

공 공 저 작 물 실 태 조 사

사이트맵 수집·정리 매뉴얼

공공기관 홈페이지를 사이트맵·GNB 단위로 점검해 엑셀로 정리하
는 재현 가능한 절차서

DB수집 프로젝트 · 2026

공공저작물 실태조사 대상 기관의 홈페이지를 사이트맵 단위로 점검해 엑셀에 정리하는 작업의 **재현 가능한 절차서**. 한 기관이 끝나면 다음 기관도 이 문서만 보고 동일하게 만들 수 있어야 한다.

기준 양식: `자료_취합_예시.xlsx` (시트명 예: `04_구소영`). 완료 예시: `충청남도\1.계룡시\계룡시.xlsx`.

절대규칙 — ☒ 검수완료 기관은 손대지 마라

작업 전 반드시 `작업파일\_작업현황.md` 를 먼저 읽고, ☒ 검수완료 로 표시된 기관은 재수집·수정·재계산·일괄작업·병합 무엇도 하지 않는다(읽기만). 일괄 스크립트는 검수완료 대상을 반드시 제외한다. (현재: 계룡·공주·금산·논산)

0. 입력·산출물

입력

- 기관명 (예: "계룡시", "공주시")
- 베이스 URL (예: `https://gyeryong.go.kr/kr/`)
- 저장 경로 (예: `충청남도\2.공주시\`)
- (선택) 담당자명 — 시트명 접두에 사용

산출물

- `{저장경로}\{광역}_{기관명}.xlsx` (예: `충청남도_계룡시.xlsx` . ★2026-06-01 파일명에 광역(도) 접두 부여 — 폴더명 `{번호}.``{기관}` 은 유지)
- 시트명: `{번호}_{기관명}` (예: `01_계룡시`)
- 행 1~2: `자료_취합_예시.xlsx` 의 헤더 그대로 유지
- 행 3~끝: 사이트맵에서 추출한 데이터

1. 단계 개요

단계	결과	진행 여부
Phase 1	사이트맵 → D~K열 채우기 + 병합·정렬·스타일	☑ 정립 완료
Phase 2	L(게시판형태) · M(수량)	☑ 정립 완료
Phase 3	N(저작물 유형)	☑ 정립 완료
Phase 4	O(공공누리 부착) · P(마크 부착위치) · Q(마크 하이퍼링크)	☑ 정립 완료
Phase 5	R(공공누리 연계) · S(비고) · T~W(검토) · Y~AA(완료체크)	📄 사용자 수기 영역

Phase 1 — 사이트맵 수집·정리 (현재 정립된 절차)

❌ 1-0. 권위 트리 = 사이트맵이 아니라 "본 페이지 메뉴(GNB/LNB)" (프리랜서2 #3·#16 — 최우선)

사이트맵에는 중간 메뉴가 누락된 경우가 매우 많다. 무조건 실제 홈페이지 메뉴(GNB 메가메뉴 / 좌측 LNB)를 권위 트리로 삼아 전수 작성한다. 사이트맵은 보조 참고일 뿐. 사이트맵만 믿으면 페이지·게시판 누락이 대량 발생(프리랜서2 지적 #16). - 절차: GNB/LNB를 라이브로 펼쳐(필요 시 Playwright JS렌더) 대→중→소→세부 트리를 추출 → 사이트맵과 **합집합**(사이트맵에만 있는 것도 보존). 누락 0이 목표. - 본문 탭(1-5b/c)·카테고리탭까지 합쳐야 진짜 전수. "사이트맵에 없으니 없다"는 금지. - 실제 적용: 장수(41→356)·교육청 3곳·전주 GNB 4차 등 — 전부 사이트맵 누락을 GNB/LNB로 복구.

1-1. 사이트맵 페이지 찾기 (보조 입력)

베이스 URL의 메인 페이지 푸터/헤더에서 "사이트맵" 링크를 찾는다. (단 1-0대로 본 페이지 메뉴가 권위, 사이트맵은 보조) - 일반 패턴: `/kr/html/guide/0701.html` (계룡시 등 다수 충남 시·군) - 다른 패턴 예: `/sitemap.do`, `/html/sub06/sitemap.html`, `/sitemap.html` - 못 찾으면 메인 페이지의 모든 `a` 태그에서 텍스트에 "사이트맵" / href에 `sitemap` 이 들어가는 것을 검색

1-2. 사이트맵 HTML 구조 파악

대부분의 시·군 홈페이지는 다음 두 패턴 중 하나를 따른다.

패턴 A — `dl > dt > dd > ul > li` (계룡시 형)

```
.sitemap
├─ dl (대분류 1개)
│  └─ dt → 대분류명          → D 메뉴명
│     └─ dd (중분류 1개)
│        └─ b > a → 중분류명    → E 카테고리(대)
│           └─ ul
│              └─ li > a → 소분류    → F 카테고리(중)
│                 └─ ul (중첩!) → li > a → G 카테고리(소)
│                    └─ ul → H, I, J ...
```

패턴 B — `ul.depth1 > li.depth2 > ul.depth3` (단순 중첩 ul) - depth 1 → D, depth 2 → E, depth 3 → F, depth 4 → G ...

핵심: `<li>` 안에 `<ul>` 이 또 있을 수 있다(예: 계룡시의 `<ul class="s_4th_ul">`). 평면적으로 파싱하면 자식들이 F=빈칸으로 떨어지니, 반드시 재귀로 깊이를 추적한다.

1-3. 컬럼 매핑

엑셀 열	의미	채워지는 값
B	순번	1부터 자동 부여, 행 삭제·삽입 후 재정렬
C	사이트명	기관명 고정 (예: "계룡시")
D	메뉴명	사이트맵 최상위 (dt)
E	카테고리(대)	그 다음 깊이
F	카테고리(중)	그 다음 깊이
G	카테고리(소)	그 다음 깊이
H	카테고리(소·세부사항)	그 다음 깊이
I	카테고리(소·세부사항2)	그 다음 깊이
J	카테고리(소·세부사항3)	가장 깊은 깊이 (그 이상은 J에 합침)
K	주소(url)	앞 노드의 href, 상대경로는 베이스로 절대화
S	비고	외부 도메인이면 "외부링크"

1-4. URL 처리 규칙

- `javascript:...` , `#` → 빈 값
- `http://` , `https://` 시작 → 그대로
- `/` 시작 → `https://{도메인}` + path 로 절대화
- 그 외 상대경로 → `urljoin(베이스, href)`
- 외부 판정: `http(s)://` 이면서 도메인이 기관 도메인이 아니면 외부 → S열에 "외부링크"

1-5. 부모-자식 URL 중복 제거 (G 활성화 룰)

`<li>` 안에 중첩 `<ul>` 이 있는 경우 부모 li의 a와 첫 자식 li의 a가 **같은 URL**을 가리키는 경우가 많다(예: 계룡시 "자동차 민원안내" = 01050601, 첫 자식 "신규등록" = 01050601).

규칙(★2026-05-30 전 깊이로 일반화): 1. 사이트맵 파싱 결과를 단순 행 리스트로 만든다 (부모 행 1개 + 자식 행 N개). 2. **임의 깊이**에서, 부모 행의 가장 깊은 카테고리 컬럼을 lc라 할 때(lc+1은 빈칸), 직후 자식 행이 ① D~lc 값이 모두 같고 ② lc+1 컬럼이 채워져 있고 ③ **K열 URL이 정확히 같으면** → **부모 행 삭제**. (D→E·E→F·F→G 모두 동일 적용. D=E=F가 같은 URL인 다중 중첩은 가장 깊은 행만 남김) 3. 카테고리 텍스트 (D/E/F)는 자식 그룹 병합으로 살아남는다. 4. URL이 다르면 부모 행 그대로 둠.

기존 구현(phase1)은 **F→G 한 깊이만** 처리했다. 중분류(E)·대분류(D)가 클릭 랜딩이고 그 URL이 첫 하위행과 같은 사이트(서산시 등 다수 eGov)는 D→E·E→F 중복이 남아 있었음 → `_스크립`

`트/_dedup_all.py` (dry/run)로 전 기관 일괄 정리(2026-05-30, 26개 기관 -791행). 신규 수집 시 phase1 단계에서 이 일반화 규칙을 적용할 것.

1-5b. 본문 탭(서브내비) → 하위 카테고리 확장 ★

사이트맵에는 한 줄로만 나오지만, 그 페이지 본문에 탭(서브내비)이 있어 실제로는 여러 하위 페이지를 거느리는 경우가 많다(예: 공주시 "납세자보호관" 페이지에 `납세자보호관 / 납세자관리현장 / 납세자보호관 업무추진 실적` 3개 탭). 이 탭들도 한 단계 아래 카테고리로 수집한다. (leaf가 F면 탭은 G, E면 F ...)

