페이지 | 1 |
| | page_end | int | 종료 페이지 | 1 |
| | keywords_found | string | 찾은 키워드 | “입주 대상”,”입주자격” |
| | has_positive | bull | true/false | true |

| | | | | |
|---|---|---|---|---|
| | has_negative | bull | true/false | false |

**(5) API** 응답 명세서

| 필드명 | 샘플데이터 | 항목설명 |
|---|---|---|
| intg_pbanc_yn | N | 통합 공고 여부 |
| intg_pbanc_biz_nm | | 통합 공고 사업명 |
| biz_pbanc_nm | 창업보육센터 입주기업 수출상담회 | 지원 사업 공고명 |
| pbanc_ctnt | | 공고 내용 |
| supt_biz_clsfc | 행사·네트워트 | 지원 분야 |
| aply_trgt_ctnt | | 신청 대상 내용 |
| supt_regin | 서울특별시 | 지역명 |
| pbanc_rcpt_bgng_dt | 2012-11-29 00:00:00 | 공고 접수 시작 일시 |
| pbanc_rcpt_end_dt | 2012-12-01 00:00:00 | 공고 접수 종료 일시 |
| pbanc_ntrp_nm | | 창업 지원 기관명 |
| sprv_inst | 공공기관 | 주관 기관 |
| biz_prch_dprt_nm | | 사업 담당자 부서명 |
| biz_gdnc_url | | 담당자 연락처 |
| prch_cnpl_no | | 사업 안내 URL |
| detl_pg_url | www.k-startup.go.kr/web/contents/web/co | 상세페이지URL |

| | | |
|---|---|---|
| | ntents/bizpbanc-ongoing.do?schM=view&pbancSn=14212 | |
| aply_mthd_vst_rcpt_istc | | 신청 방법 방문 접수 설명 |
| aply_mthd_pssr_rcpt_istc | | 신청 방법 우편 접수 설명 |
| aply_mthd_fax_rcpt_istc | | 신청 방법 팩스 접수 설명 |
| aply_mthd_eml_rcpt_istc | | 신청 방법 이메일 접수 설명 |
| aply_mthd_onli_rcpt_istc | | 신청 방법 온라인 접수 설명 |
| aply_mthd_etc_istc | | 신청 방법 기타 설명 |
| aply_exclt_trgt_ctnt | | 신청제외대상내용 |
| aply_trgt | 청소년,대학생,일반인 | 신청 대상 |
| biz_enyy | 7년미만,5년미만,3년미만,2년미만,1년미만,예비창업자 | 창업 기간 |
| biz_trgt_age | 만 20세 미만, 만 20세 이상 ~ 만 39세 이하, 만 40세 이상 | 대상 연령 |
| prfn_matr | | 우대 사항 |
| Rcrt_prgs_yn | Y | 모집진행여부 |
| pbanc_sn | 1234 | 공고 일련번호 |

**(6)** 창업**/**지원 관련 가이드

| 파일명 또는 테이블명 | 필드명 | 데이터 타입 | 설명 | 예시 |
|---|---|---|---|---|
| **startup_procedures.json** | type | string | 데이터 유형(창업 절차 | “창업절차가이드” |

| | | | | |
|---|---|---|---|---|
| | | | 가이드 식별자) | |
| | source | string | 데이터 출처 사이트 | "생활법령정보, 국세청" |
| | collected_at | date (string) | 데이터 수집 일자 | "2026-01-23" |
| | total_categories | integer | 포함된 업종 카테고리 수 | 10 |
| | categories | array (object) | 업종별 창업 가이드 목록 | [ {...}, {...} ] |
| categories | category | string | 업종명 | “미용실” |
| | csmSeq | int | 생활법령 정보 콘텐츠 식별 번호 | 1009 |
| | source_url | string (URL) | 원문 출처 링크 | "https://easylaw.go.kr/CSP/CnpClsMain.laf?csmSeq=1009" |
| | items | array (object) | 세부 절차/규정 섹션 목록 | [ {...}, {...} ] |
| items | section | string | 정보 구분 섹션 제목 | “자격요건” |
| | content | string / array(object) | 설명 텍스트 또는 하위 항목 목록 | "미용사 면허..." |
| | steps | array(string) | 단계별 절차 목록 | ["면허취득", "입지선정", ...] |
| | procedure | object | 신고/등록 절차 상세 정보 | { "신고처": "..."} |