사이트마다 탭 마크업이 다르므로 매 기관 본문을 직접 열어 패턴을 확인한 뒤 셀렉터를 등록한다. 1-10에 기관별 탭 셀렉터를 기록.

확인 절차: `_tab_expand.py` 는 흔한 탭 클래스만 자동 탐지하므로, dry-run에서 "**확장 0건**"이 나와도 탭 없음으로 단정하지 말 것. 대표 페이지 2~3개의 본문을 직접 열어 탭 UL의 class를 확인 → `TAB_CLASS_PATS` 에 없는 이름이면 추가 후 재실행. 본문 탭이 실제로 없을 때만 "탭 없음"으로 확정한다.

확장 대상 판별 규칙(4조건 모두 충족할 때만 확장): 1. 탭 컨테이너 클래스가 본문에 있다. (공주시: `ul.tab-ul` . 흔한 후보: `tab-ul` · `tabmenu` · `tab-menu` · `tablist` · `subtab` 등. UL 자신의 class로 판별 — div 래퍼만 보고 오인하지 말 것) 2. 탭이 **2개 이상**이고, 모든 탭 href가 실제 페이지 링크다. → `#anchor` (인페이지 스크롤 탭, 예 `#nav1`) · `javascript:` · 빈 href-쿼리스트링만 다른 필터 탭(예 `?catecd=01`)은 제외. 3. 현재 페이지 URL이 탭 URL 집합에 포함된다(= 자기 자신의 탭 그룹). 자기와 무관한 상위 nav 오인 방지. 4. 탭 중 엑셀에 아직 없는 URL이 1개 이상. → 탭들이 이미 전부 사이트맵 행으로 존재하면(상위 공통 nav) 제외(중복 방지). 예: 공주시 정보공개 6탭 nav는 이미 F행으로 다 있으므로 재확장 안 함.

확장 방식: - leaf 컬럼(E~J 중 가장 깊은 값)을 부모로 두고, 탭들을 leaf+1 컬럼에 탭 순서대로 채운다. 부모 F는 자식들에 걸쳐 병합된다. - 현재 페이지와 같은 URL인 탭("on" 탭) = 원래 행 재사용 → 이미 채운 L~T(Phase 2~4) 데이터 보존, leaf+1 컬럼에 그 탭 라벨만 기입. - 나머지 새 탭 = 신규 행(L~T 비움 → Phase 2~4 별도 수집 대상). - "on" 탭 라벨이 부모(F)와 같아 G에 동일 텍스트가 한 번 더 들어가도 그대로 둔다(사용자 확장 방식).

일괄 적용(2026-05-29): 수집완료 39개 기관 전체에 `_스크립트/_tab_batch.py` (probe-dry-run 모드)로 일괄 수행. base/domain은 엑셀 K열 자동도출, 영동군만 `--weak-ssl` . 확장 11곳(+1,240행). **전북 14곳·제주·충남·충북 나머지는 본문 서브탭 실제 없음**(클래스 무관 탐지 `_tab_classfind.py` 로 검증 — 반복 UL은 전역 패밀리/유틸 nav). `--weak-ssl` 옵션을 `_tab_expand.py` · `_tab_phase234.py` 에 추가, dry먼저→신규행>0일 때만 write로 0건 불필요 재기록 방지.

스크립트: `_tab_expand.py` <엑셀> <BASE_URL> <도메인키워드> [`--write`] [`--weak-ssl`] — 인자 없이 실행하면 확장 계획만 출력(dry-run), `--write` 로 백업(`*_backup_tab전.xlsx`) 후 기입. 신규 행 L~P 수집은

`_tab_phase234.py` <엑셀> <Phase234 모듈> <도메인> [body_sel콤마] :

- **개별형**(공주시 `_phase234.py`, `get_body(soup)`): body_sel 생략 — 예 `_tab_phase234.py` 공주시.xlsx 충청남도/2.공주시/_phase234.py gongju.go.kr
- **일괄형**(`_chungnam_phase234_all.py` 등, `get_body(soup, sel)`): body_sel 지정 — 예 `_tab_phase234.py` 논산시.xlsx `_chungnam_phase234_all.py` nonsan.go.kr "#txt,#contents,main"
- `get_body` 인자 개수를 자동 감지해 두 형태 모두 대응. body_sel은 해당 모듈 `SITES[기관]` ['body_sel'] 값을 그대로 쓰면 됨.

△ fetch는 바이트(`resp.content`)로 받아 BeautifulSoup가 meta charset으로 인코딩 자동판별할 것. `apparent_encoding` 은 일부 페이지(공주시 읍면동)에서 오판해 한글이 깨진다.

1-5c. 탭 정밀화 — CMS 위젯 클래스·인페이지탭·카테고리탭 ★ (누적 재작업 교훈)

△ "탭없음 0건"은 신뢰 금지. 서산(0→577)·부여(222→351)·서귀포(367→502)·제주시(271→363) 모두 "탭없음" 오판이었다. 반드시 대표 페이지 본문 UL을 직접 확인.

- **위젯 클래스 변형** — `[class*=tab]` 는 `taps / tap (b 없는)` 위젯을 통째로 놓친다. 제주교육청 `ul.taps-dt4/taps-dt5(tapScroll)` 가 broad `[class*='tab']` 에서 누락됨(2026-06-12 발견). broad 스캔은 `[class*='tab'], [class*='tap']` 둘 다 매칭. 벤더별 실제 클래스: `tab-ul` (공주)·`tab_menu` (서산)·`basic_tab` (부여)·`ul.taps` (장수)·`div.tabmenu` (제주시·제주교육청)·`menu_tabmenu` (단양)·`ul.item3/item4` (진안)·`tabTy2/tab_st` (충남교육청)·`taps-dt4/dt5` (제주교육청).
- **인페이지(#앵커) 탭 2종 구분**:
 - 단순 스크롤 내비(섹션 점프, 본문이 한 페이지) → 그 자체는 행 추가 안 함(어문 섹션).
 - **콘텐츠 섹션탭**(`tapScroll/taps-dt5` 등 실제 하위 설명섹션을 탭으로 나눔) → **페이지 행이면 M=탭수** 기입 (인페이지탭=설명섹션). **게시판이면 M=총게시물 유지**(탭은 부가).
 - **카테고리 탭(categoryCode)**: 한 게시판(boardId 동일)을 카테고리로 나눈 탭(`categoryCode1=A/B/C/D` , 예 학교자료 운영계획서/운영현황/자체평가보고서/학교현장) → **각 카테고리를 leaf+1 별행**(게시판, M=카테고리별 총건수, boardId는 같아도 categoryCode로 구분). "전체"(categoryCode 없는 board)·주석처리(`<!-- -->`)된 li = 제외.
- △ **탭 ul 안의 li를 전부 분류하라**. 카테고리 탭 그룹에 **선행 일반메뉴**(`index.jje?menuCd=` 소개/랜딩 페이지) 가 섞여 있으면 그것도 별행(페이지)으로 수집 — 카테고리 4개만 보고 랜딩을 빠뜨리기 쉽다(학교자료 글로벌역량학교 등 15개 누락 사례, 2026-06-12). △ 사이트 자체 오링크 주의(미래기술인재학교 첫 li 가 타학교 menuCd를 가리킴 → 권위값=메뉴트리/소개페이지 menuCd 사용).
- △ **탭 앵커는 행번호가 아닌 URL로 배치**. 랜딩셀 제거 등으로 행이 밀리면 옛 행번호(stale anchor) 기반 삽입이 탭을 엉뚱한 부모 밑에 흘뿌린다. 메뉴페이지를 라이브 broad스캔→self-탭 URL이 그 행이면 거기 앵커, 나머지는 자식. 기존 L/M/N/O는 **URL 매핑으로 보존**.
- **중첩 탭(탭의 탭)·한 페이지 다중 탭그룹**: self-그룹 전부 union 후 **fixpoint 반복 전개**(새 탭 0까지). 충남교육청 관·과장(바깥 부서대분류 + 안쪽 부서) 사례.

1-5d. 리다이렉트 중복 — URL(menuCd) 다른데 같은 페이지 ★

일부 사이트(예: 고창군 `index.gochang?menuCd=...`)는 메뉴 카테고리(랜딩)를 클릭하면 실제 콘텐츠가 있는 자식 페이지로 자동 리다이렉트한다. 그래서 "랜딩 행"과 "자식 행"이 URL(menuCd)만 다르고 화면은 똑같다 — 같은 페이지인데 행이 둘로 부풀려진다.

△ 1-5(URL이 정확히 동일)로는 안 잡힌다. URL 문자열이 다르기 때문. 리다이렉트로만 같은 페이지가 된다.

탐지: 각 행의 K URL을 실제로 열어(`requests allow_redirects=True`) 최종 `resp.url` 을 비교한다. 최종 URL(또는 최종 `menuCd` + `contentsSid`)이 같은데 행이 둘 이상이면 같은 페이지. - 보조 신호: 두 행의 게시물 수 M 동일 + F(또는 E)라벨 동일 + 한쪽이 카테고리 랜딩(`menuCd` 끝 `...000000` 또는 첫 자식).

처리: 리다이렉트해 떠나는 셀(자기 `menuCd` ≠ 최종 `menuCd` , 그리고 최종 `menuCd` 가 다른 행으로 이미 존재) 행을 삭제, 최종(타겟) 행은 유지. 셀의 카테고리 텍스트(D/E/F)는 자식행에 병합으로 살아남으므로 손실 없음. flatten→타겟 제외→D/E/F 재병합(1-6).

예(고창군 2026-05-31): 군정알림→공지사항·고창소식→보도자료·문화관광→관광공사·참여마당→고창군에바란다 등 13건 셀 삭제(342→329). 전 행 fetch로 최종URL 충돌 탐지. 도구 1. 고창군/_del_dup.py .

1-5e. 랜딩셀 삭제 — 병합 그룹 첫 행이 빈 다음깊이인 메뉴 꺾이기 ★

카테고리 메뉴(대/중/소...)가 자기 자식 목록만 보여주는 섹션 랜딩 페이지일 때, 사이트맵에 그 랜딩이 한 행 + 자식들이 뒤따라 들어와서 병합 그룹의 첫 행이 "꺾이기"가 된다. 이 첫 행을 삭제하고 카테고리 라벨은 첫 자식이 승계한다.

△ 1-5(URL 정확히 동일)·1-5d(리다이렉트로 같은 화면)로는 안 잡힌다. 랜딩셀은 URL이 자식과 다르고(섹션 랜딩 `/field/welfare.do` · `/info/news.htm` 式) 리다이렉트도 안 되는, 독립된 섹션 인덱스 페이지다.

삭제 조건 (4개 모두 충족): 깊이 컬럼은 D(대)→E(중)→F(소)→G→H→I→J(세부). 임의의 깊이 컬럼 C_n ($n = D \sim I$)에 대해 1. C_n 이 여러 행에 걸쳐 병합돼 있고 2. 그 병합 그룹의 첫 행에서 바로 다음 깊이 C_{n+1} (오른쪽 한 칸, 최대 J)이 비어있음 3. K가 섹션 랜딩 URL(자식 목록만 거느린 카테고리 인덱스) 4. → 그 첫 행(셀)만 삭제, C_n 라벨은 첫 자식이 승계(병합 재계산).

즉 D병합+첫행 E빈칸 / E병합+첫행 F빈칸 / F병합+첫행 G빈칸 / G병합+첫행 H빈칸 ... 전 깊이 동일하게 적용.

△ 보존 예외 (삭제 금지): - 게시판·공공누리 부착 랜딩(M 또는 O 값 있음) — 실제 콘텐츠. - 단일행 리프(병합 아닌데 다음깊이 빈칸) — 예산정보·민원발급정부24式 실제 페이지/사이트. 반드시 "병합된 그룹"일 때만 셀로 본다. - 1-5(URL 동일)·1-5d(리다이렉트)에 해당하면 그 룰로 처리.

처리: unmerge 전체 → 셀 앵커의 D~J 라벨을 첫 자식이 승계 → bottom-up 삭제 → 신규행 remap → K 하이퍼링크 value 재생성 → B 재부번 → 1-6 재병합. 검증: 잔여 셀 0-K링크 불일치/누락 0-병합검침 0-B순번 연속.

도구: `_del_shell.py` (제주식 견고삭제, 전 깊이) · `_스크립트/_collapse_landing.py --safe` (전북 일괄, 순수 메뉴셀·페이지·M≤1·미부착만).

적용 이력: 전북 일괄 2026-05-31(`_collapse_landing.py` 98행) · 김제 -2 · 남원 -1 · 전주 -27 · 제주시 34행(E병합/F빈· `/field/welfare.do`) · 서귀포 36행(E병합/F빈· `/info/news.htm`) · 완주(검수時) · 정읍 23행(G병합/H빈, javascript 가짜탭 제거 병합). △ 한 깊이 더 깊은 변형(G/H 등)도 같은 룰. 자식 전부 페이지면 부모로 **접는** 재귀합치기(`_collapse_recursive.py`, 고창 F그룹)는 **삭제가 아니라 병합**이라 별건.

1-6. 병합·정렬 규칙

병합

- **D 열:** 같은 값이 연속되는 구간 전체를 병합 (대분류 단위)
- **E 열:** 같은 D 그룹 안에서 같은 E 값 연속 구간만 병합
- **F 열:** 같은 D+E 그룹 안에서 같은 F 값 연속 구간만 병합
- **G 이상:** 병합하지 않음
- **헤더 병합 보존:** `B1:R1`, `S1:W1`, `Y1:AA1` 세 개는 절대 풀지 말 것
- ★ **병합 순서:** 반드시 **F → E → D (D를 마지막에)**. E·F 병합은 상위 컬럼(D, D+E) 값으로 그룹을 묶는데, D를 먼저 병합하면 openpyxl이 블록 2행부터 D를 None으로 비워 E 그룹키가 깨진다 → 각 대메뉴(D) 블록의 **첫 행 E가 단독으로 남는 누락** 발생. D를 맨 마지막에 병합하면 해결. (2026-05-29 발견·수정. 기존 산출물은 `_fix_emerge_all.py` 로 재크롤링 없이 보정 완료)

정렬

- 병합/단독 무관 — D · E · F · G 셀: **가로 가운데, 세로 가운데, 줄바꿈 ON**
- **K 열:** **가로 왼쪽, 세로 가운데** (링크 본문 가독성)
- B(순번), Y/Z/AA(검토/완료) 등 헤더 기본 정렬 유지

1-7. 스타일 규칙

- **모든 행 높이 15** (헤더 1~2행 포함)
- **K열 하이퍼링크 활성화:**
- 셀의 `hyperlink` 속성에 URL 지정
- 폰트: 색 `0000FF`, 밑줄 `single`, 폰트명/크기는 기존 값 유지 (없으면 맑은 고딕 11)
- 외부링크인 행: S열에 "외부링크" 텍스트
- 셀 테두리·배경 등은 양식 파일의 기본 서식을 그대로 따른다

1-7b. 사용자 부분검수 보존 · 수정/신규셀 표기 · 국소삽입 ★ (누적 교훈)

재수집·정정 작업이 거의 항상 이미 사용자가 일부 검수한 파일 위에서 일어난다. 사용자 작업을 훼손하지 않는 것이 최우선.

- **부분검수 보존**: 사용자가 "N행까지(또는 1~N 순번) 검수했다/정답이다"라고 하면 그 범위는 **0셀 변경 보존**. 변경은 그 아래(또는 사용자가 지목한 범위)만. 작업 후 **diff로 보존 범위 0변경 검증**.
- **수정/신규셀 하늘색**: 내가 변경하거나 새로 추가한 셀은 **FFBDD7EE** (하늘색, **fg+bg 둘 다 지정** — fg만 주면 안 보임)로 표기해 사용자가 확인. "**적용은 내가 시킬 때**" 류 지시면 색칠만(=확인용)·실반영은 사용자 지시 시.
- **국소삽입(행 추가)**: 전체 재병합은 사용자 수동병합/B순번 스킴을 훼손하므로 회피. 절차 = **기존 병합 기록** → **unmerge 전체** → **insert_rows** → **generic remap(merge 시작: +삽입수<r0, 끝: +삽입수≤r1** → **부모/H 병합 자동확장**) → **재병합** → **K 하이퍼링크 진행 재적용**(insert_rows 버그). 전역 B재부번은 **사용자 수기스킴 있으면 회피**(신규행 B=None, 검수 시 재부번). 워크북 간 스타일은 **cell._style** 직접복사 금지 → 스타일 객체 **copy**.
- **라이브 사본 확인**: 작업 전 **프리랜서1** vs **제출_*** 등 2~3벌 중 라이브를 mtime/md5로 확인 ([submit_package_two_copies_gotcha]).