| | | | | |
|---|---|---|---|---|
| | required_documents | array(string) | 제출 서류 목록 | ["영업신고증", ...] |
| | administrative | string | 행정처분 내용 | “영업정지” |
| | criminal | array(object) | 형사처벌 항목 | [ { offense, penalty } ] |
| procedure | 신고대상자 | string | 신고 대상 조건 | "시설기준 적합자" |
| | 신고처 | string | 접수 기관 | "구청 위생과" |
| | 신고방식 | string | 신고 방법 | "전자문서 제출" |
| | 발급 | string | 처리 결과 | "즉시 교부" |

| 파일명 또는 테이블명 | 필드명 | 데이터 타입 | 설명 | 예시 |
|---|---|---|---|---|
| **startup_guide_complete.json** | type | string | 데이터 종류 | “통합_창업가이드” |
| | version | string | 스키마/데이터 버전 | “2.0” |
| | source | array[string] | 출처 목록 | ["국세청 업종코드", "생활법령정보...", "국세청..."] |
| | total_industries | number(int) | 업종 레코드 총 개수 | 1589 |
| | detailed_categories | number(int) | 생활법령 상세가 붙는 카테고리 수 | 10 |

| | | | | |
|---|---|---|---|---|
| | common_info | object | 전 업종 공통 안내 | |
| | industries | array[object] | 업종별 가이드 리스트 | |
| common_info.사업자등록 | 개념 | string | 사업자등록 기본 개념/주의 | "…사업자등록이 의무…명의를 빌려주는 행위는…금지" |
| | 신청시기 | string | 신청 기한 | "사업 개시일부터 20일 이내 (사전 신청 가능)" |
| | 신청장소 | string | 신청처 | “사업장 관할 세무서장” |
| | 발급기간 | string | 발급 소요 | “신청일부터 2일 이내 …” |
| | 명의대여_위험 | array[string] | 명의대여 리크 목록 | ["세금 책임…", "신용악화…", ...] |
| | 문의 | string | 문의처 | “국세상담센터 126” |
| common_info.사업자등록_제출서류 | 개인사업자 | array[string] | 개인사업자 제출서류 | ["사업자등록신청서", "임대차계약서 사본…", ...] |
| | 개인사업자_외국인추가 | array[string] | 외국인 추가서류 | ["재외국민등록부 등본", ...] |
| | 영리법인_본점 | array[string] | 영리법인 본 | |
| | 비영리내국법인_본점 | array[string] | 비영리 내국법인 본점 | |
| | 내국법인_지점 | array[string] | 내국법인 지점 | |
| | 외국법인_국내사업장 | array[string] | 외국법인 국내사업장 | |

| | | | | |
|---|---|---|---|---|
| | 종교단체 | array[string] | 종교단체 | ["사업자등록신청서", "…단체직인"] |
| industries | code | string | 업종코드 | “011000” |
| | name | string | 업종명 | "곡물 및 기타 식량작물 재배업" |
| | classification | object | 산업분류(대/중/소/세/세세) | |
| | license_type | string | 인허가 성격(신고/등록/허가 등) | “신고/등록” |
| | required_licenses | array[object] | 필요한 인허가/신고 목록 | |
| | related_laws | array[string] | 관련 법령 | ["축산법", "수산업법", ...] |
| | common_procedure | object | 공통 창업 흐름(step 1~7) | |
| | detail_level | string | 상세 수준 (basic/detailed) | `"basic"` 또는 `"detailed"` |
| | matched_category | string (optional) | (detailed 일 때) 매칭된 생활법령 카테고리 | "커피전문점(카페)" |
| | detailed_procedures | object (optional) | (detailed 일 때) 상세 절차 본문 | |
| ndustries[].classification | large.code / name | string | 대분류 코드/명 | "A" / "농업, 임업 및 어업" |
| | medium.code / name | string | 중분류 | "01" / "농업" |