1-8. 검증 체크리스트

작업 완료 후 다음을 확인한다:

- [] 행 1~2 헤더가 원본과 동일하고 병합 3개(**B1:R1** **S1:W1** **Y1:AA1**) 유지
- [] 시트명이 **{번호}_{기관명}** 형식
- [] D 열 병합 개수 = 대분류 수 (계룡시: 7개)
- [] E 열 병합이 D 그룹을 넘지 않음
- [] F 열 병합이 D+E 그룹을 넘지 않음
- [] G 활성화된 부모 그룹은 첫 자식이 부모와 같은 URL이 아님 (중복 제거됨)
- [] K 열에 빈 값 / **javascript:** / **#** 가 없음
- [] K 열 셀 클릭 시 브라우저로 이동되는지 (하이퍼링크 활성화)
- [] 외부 도메인 행만 S='외부링크'
- [] 행 높이가 전부 15
- [] 순번(B)이 1부터 끝까지 연속

1-9. 작업 메모용 표 (다음 기관에서 채우기)

항목	값
기관명	(예: 공주시)
베이스 URL	(예: https://www.gongju.go.kr/)
사이트맵 URL	
대분류(dt) 개수	
총 잎 노드 수	
부모-자식 URL 중복으로 삭제된 행 수	
최종 행 수	
외부링크 행 수	
F 병합 그룹(자식 있는 부모) 수	
특이 사항	(예: 사이트맵 위치가 다른 패턴, 중첩 ul 클래스명 등)

1-10. 기록된 사이트별 셀렉터·통계

계룡시 (gyeryong.go.kr) — 2026-05-28 완료

- 사이트맵: `/kr/html/guide/0701.html`, 셀렉터 `.sitemap > dl > dt|dd > ul > li`, 중첩 `<ul class="s_4th_ul">`
- 본문(페이지/게시판 공통): `#txt`
- 페이지징: `.pagination`
- 검색 input: `<input type="text" name="sval" placeholder="검색어를 입력해주세요">`
- 전체 건수 텍스트: 총 게시물 N 건 또는 총 N 건 (정규식 `총\s*(?:게시물\s*)?([\d,]+\s*)`)
- 게시판 상세글 URL 패턴: `?mode=V&mng_no=...` 또는 `?mode=V&no=...&GotoPage=...`
- 통계 — 대분류 7, 총 행 442, F 병합 그룹 71, 외부링크 31, L 분포(페이지 288 · 게시판 122 · 사이트 31 · 접근 실패 1)
- 공공누리 부착 14건 (모두 Q=Y, 대부분 P=게시물). 미부착 396건
- O 이미지 기준 재판정(2026-05-29): 1유형 11·2유형 1(감사결과)·4유형 2(개인정보·공지사항). 링크 하드코딩(`licenseType4`) 불일치 12건 → S='링크주소 오기'

공주시 (www.gongju.go.kr) — 2026-05-28 완료

- 사이트맵: `/kr/sitemap.do`, 셀렉터 `.sitemap dl > dt|dd > ul > li`
- 특이점: a 태그 href가 더미(`#enURI1` 등), 실제 URL은 onclick의 `encodeURIComponent('...')` 인자. 정규식 `location\.href\s*=\s*encodeURIComponent\(\s*['"](?:[^\\"']+|\\.)*\s*\)`
- 사이트맵 중첩 ul 없음 (F 병합 그룹 0개)

- 본문(페이지/게시판 공통): `#txt`
- 페이지: `.pagination`
- 검색 input: `<input type="text">` — 단, 페이지 헤더 검색이 본문(#txt) 안에는 없으므로 #txt 내 검사로 한정 시 정상
- 전체 건수 텍스트: `총 게시물 N , 페이지 1 / N` 형태. "건" 미포함 → 정규식에 `총\s*게시물\s*([\d,]+)` 추가 필요
- 게시판 상세글 URL: `<a onclick="fn_search_detail('nttId'); return false;">` + `list.do` → `view.do?nttId={id}` 로 path 변환. 또는 첨부파일 td의 직접 `view.do?nttId=` 링크
- 통계 — 대분류 7, 총 행 306, F 병합 그룹 0, 외부링크 7, L 분포(페이지 218 · 게시판 81 · 사이트 7 · 접근 실패 1)
- 공공누리 부착 6건 (Q=Y 5건, Q=N 1건 — sub02_26_01.do 공공저작물 정책소개 1·2·3·4유형 마크 이미지 가 a 태그 밖)
- 본문 탭(서브내비): `ul.tab-ul` (실제 하위페이지). 2026-05-29 91개 그룹 확장 → **306행 → 592행(신규 286)**. `_tab_expand.py` (--write) + `_tab_phase234.py` (개별형, body_sel 생략)로 처리. 제외 케이스: `tab-ul type3 col*` (= #nav 인페이지 스크롤 탭)·정보공개 상위 6탭 nav(이미 사이트맵에 있어 중복)·쿼리필터 탭. 읍면동 페이지는 인코딩 자동판별(바이트) 필요. 신규 286행 L 분포(페이지 260 · 게시판 27), O 전부 미부착

금산군 (www.geumsan.go.kr) — 2026-05-28 완료

- 사이트맵: `/kr/html/sub07/0701.html`, 선택터 `.sitemap > dl > dt|dd > ul > li`, 중첩 `<ul class="s_4th_ul">` (계룡시와 동일 구조 — sub 번호만 차이)
- 본문(페이지/게시판 공통): `#txt`
- 페이지: `.pagination`
- 전체 건수 텍스트: `총 N 건` (계룡시 정규식 그대로 적용)
- 게시판 상세글 URL: `?`
`mode=V&site_dvs_cd=kr&menu_dvs_cd=...&code=...&no=...&GotoPage=...&group_at=...` — `mode=V` 조건만 봐도 충분히 추출됨
- Phase 1 통계 — 대분류 8(민원안내·정보공개·소통참여·금산소개·분야별정보·읍면·누리집가이드·부서안내), 원본 617행 → 부모-자식 URL 중복 93행 제거 → 최종 524행, F 병합 그룹 92, 외부링크 26
- Phase 2~4 통계 — L 분포(페이지 369 · 게시판 125 · 사이트 26 · 접근 실패 4), 공공누리 부착 65건(Q=Y 64·N 1), 미부착 429건
- 특이사항: DOM·게시판·선택터가 계룡시와 사실상 동일하여 계룡시 `_phase234.py` 를 XLSX 경로만 바꿔 재사용

무주군 (www.muju.go.kr) — 2026-06-02 F전개+탭+게시물KOGL 재수집

- 사이트맵: `/index.9is?contentUid=ff8080816db80238016dc8e98fa10ef6`, `div#sitemap > div > (h4=D) + ul>li(E) > ul.sitemap_blist>li(F)` **3단계**. (구 `parse_muju` 는 2단계만 파싱해 F누락 → **5.무주군/_muju_rebuild.py** 로 재구성)

- **정보공개 섹션 보강:** 사이트맵 h4가 `농특산물(s07)=실제 정보공개` 로 올라벨 + 해당 8개 E의 F가 접힘 → **GNB 메가메뉴**(`#gnb li.m0X>div.gnbW>ul.depth2>li(E)>ul.depth3>li(F)`)로 F 보강. D 라벨은 사이트맵 원본 유지.
- **리다이렉트 셀(1-5d):** 중분류 랜딩 fetch 최종 contentUid가 첫 자식과 같으면 셀 삭제. 외부는 netloc+path+query 비교.
- ★ **본문 하위탭 =** `ul.deviceN` (device2/3/4/5/6...). `li>a` 중 `class="on"` =현재페이지(=자기그룹). 조건: 탭≥2·실제 index.9is 링크·`on` 존재·신규 contentUid≥1 → **leaf+1 컬럼으로 전개**(F leaf면 G, E leaf면 F). 단일 sibling 그룹(중첩 없음 — 탭페이지가 보이는 device는 같은 그룹의 on이라 contentUid 가드로 재전개 안 됨). 47개 그룹 전개.
- 게시판 상세글 URL: `view.9is?dataUid=...` (목록 `list.9is` → `./view.9is`). **상세 10개**(없으면 건수대로).
- ★ **공공누리(O/P/Q) 무주 검출:**
- **게시판/페이지 자체 마크:** 본문의 `img_open(type|code){1~4}.(png|jpg|gif)` + `kogl.or.kr/info/licenseType{n}` 링크. (※ `01~04` 전 유형 = 공공저작물이용안내 **범례** → **미부착**, 4-1 규칙2)
- **게시물 마크:** 상세페이지(`div.bbsView`) 하단 `div.kogl-wrap` 의 koglType1~5가 전부 `display:none` 이고 스크립트 `$("#koglType{N}").css("display","block")` 로 활성화형 결정. **N=1~4=실제유형, N=5(opencode0)=자유이용불가=미부착, 숫자없음=미부착.**
- **O** = 게시판·게시물에서 발견된 유형 **전부 합집합**, 숫자 오름차순 콤마(공백없음).
- **P** = 게시판마크 있으면 **게시판** (둘 다 있어도 게시판 우선) / 게시물에만 있으면 **게시물** .
- **Q** = 게시판·게시물 중 **하나라도 licenseType 링크가 있으면 Y, 전부 이미지/텍스트뿐이면 N**. (무주 게시물 koglType은 링크 없음 → 게시물만 부착이면 보통 N)
- 도구: `5.무주군/_muju_phase234.py` (L/M/N/O/P/Q, device-aware, 상세10·bbsView·koglType스크립트).
- 처리 순서: `_muju_rebuild.py` (P1+탭) → `_muju_phase234.py` (P2~4) → `_스크립트/_redo_N_all.py run 무주군` (N 정교화).

Phase 2 — 게시판형태(L) · 수량(M)

2-1. L 판별 규칙

K열 URL에 접속하여 본문에서 다음 3요소 중 하나라도 있으면 "게시판", 아니면 "페이지"로 판별.

요소	의미	흔한 DOM 신호
전체 건수	"전체 N건" / "Total N" 같은 카운터	<code><div class="total"></code> , <code>#totalCount</code> , <code>span.total b</code> , <code><p class="page"></code> <code></code>
페이징	1, 2, 3, ... 다음 페이지 버튼	<code>.paging</code> , <code>.pagination</code> , <code>nav.paging</code> , <code>?GotoPage=</code> / <code>?pageIndex=</code> 형태 query
검색 기능	키워드 검색창 + 검색 버튼	<code><input type="search"></code> , <code><input name="searchKeyword"></code> , <code>.search_box</code> , <code><form class="search"></code>

판정 결과: - 3요소 중 하나라도 발견 → L = "게시판" - 하나도 없음 → L = "페이지" - S='외부링크'인 행 → L = "사이트" (URL 미접속, 위 검사 스킵)

★ 2-1a. 사이트(L) 판별 정밀화 (누적 교훈 + 프리랜서2 #7) - 현재 창이 아닌 다른 창/사이트로 넘어 가면 무조건 L="사이트" (새창 `target="_blank"` 포함, M/N/O/P/Q 전부 빈칸).

[[feedback_blank_newwindow_site]] - 예외 — 분야별 정보: 분야별 정보 "페이지"가 그 사이트 안에 따로 구성돼 있으면 그 페이지는 정상 작성. 다만 분야별 정보 항목 안에서 또 다른 사이트로 넘어가는 링크는 사이트 로 표기. - 외부·서브사이트 판별 = 최종 등록도메인이 기관 도메인과 다르거나, 같은 도메인이라도 비-www 서브사이트(news.school.jinhak. 등). fetch(allow_redirects) 후 최종 도메인 기준이 안전 — 로고 alt 휴리스틱은 부서명 이미지를 로고로 오인(오탐). 교육청 룰: 좌상단 로고가 그 기관(○○교육청)이 아니면 사이트. - epeople·국민신문고·정부24·위택스·부동산공시가격 등 외부 포털 임베드/이동 = 사이트.

★ 2-1b. 게시판 형태 판별 보강 (프리랜서2 #5·#10·#13) - 게시글 목록이 비어 보여도(글이 안 들어가도) 게시판 형태면 L="게시판" (예: 설문조사 페이지). 게시판 형태 정의는 2-1 3요소. - 게시판은 행 1개만. 게시판 안의 게시물(상세글)을 별도 행으로 추가하지 말 것(예 진안 BBS_0000057 문화/전시시설 보드의 용담호사진문학과 등은 행 추가 X → 게시판 1행, M=글수, P=게시물). - 한 페이지에 게시판이 여러 개면 전부 작성(예 서귀포 </info/news/law/law.htm> 게시판 3개). 1-5c 탭/섹션 스캔으로 누락 방지.

2-2. M(수량) 채움 규칙

L 값	M 값
게시판	페이지에 표시된 총 게시물 수 (숫자만) — 콤마·"건" 같은 단위 제거하고 정수로
페이지	1
사이트	빈칸

총 건수 추출 (우선순위): 1. "전체/총 N건" 카운터 셀렉터에서 텍스트 → 정규식 `(?:전체|총)\s*:\s*` `([\d,]+)` 숫자만 정수 변환 2. 카운터가 없으면 첫 글행의 번호(최신글번호 = 총건수). △ 삭제분이 있으면 실제게시물수보다 클 수 있음(사용자 컨벤션 = 번호 기준). 3. 페이지링만 있으면 마지막 실패 페이지 번호 × 페이지당. 불가능하면 0 + 비교(S)에 "수량 미확인".

★ 2-2a. M 카운트 함정 (누적 재작업 교훈) - △ 원본이 글수 대신 "페이지수"(≈글수/10)를 저장한 경우(아산시) → list를 in-page fetch해 첫 글행 td.no 최댓값으로 정정. 게시판 M이 비현실적으로 작으면 의심. - △ 번호 기반 카운트: 통계/소통 게시판은 첫 데이터행의 번호(`td.hidden-md-down` 등)가 최신글 번호=총건수. `act=view` ·사이드바 "N건" 노이즈로 오집계 주의(서귀포시). - △ 대형 edufine/외부연계 리스트 = 가짜 페이지: `linkPage(N)` 이 고정 상한이고 `pageIndex`가 범위를 넘어도 서버가 마지막 페이지를 반복 반환한다. 마지막 페이지 데이터가 바뀌는 지점(다음 페이지가 같은 데이터면 거기가 끝)으로 실제 끝을 확인. 여러 게시판이 동일한 거대값(예 6개 전부 16100)이면 = menuCd 필터가 안 걸린 통합뷰(같은 데이터)이므로 그대로 기록하되 의심·확인(서귀포교육청 계약현황). - 빈 게시판: M=0, 본문/상세 글 0이면 N= 없음 .

2-3. 예외 처리

- ★ 본인인증/로그인을 해야만 볼 수 있는 페이지 → 행 자체를 제거(프리랜서2 #15. 기존 "L=페이지 유지" 정정). 단순 게시판인데 일부 글만 로그인인 경우는 게시판 유지.
- 응답 실패 / 타임아웃 → `L=""` , `M=""` , `S` 에 "접근 실패" 기록
- 자물쇠 페이지(게시물 0건 포함) → `L` 은 정상 판별, `M=0` , N= 없음 , Phase 4의 O는 자동으로 "미부착"

Phase 3 — 저작물 유형(N)

3-1. 기본 룰

- 본문에 글자가 조금이라도 있으면 → "어문" 기본 포함
- 추가 미디어 감지 시 콤마(공백 없음)로 이어 붙임. 순서: 어문 → 이미지 → 영상 → 오디오
- ★ N 유형은 8종에서만 고른다(2026-06-12 프리랜서2 확정): 어문 · 이미지 · 영상 · 오디오 · 글꼴 · 3D · 기타 · 없음 . 카테고리 밖 표기 금지(예 "동영상"X → 영상).
- 글꼴 = 폰트 파일/서체 배포, 3D = 3D모델·VR, 기타 = 어느 유형도 아닌 것(아래 3-1b 링크전용·그래프 단독 등).
- 결과 형식 예: 어문 / 어문, 이미지 / 어문, 영상 / 어문, 이미지, 영상 / 어문, 이미지, 오디오
- 글자도 미디어도 전혀 없음 → 없음 . 게시판인데 게시물 0건 → 없음 (프리랜서2 #17).

★ 3-1a. "이미지" 판정 핵심 원칙 (2026-06-01 당진 / 2026-06-12 프리랜서2 정정)

이미지는 "별도 정보를 주는 그림"만. 글을 그대로 옮긴 것·그래프는 이미지가 아니다. - 이미지로 인정: 사진(전경·모습·행사), 평면도/배치도, 지도/약도, QR코드, 악보, 소식지·연보 표지, 상징물(심벌마크·시그너처·캘리그래피), 게시판 업로드 사진, 포스터·카드뉴스(그림+글자 함께). → 그림+글자가 같이 있으면 이미지. - 이미지 아님 = 어문: ①본문 텍스트를 그대로 이미지화한 "○○ 한눈에"식 인포그래픽·흐름도·절차도·조직도, ②글자만 있는 이미지(보고서 캡처·hwp 캡처본 등 = 글뿐이라 어문). - ★ 그래프(막대·선·파이)는 이미지가 아니다(프리랜서2 #4 — 기존 "그래프=이미지/어문" 모두 정정): 그래프 + 글자 = 어문 , 그래프만(글자 거의 없음) = 기타 . - 판단 기준: 그림이 텍스트에 없는 정보(공간배치·위치·실사진·악보 등)를 담으면 이미지, 텍스트를 그림으로 옮겼을 뿐이면 어문, 데이터 차트면 어문/기타.