| | | | | |
|---|---|---|---|---|
| | small.code / name | string | 소분류 | "011" / "작물 재배업" |
| | detail.code / name | string | 세분류 | "0110" / "곡물 및 기타 식량작물 재배업" |
| | sub_detail | string | 세세분류(텍스트) | "곡물 및 기타 식량작물 재배업" |
| industries[].required_licenses[] | type | string | 인허가/등록/신고 명칭 | “사업자등록” |
| `detail_level`이 `"detailed"`일 때만 존재 | required | bool | 필수여부 | true/false |
| | condition | string (optional) | 조건부일 때 조건 설명 | "축산업의 경우" |
| | authority | string | 담당 기관 | "세무서", "시군구청" 등 |
| industries[].common_procedure | step1 ~ step7 | string | 공통 창업 단계 문장 | "사업 계획 수립 및 입지 선정", "사업자등록 신청…", ... |
| industries[].detailed_procedures | category | string | 생활법령 카테고리 명 | "커피전문점(카페)" |
| | csmSeq | number(int) | 생활법령 카테고리 ID | 706 |
| | source_url | string(url) | 원문 URL | "https://easylaw.go.kr/...csmSeq=706" |
| | items | array[object] | 섹션 단위 상세 내용 | |
| items[] | section | string | 섹션 제목 | "개요", "창업 절차", "사전준비 단계"… |
| | content | string (optional) | 설명 본문 | "커피전문점은 … 휴게음식점에 해당" |

| | | | | |
|---|---|---|---|---|
| | steps | array[string] (optional) | 절차 목록(문자열 배열) | ["사전준비: …", "창업준비: …", ...] |
| | items | array[string] (optional) | 체크리스트/주의사항 목록 | ["독립창업 vs 가맹사업…", ...] |

| 파일명 또는 테이블명 | 필드명 | 데이터 타입 | 설명 | 예시 |
|---|---|---|---|---|
| **industry_startup_guide.json** | type | string | 데이터 종류 식별자 | "업종별_창업가이드" |
| | source | string | 데이터 출처 설명 | "국세청 업종코드 + 인허가정보 매핑" |
| | total_industries | number(int) | 전체 업종 레코드 수 | 1589 |
| | industries | array(object) | 업종별 창업 가이드 목록 | [ {…}, {…} ] |
| industries | code | string | 업종 코드 (국세청 표준) | "011000" |
| | name | string | 업종명 | "곡물 및 기타 식량작물 재배업" |
| | description | string | 업종 설명 (선택) | "" |
| | license_type | string | 인허가 유형 요약 | "신고/등록" |
| | required_documents | array(string) | 필요 서류 목록 (없으면 빈 배열) | [] |
| | related_laws | array(string) | 관련 법령 목록 | ["축산법", "수산업법"] |

| | | | | |
|---|---|---|---|---|
| | facility_standards | string | 시설 기준 요약 | “” |
| | qualification | string | 자격 요건 | “” |
| | note | string | 기타 참고사항 | “” |
| | classification | object | 산업 분류 체계 | { large, medium, small, detail, sub_detail } |
| | required_licenses | array(object) | 필요 인허가 목록 | [ {type, required…} ] |
| | common_procedure | object | 공통 창업 절차 단계 | { step1 ~ step7 } |
| classification | large.code | string | 대분류 코드 | "A" |
| | large.name | string | 대분류 명 | "농업, 임업 및 어업" |
| | medium.code | string | 중분류 코드 | "01" |
| | medium.name | string | 중분류 명 | "농업" |
| | small.code | string | 소분류 코드 | "011" |
| | small.name | string | 소분류 명 | "작물 재배업" |
| | detail.code | string | 세분류 코드 | "0110" |
| | detail.name | string | 세분류 명 | "곡물 및 기타 식량작물 재배업" |
| | sub_detail | string | 세세분류 텍스 | "곡물 및 기타 식량작물 재배업" |