★ 3-1b. 링크전용 페이지 = 기타 (프리랜서2 #2)

페이지(게시판 아님)에 링크/링크모음만 있으면 N = 기타 . 단 링크 + 어문(설명글)이 같이 있으면 어문 . (링크 목록만 나열한 바로가기 페이지 등)

3-2. 검사 대상

L 값	검사 위치
게시판	(1) 목록의 썸네일 <code></code> 존재 여부 → 이미지 (2) 상세글 상위 6~10개 본문에서 이미지·유튜브 (3) 상세글 본문에 유튜브 iframe/링크(<code>youtube.com</code> , <code>youtu.be</code>) → 영상
페이지	본문 영역 안의 <code></code> , <code><iframe src=youtube...></code> , 유튜브 링크
사이트	검사 안 함 (N 빈칸)

★ 3-2a. N 이미지 정밀 검출 (누적 재작업 교훈) - 게시판 N 이미지 = 상세글 표본의 "본문 컨테이너 div" 내부 콘텐츠 img만. 본문영역 셀렉터로 스코프 한정(제주교육청 `div.bdvCntWrap` ·서귀포 `/files|/storage|/board/`). △ 본문 밖 webCheck/추천위젯·QR·고정ID 위젯·썸네일·첨부 PDF뷰어·아이콘은 제외(스코프 안 잡으면 위젯 오탐 127→실제 24, 제주교육청). 첨부가 아닌 본문에 보이는 진짜 사진/갤러리썸네일만. - 페이지 N = HTML `<img>` 휴리스틱만으로 부족 — 지도타일·통계차트·인포그래픽 등 JS그래픽은 `<img>` 가 아니라 누락된다. → Playwright 렌더로 콘텐츠이미지(≥170×110, 노이즈 제외)+지도컨테이너(`.root_daum_roughmap` /iframe map)+PDF임베드 검출 후 몽타주 시각판정(과표시 방지). - 시각판정 기준 재확인(3-1a 보강): 지도·평면도·악보·상징물(시기/시화/캐릭터)·소식지표지·캐릭터 일러스트·실사진=이미지 / 흐름도·조직도·절차도."한눈에"식 인포그래픽·대진표 등 표·이미지=어문. - 과표시 방지: 역방향(이미지 과표시 → 어문) 정정은 사용자 검수영역. 게시판은 상위 표본이라 깊은 글 사진은 누락 가능(표본 한계 명시).

3-3. 노이즈 제외 (이미지 판정에서 빼는 것)

- 공공누리 마크 이미지 자체(`new_img_opentype` , `img_opentype`)
- 사이트 공통 헤더·푸터·로그 이미지 (본문 셀렉터로 한정하면 자동 제외)
- 장식/공통UI: `move.png` (스크롤표시)· `no-img` / `noimg` 플레이스홀더·아이콘·불릿·버튼·배경(`ico_` · `bul_` · `btn_` · `bg_` · `see_btn` 바로보기 버튼)
- 로고류: 사이트 로고·파트너/연계기관 로고(국민신문고 `epeople` · `open.go.kr` ·노동권익센터 등)· `relation_item` ·카피라이트(`copy_logo`)
- 헤드라인 텍스트이미지(`headline` — 글자를 그림으로)·팝업배너(`popup` /popupZone)·웹접근성 마크(`wa_mk`)·트래킹픽셀(`tracer`)
- "한눈에"식 설명 인포그래픽(3-1a) — 한눈에 · 흐름도 · 절차도
- △ "로고송"은 노래(오디오) 이므로 로고 키워드로 제외하지 말 것 (로고(?!송))

3-4. 영상 판별 기준

- `<iframe>` src에 `youtube.com/embed` , `youtu.be` , 또는 일반 비디오 호스팅(예: `vimeo.com`)
- `<a href>` 가 `youtube.com/watch?v=` , `youtu.be/...`

- `<video>` 태그 또는 `.mp4`, `.webm`, `.mov` 파일 링크

3-4b. 오디오 판별 기준 (2026-06-01 신설)

- `<audio>` 태그
 - `.mp3` / `.wav` / `.m4a` / `.ogg` / `.flac` 파일 링크 — 정규식 `\.(?:mp3|wav|m4a|ogg|flac)\b`
 - ⚠ 다운로드 핸들러 쿼리형도 포함: `download.do?fileName=...in04_02_04_A.wav&...` (확장자 뒤 `&` 가 와도 잡히게 `\b` 사용)
 - 예: 당진 "플러그인당진 브랜드송"(로고송 wav 5개) → `어문, 이미지, 오디오`
-

Phase 4 — 공공누리(O · P · Q)

4-1. 유형 추출 패턴

패턴 A — 링크 (교차검증 전용, 유형 판정에는 쓰지 않음) - `<a href="...kog1.or.kr/info/licenseType{n}">`
에서 끝 숫자 `n` (1~4) 추출 → `{n}유형`

패턴 B — 이미지 파일명 (★ 1순위 유형 판정 기준) - 정규식: `(?:new_)?img_open(?:type|code)(\d{1,2})\. (?:png|jpe?g|gif)` — `<img src>` · `style` 배경이미지 · raw HTML 전체에서 검사. 추출 숫자가 1~4일 때만 유효. - `opentype` · `opencode` 두 변형 모두, 확장자 png/jpg/gif, 1~2자리 숫자 모두 허용. - 예:

`img_opentype01.png` → 1유형, `new_img_opentype04.png` → 4유형, `new_img_opencode4.jpg` → 4유형 (서천군)

★ 유형 판정 우선순위 (2026-05-29 확정): 1. 이미지명(패턴 B) 있으면 이미지 우선(권위값). 링크가 존재하면서 이미지≠링크면 4-3b에 따라 비교 링크주소 오기. 일부 사이트는 게시물 템플릿에 링크를 하드코딩해 어긋난다(계룡시: 모든 링크 `licenseType4` 고정). 2. 이미지명으로 판정 불가 + 링크/텍스트 부착만 있으면 → 링크(패턴 A) 유형으로 인정. 배지가 비표준 파일명(`sub040501_img11.png`)이거나 CSS 마크(`<div class="mark--area mark2">공공누리 2유형...`)인 사이트(천안시). 단 링크 유형이 1·2·3·4 전부면 "공공저작물 개방 설명(범례) 페이지"이므로 미부착 처리(증평군 `sub01_07_05.do`). 3. 이미지·링크 둘 다 없으면 미부착.

★ 4-1a. O 판정 함정 (누적 재작업 교훈 — 반드시 준수) - ❌ 본문 "공공누리 제N유형" 텍스트 정규식으로 판정 금지. 본문 안내문·설명문("...제4유형으로 개방합니다")을 마크로 오탐해 대량 false 부착이 난다(진안·제주시 반박). 이미지마크(`img_opentype` / `new_img_opentype`) 또는 `cclBox` `span.ccl_img` 클래스(`open0N`) 만 권위 신호. `cclBox` `<p>` 문구를 굳이 쓰면 "조건에 따라 이용할 수 있습니다"만 진짜, `open05` =자유이용불가=미부착(진안 `[[jinan_kogl_cclbox_open05]]`). - 게시판 상세 추적 시 `boardId/list URL` = 상세 view링크의 최빈값으로 정확히 잡는다. △ 크롤러가 홈아이콘/공통 `boardId`를 dominant 계산에서 잘못 제외하거나(자료실 BBS_0000157 미검출), 페이지 `menuCd`와 안 맞는 `boardId`를 보면 진짜 보드를 놓쳐 미부착 오판(진안 `r190-r205`). `boardId`는 그 페이지 `menuCd`와 일치하는 view링크 최빈으로. - 기관군 예외 — 공공누리 전혀 안 쓰는 곳: 시·도 교육청은 홈·푸터·목록·상세 표본(≤10)을 다 봐도 마크 0 → 전부 미부착(O=미부착, P/Q 빈칸). (시·군은 푸터 선언 혼함과 대조.) 이런 기관군은 마크 0 확인 후 전 행 미부착 일괄. - anti-scraping 서버: 대량 동시요청 시 '마크 든 가짜 페이지'를 섞어 반환하는 사이트(진안) → `rate-limit(sleep)+canary(마크없는 글 주기확인)+calm` 재검증으로 false 4유형 제거. `throttle` 한계로 대형보드는 `calm` 표본만 확정.

4-2. 탐색 전략

1. K URL에 접속 → 본문 영역에서 패턴 A·B 검사
2. L="게시판"인데 본문에서 미발견 → 게시판 목록에서 상세글 URL 상위 10개 추출 → 각 상세글 본문에서 재검사 (건수가 많은 게시판일수록 10개 권장. 2026-06-02 5→10 상향)
3. 두 단계 합산해서 발견된 모든 유형 수집

본문 영역 셀렉터(사이트별 다름, 부록 참조 + 작업 시 확인): - 계룡시 페이지/본문 — #txt , .program-- contents - 사이트별로 셀렉터가 다르면 그 사이트 작업 시 "1-9. 작업 메모용 표"에 기록

4-3. O열 (공공누리 부착) 결과 포맷

상황	O 값
미발견 (자물쇠·게시물 0건 포함)	미부착
단일 유형	1유형 (등)
다중 유형	이미지(패턴 B) 기준 발견된 모든 유형 숫자 오름차순, 콤마만 (공백 없음) 으로 결합 → 1유형,4유형

O열은 이미지명(패턴 B) 우선으로 집계한다. 이미지명으로 판정 불가하고 링크/텍스트 부착만 있으면 링크 유형으로 인정(4-1 우선순위 2). 다중 유형은 실제로 서로 다른 유형이 쓰인 경우에만 나온다.

4-3b. 링크-이미지 불일치 검증 → 비교(S)

이미지(패턴 B) 유형 집합과 링크(패턴 A) 유형 집합을 비교한다.

상황	처리
링크가 존재하면서 이미지 집합 ≠ 링크 집합	S열 비교에 링크주소 오기 기재 (O열은 이미지 기준 그대로 유지)
이미지 있음 + 링크 없음(licenseType 링크 자체가 페이지에 없음)	비교 비워둠. 이미지가 권위값이므로 O열만 확정, 정상 처리
이미지 집합 = 링크 집합	패스 (비교 건드리지 않음)
미부착(이미지·링크 모두 없음)	검증 대상 아님

★ "링크주소 오기"는 링크가 존재하면서 이미지와 다를 때만 기재한다. KOGL 배지 이미지만 있고 kogl.or.kr/info/licenseType 링크가 아예 없는 경우(충북 음성군·청주시 등 다수)는 "오기"가 아니므로 비교를 비워둔다. (2026-05-29 사용자 결정)

예: 이미지= 1유형 인데 링크가 licenseType4 → 불일치 → S = "링크주소 오기". 계룡시 부착 14건 중 12건 이 링크 하드코딩 오류로 불일치 → 비교 기재.

4-4. P열 (마크 부착위치)

발견 위치	P 값
L=페이지의 본문에 마크	게시물 (프리랜서2 #1 — 페이지는 게시판이 아니므로 게시판 이 아니라 게시물 . 진안때 "페이지마킹→P=페이지" 정정됨)
L=게시판의 목록 본문에 마크	게시판
L=게시판의 상세글에 서만 마크	게시물
게시판에서 양쪽 모두 발견	게시판 (본문 우선)
미부착	빈칸

4-5. Q열 (마크 하이퍼링크 활성 여부)

마크가 클릭 가능한 a 태그 안에 들어 있는지 확인:

케이스	Q 값
패턴 A(<code></code>)로 발견	Y
패턴 B(이미지)로 발견했고, 그 <code></code> 의 부모 체인에 <code></code> (의미 있는 링크)가 있음	Y
패턴 B로 발견했지만 <code></code> 가 a 태그 밖에 있거나, 부모 a의 href가 <code># / javascript:</code>	N
미부착 (O=미부착)	빈칸

다중 유형이 발견되어 일부는 Y, 일부는 N인 경우 → **하나라도 Y면 Q=Y**.

4-6. 게시판 상세글 URL 추출

게시판 목록에서 상세글 링크 10개를 모아오는 패턴(사이트별 상이): - 일반: `table.bbs_table tbody td.title a`, `table.tstyle_list td.txt_left a` - href에 `View.do`, `view.do`, `seqRepeat=`, `mode=V`, `?bbsSn=` 같은 query - 상대경로 → 현재 페이지를 base로 `urljoin`

추출 못 하면 → 상세글 검사 스킵, 본문 결과로만 판정.

Phase 5 — 수기 입력 컬럼

자동화 대상 아님. 검토 단계에서 채운다.

열	의미
R	공공누리 연계
S	비고(특이사항) — 자동: Phase 1 "외부링크", Phase 4 "링크주소 오기"(링크 존재 & 이미지≠링크일 때만; 링크 없음은 비움). 나머지는 수기
T	신유형 적용 가부
U	적용 사유
V	검토 대상
W	검토 확정(기관협의결과)
Y	완료여부(Y, N)
Z	확인 날짜
AA	수집 담당자

5-1. 결과영역 Y·Z·AA — 검수완료 시 일괄 작성 ★

Y1:AA1 은 헤더 "결과", 2행 라벨은 **완료여부(Y,N)** / **확인 날짜** / **수집 담당자** . **기관 검수가 끝나(☒ 검수 완료) 데이터가 확정되면** 이 세 열을 **모든 데이터 행에 동일하게** 채운다. (검수완료 5곳 계룡·공주·금산·논산·보령에서 확정된 패턴)

열	값	규칙
Y 완료여부	Y	모든 데이터 행에 Y (빠짐없이 전부). 미완료 행을 따로 N 으로 두지 않음
Z 확인 날짜	검수일	검수완료 처리한 날짜 (datetime, 표시형식 yyyy-mm-dd). 기관 전체를 한 날짜로 통일(예 보령 2026-05-31)
AA 수집 담당자	담당자	광역별 담당자 : 충남·충북= 프리랜서1 , 전북·제주= 프리랜서2 (대시보드 담당자와 동일). ※ 기존 충남 4곳은 plain 프리랜서 로 돼 있으나 신규는 1/2 구분 표기

- **대상 행 판별**: **B** (순번)이 정수인 행만(3행~). 트레일링 빈행·잡행은 제외(공주 사례: 실데이터 400행만 채움, 이하 빈행 미기입).
- **도구**: `작업파일/_fill_YZAA.py <엑셀> <YYYY-MM-DD> <담당자> [--write]` — dry-run으로 행수 확인 후 `--write` (백업 `*_backup_YZAA전.xlsx` 생성, 기존 셀 스타일/테두리 보존, 값·표시형식만 기입).

-   **검수완료**로 이미 잠긴 기관은 재가입 금지(읽기만). 이 작업은 **그 기관을 검수완료로 막 전환하는 시점에 1회** 수행.
-

부록 — 자주 마주치는 변수

사이트맵 셀렉터 후보

1. `.sitemap` (계룡시)
2. `#sitemap`, `div.sitemap_wrap`, `.site_map`
3. `main .content ul` (드물게 직접 ul만 있는 경우)
4. 사이트맵 페이지가 없으면 → 메인 GNB(글로벌 내비) 메뉴 사용

본문(상세 페이지) 셀렉터 후보

- 계룡시: `#txt`, `.program--contents`
- 보건복지부: `#contents_body`, `article.board_view`
- 외교부: 본문 영역 클래스 다양 — 사이트별 확인 필요

URL 절대화 시 주의

- `urljoin` 사용 시 베이스에 슬래시(`/`)를 정확히 붙여야 한다 (`BASE + '/'`)
- `?bbsn=` 같은 query만 다른 상대 URL — 현재 페이지를 베이스로 결합

외부 도메인 화이트리스트(시·군 공통)

정부 공통 서비스. 외부링크이지만 자주 등장하므로 인지만: - `gov.kr` (정부24) - `cheongwon.go.kr` (청원24)
- `wetax.go.kr` (위택스) - `epeople.go.kr` (국민신문고) - `safetyreport.go.kr` (안전신문고) - `car365.go.kr`
(자동차민원대국민포털) - `kogl.or.kr` (공공누리)