| | | | | |
|---|---|---|---|---|
| required_licenses[] | type | string | 인허가 또는 신고 명칭 | “사업자등록” |
| | required | bool | 필수 여부 | true |
| | condition | string(optional) | 조건부 필요 시 조건 설명 | “축산업의 경우” |
| | authority | string | 담당 기관 | “세무서”,”시군구청” |
| common_procedure | step1 | string | 창업 1단계 | "사업 계획 수립 및 입지 선정" |
| | step2 | string | 창업 2단계 | "업종별 인허가 요건 확인" |
| | step3 | string | 창업 3단계 | "시설 기준 충족 및 필요 서류 준비" |
| | step4 | string | 창업 4단계 | "관할 관청에 인허가 신청" |
| | step5 | string | 창업 5단계 | "사업자등록 신청" |
| | step6 | string | 창업 6단계 | "4대 보험 가입 (직원 채용 시)" |
| | step7 | string | 창업 7단계 | "영업 개시" |

**3.2** 데이터 양

- 전체 수집 데이터 건수:
    - 법령 정보 : 18638개 (현행법 5539개, 각 법령 해석 8604개, 판례 4495개)
    - 지원 사업 정보 : 1280개 (기업마당 978개, K-Startup 302개)
    - 창업/지원 : 3개 (국세청 업종코드 + 인허가정보 매핑 + 생활법령정보센터)

- 재무/세무 : 2개 ([PDF] 2025 중소기업세제·세정지원 제도-불필요페이지삭제 138page, [CSV] 국세청_세무일정_20260101_(2025.12~2017.03) )
- 인사/노무 : 2개 ([PDF] 근로기준법 질의회시집(2018.4.~2023.6.) 615page, [PDF] [PDF]중소벤처기업 4대보험 신고 9page)

**3.3** 저장 위치 및 포맷

- 저장 경로

```
ata/processed/
├── laws/
│   ├── laws_full.jsonl          # 5,539건
│   └── law_lookup.json          # 법령명→ID 매핑
├── interpretations/
│   ├── smba_interp.jsonl        # 중소벤처기업부 500건
│   ├── labor_interp.jsonl       # 고용노동부 200건
│   ├── nts_interp.jsonl         # 국세청 200건
│   └── ftc_interp.jsonl         # 공정거래위원회 200건
├── guides/
│   ├── common_info.jsonl        # 공통 창업정보
│   ├── industries.jsonl         # 업종별 가이드 1,589건
│   └── pdf_guides.jsonl         # PDF 추출 섹션
└── schedules/
    └── tax_schedule.jsonl       # 세무일정 238건
```

- 저장 포맷: Json, PDF, csv
- 인코딩: UTF-8

**5.** 법적·윤리적 검토

- 개인정보 포함 여부:
  - 미포함

- 출처 및 사용권:
  - 공개 여부: 공개
  - 라이선스 또는 약관 검토 여부:
    무료로 국민들에게 제공되는 공공 데이터

**6.** 데이터 품질 및 정합성 관리 방안

- 중복 제거 기준:

    - 공고 내용의 맥락/의미가 동일할 시 제거

- 정합성 검증 방법:
    - 실제 공고 내용과 교차 검증

- 청킹 전략:
    - **[**법령**/PDF** 종류**]** 기계적인 길이(Character) 기준 분할을 지양하고, 의미 단위인 '조(Article)' 단위로 분할하여 문맥(Context) 보존.
    - **[**지원사업**]** 공고문의 '지원 자격', '제외 대상', '지원 내용' 등 의미론적 섹션(Section) 단위로 분할.

**7.** 변경 이력 및 보완 내역

| 변경일 | 변경자 | 변경 내용 | 비고 |
|---|---|---|---|
| 2026-01-19 | 정소영 | 수집 데이터 산출물 초안 작성 | 데이터 수집이 진행되며 수시로 수정 진행 |
| 2026-01-20 | 정소영,이경현 | API를 통해 받아온 법령 데이터 분석하여 컬럼 파악 | |
| 2026-01-21 | 정소영 | 저장위치 및 포맷 수정, 지원사업 데이터 수정 | |
| 2026-01-23 | 정소영 | 지원사업 데이터 수집 방향 수정되어, 산출물 전반적으로 수정 진행 | |
| 2026-01-26 | 정소영 | 추가 산출물 관련 컬럼 작성 진행 | |