변경 이력

- 2026-05-28 — Phase 1 정립 (계룡시 1차 작업 기준). Phase 2~4 자리만 잡음.
- 2026-05-28 — 공주시 Phase 1~4 완료 (306행, 부차 6건). 사이트맵 URL이 onclick encodeURI에 들어가는 패턴 발견, 추출 로직에 반영. 매뉴얼 1-10에 기록.
- 2026-05-28 — 계룡시 Phase 2~4 자동 채우기 완료 (442행, 부차 14건). 셀렉터 1-10에 기록.
- 2026-05-28 — Phase 2~4 정립:
- L: 전체 건수 / 페이징 / 검색 **3요소 중 하나라도 있으면 게시판**
- M: 게시판은 총 게시글 수(숫자만), 페이지는 1, 사이트는 빈칸
- N: 어문 기본 + 이미지/영상 콤마(공백 없음) 결합. 게시판은 상세글 상위 10개까지 검사
- O: `licenseType{n}` 링크 + `new_img_opentype{NN}.png` 이미지 양쪽 검사, 다중 유형은 오름차순 콤마(공백 없음) — 예: `1유형,4유형`
- P: 본문 발견 → `게시판`, 상세글 발견 → `게시물` (양쪽이면 본문 우선)
- Q: 마크가 a 태그(유효 href) 안에 있으면 `Y`, 아니면 `N`
- 2026-05-29 — **Phase 4 O 판정 기준 변경**: O열 유형은 **이미지(`img_opentype{NN}`)만으로 판정**(링크 숫자 합산 폐지). 링크(`licenseType{n}`)는 교차검증용으로만 사용하고, **이미지 숫자 ≠ 링크 숫자면 S열 비교**에 **링크주소 오기 기재**(같으면 패스). 계룡시 게시물 템플릿이 KOGL 링크를 `licenseType4` 로 하드코딩해 실제 배지(이미지)와 어긋나는 오탐을 막기 위함. 4-1-4-3-4-3b 반영.
- 2026-05-29 — **"링크 없음" 처리 규칙 추가**: `licenseType` 링크가 페이지에 아예 없는 경우(이미지만 존재)는 **링크주소 오기** 비교를 **달지 않는다**(비움). "링크주소 오기"는 링크가 **존재하면서** 이미지와 다를 때만 기재. 충청남도 2~15 + 충청북도 1~11 일괄 재검토(`_recheck_kogl_all.py`)에 적용. 4-3b 반영.
- 2026-05-29 — **본문 탭(서브내비) 하위 카테고리 수집 규칙 정립(1-5b 신설)**. 사이트맵에 한 줄로만 나오는 페이지가 본문 탭으로 하위 페이지를 거느리는 경우, 4조건(탭 클래스·실제링크·자기그룹·신규URL) 충족 시 leaf+1 컬럼으로 확장. 인페이지 `#anchor` ·쿼리필터·이미 사이트맵에 있는 상위 nav는 제외. `_tab_expand.py` (확장) + `_tab_phase234.py` (신규행 L~P 수집). 공주시 적용: 91개 그룹 확장, 306행 → 592행(신규 286). 다른 시·군은 보류 후 동일 로직 적용 예정.
- 2026-05-31 — **1-5d 신설(리다이렉트 중복)**: menuCd 등 URL은 다른데 자식으로 자동 리다이렉트되어 화면이 같은 페이지인 셀 행 삭제. 전 행 fetch로 최종 `resp.url` 비교. 고창군 13건(342→329).
- 2026-06-01 — **Phase 3 N 판정 정교화(당진시 사용자 확정)**: ①**3-1a 이미지 판정 핵심 원칙** — 별도 정보를 주는 이미지(사진·평면도·지도·QR·악보·소식지표지·상징물)만 "이미지", **글을 그대로 옮긴 "한눈에"식 설명 인포그래픽은 어문**. ②**3-3 노이즈 제외 확장** — 로고·파트너/연계기관 로고·아이콘·헤드라인 텍스트 이미지·바로보기 버튼·팝업배너·웹접근성마크·트래커 추가('로고송'은 노래라 예외). ③**3-4b 오디오 유형 신설** — `<audio>` · `.mp3/.wav/.m4a` 링크(다운로드 쿼리형 포함) → **오디오** . 도구 `5.당진시/_redo_N.py` (r22~끝 19행 변경). [[feedback-n-image-rule]]
- 2026-06-01 — **작업파일명 광역 접두**: `{기관}.xlsx` → `{광역}_{기관}.xlsx` (예 `충청남도_계룡시.xlsx`) 42개 일괄 rename(`_스크립트/_rename_prefix.py`). 폴더명은 유지. 참조 스크립트 전부 갱신. 리터럴 경로 39파일·64건 치환 + 동적 빌더 5곳

(`_dedup_all.xpath` · `_tab_batch.xlsx_path` · `_jeonbuk_phase234.site` · phase1 3개 · `_fix_emerge`)에
`{prov}_` 접두 삽입. dd.xpath-3개 phase234 SITES 경로 누락 0 검증. (백업 `*_backup_*.xlsx` 는 미개명)

- 2026-06-01 — **수집완료 전 기관 N 일괄 재검토**(검수완료 5곳+당진 제외 **36곳**, `_스크립`
`트/_redo_N_all.py`). 3-1a/3-3/3-4b 규칙으로 약 **1,211행 N 변경**(장식·로고·파트너로고·아이콘·SNS공유버
튼·KOGI마크·"한눈에/흐름도/절차도" 인포그래픽·파일아이콘 → 어문 강등 / 오디오·영상 추가 / **거짓**
M=0 게시판은 내용기반 판정=글 있으면 어문, 진짜 글0개만 없음). 벤더별 fetch 시그니처(충남
`fetch(url)` vs 충북·전북 `fetch(session,url)` +make_session) 자동 흡수. 동시성4+재시도3으로 throttle
거짓'없음' 차단. 기관별 백업 `*_backup_N재검토전.xlsx`.
- 2026-06-11 — **1-5e 신설(랜딩셀 삭제)**: 카테고리 컬럼 $C_n(D \sim I)$ 이 병합돼 있고 그 첫 행의 다음 깊이
 C_{n+1} (최대 J)이 빈 "섹션 랜딩 업데이트"면 첫 행 삭제(라벨은 첫 자식 승계). 1-5(URL 동일)·1-5d(리다이렉트)
와 달리 URL이 자식과 다르고 리다이렉트도 안 되는 독립 섹션 인덱스. 보존=게시판/공공누리 부착·단일
행 리프(병합 아님). 도구 `_del_shell.py` (전 깊이) · `_collapse_landing.py --safe`. 기적용: 전북 일괄(98)·김
제-2·남원-1·전주-27·제주시34·서귀포36·완주·정읍23(G/H). 이전부터 사이트별로 반복 요구돼온 규칙을
정식 등재.
- 2026-06-12 — ★ 프리랜서2(발주처) 공식 검수 지침 17항 반영(`[[feedback_pro2_review_guidelines]]`),
기존 관행 덮어쓰는 정정: ① **1-0 권위트리=사이트맵 아니라 본 페이지 메뉴(GNB/LNB)**(누락 多,
#3·#16) · **3-1 N유형 8종**(어문/이미지/영상/오디오/글꼴/3D/기타/없음, #6) · **3-1a 이미지 세분**(포스터·카
드뉴스=이미지·글자만 이미지=어문·그래프만=기타·그래프+글자=어문, #4 — 기존 "그래프=이미지" 정
정) · **3-1b 링크전용 페이지=기타**(링크+어문=어문, #2) · **2-3 로그인/본인인증 페이지=행 제거**(#15, 기존
"L=페이지 유지" 정정) · **2-1b 게시글 안 보여도 게시판형태=게시판**(#5)·**게시판은 1행, 게시물 행 추가 금**
지(#10)·**한 페이지 다중게시판 전부**(#13) · **2-1a 다른창=사이트+분야별정보 예외**(#7) · **4-4 P열 페이지마**
크=게시물(#1, 진안 "P=페이지" 정정). #8 아이콘제외·#11 사이트뒤공백·#14 게시물수·#17 게시물0=없음
은 기존 반영 재확인.
- 2026-06-12 — **누적 재작업 기반 정밀화 5종 등재**(시·군 30+곳·교육청 3곳 경험): ① **1-5c 탭 정밀화** —
`[class*=tab]` 가 `taps` / `tap` (b없는) 위젯 누락(`tab` + `tap` 둘 다)·인페이지#앵커 콘텐츠탭(페이지 M=탭
수)·카테고리탭(`categoryCode` 각 별행 게시판)·탭 u의 선행 일반메뉴(랜딩) 누락 주의·앵커는 URL로 배치·
중첩탭 fixpoint. ② **2-1a 사이트 판별** — `_blank` =무조건 사이트·최종도메인/서브사이트 기준·교육청 로고
를. ③ **2-2a M 합정** — 번호=최신글번호·페이지수 저장버그(아산)·edufine 가짜페이지(동일 거대값=통합
뷰)·빈게시판 M0. ④ **3-2a N 검출** — 게시판=본문 div 스코프 img만(위젯·첨부 제외)·페이지=렌더+몽타주
시각판정(JS그래픽 누락 방지)·표이미지/조직도=어문. ⑤ **4-1a O 합정** — 본문 "제N유형" 텍스트정규식 금
지(이미지마크/ccl_img만)·boardId 상세 view최빈·교육청 등 기관군 전부 미부착·anti-scraping rate-
limit+canary. + **1-7b 사용자 부분검수 보존**·수정/신규셀 하늘색(fg+bg)·국소삽입(generic remap).
SKILL.md 핵심규칙·절대규칙·합정에도 요약 반영.