역사 큐레이션형 키워드 뉴스 포맷을 실현하려면, 단순 뉴스 크롤링을 넘어 키워드와 관련된 다양한 콘텐츠를 장기적으로 축적하고 구조화할 수 있는 수집·정제 프레임워크가 필요합니다. 아래에 퀄리뉴스를 리팩터링하여 "편집장이 하루 한 키워드"를 중심으로 운영하는 시스템을 구축하기 위한 기술적 제안을 정리했습니다.

---

1. 핵심 개념: "키워드 단위 역사 데이터베이스" 구축

1. **장기적 관점**: 표준·기술의 역사는 수 년 또는 수십 년에 걸친 데이터가 필요합니다. 하루 수집으로 끝나는 것이 아니라, 키워드별로 과거 자료를 꾸준히 쌓아가는 데이터베이스를 설계합니다.
2. **다채로운 콘텐츠 유형**: 뉴스 기사뿐 아니라 학술 논문, 표준/정책 문서, 기업 블로그, 커뮤니티 토론, 제품 리뷰 등 다양한 유형의 자료를 수집해야 합니다. 각 유형마다 수집 방법과 메타데이터가 다르므로 수집 파이프라인을 분기합니다.
3. **키워드 중심의 색인 구조**: 모든 자료는 키워드를 중심으로 태그를 붙여 데이터베이스에 저장하고, 시간순·유형별·출처별로 조회할 수 있어야 합니다.

---

2. 수집 프레임워크 설계

2.1 콘텐츠 유형별 수집 모듈

1. 뉴스/블로그 모듈

- **공식 뉴스**: 앞서 제시한 산업 매체(RSS/HTML) 크롤러를 사용해 표준/전자제조/AI 관련 뉴스와 블로그 글을 수집합니다. 제목과 본문에서 키워드가 포함된 경우 해당 키워드에 연결합니다.
- **한계**: 뉴스 사이트에서는 과거 기사 검색이 어려울 수 있으므로, Google/Bing News API나 Wayback Machine API를 활용해 과거 글을 보강하는 방법을 고려합니다.

학술 논문 모듈

- **소스**: arXiv, IEEE Xplore, Google Scholar, Semantic Scholar 등 키워드 검색이 가능한 API 또는 RSS를 활용합니다. 예를 들어 ArXiv는 키워드 검색을 지원하며, 논문 메타데이터와 추상(Abstract)을 가져올 수 있습니다.
- **키워드 매칭**: 논문 제목, 키워드, 초록에 키워드가 포함되어 있으면 수집 대상에 포함합니다. 학술 데이터는 DOI, 저자, 발표 연도 등의 메타정보를 함께 저장합니다.

표준/정책 문서 모듈

- **소스**: IPC, ISO, IEC, JEDEC 등의 표준 기관 웹사이트와 "국가법령정보센터" 등 정책 문서 저장소에서 키워드 검색을 수행합니다. 표준 문서는 보통 PDF로 제공되므로, 제목/요약/출처 링크를 저장하고, PDF 전문은 서버에 캐시하거나 링크만 저장합니다.
- **수집 방식**: RSS가 없으므로, 키워드 검색 페이지를 스크래핑해야 합니다. "IPC-J-STD-001" 등 표준 번호를 키워드로 등록하고, 연도별 수정 내역을 수집합니다.

커뮤니티/포럼 모듈

- **소스**: Reddit 서브레딧, EEVBlog, Stack Exchange (electronics), Hacker News 등. 키워드 기반 검색 혹은 태그 기반으로 과거 토론을 검색합니다.
- **비정형 데이터**: 커뮤니티 게시글은 잡음이 많으므로 추천 수, 댓글 수, 조회수 같은 지표로 신뢰도를 보정하고, 키워드가 포함된 문장 수에 따라 가중치를 부여합니다.

제품/리뷰 모듈 (선택)

- 기술 제품 출시나 리뷰 정보도 키워드 역사에 포함될 수 있습니다. Amazon, DigiKey, 제조사 블로그 등에서 키워드 기반 제품 정보를 크롤링하되, 저작권과 서비스 약관을 준수해야 합니다.

2.2 데이터 저장 및 색인

- **데이터베이스 설계**: 관계형 DB 또는 Elasticsearch처럼 검색 최적화된 시스템을 이용합니다. 테이블/인덱스 구조 예시는 다음과 같습니다.
  - keywords: 키워드 목록, 설명, 최초 등록일 등.
  - documents: 문서 ID, 키워드 ID, 출처 유형(뉴스/논문/표준/커뮤니티 등), 제목, 요약, 링크, 발행일, 수집일, 메타데이터(JSON).
  - citations(선택): 논문과 표준 문서 간 상호 참조나 관련 링크.

시간 정보: 수집일과 발행일(작성일)을 분리하여 저장함으로써, 발행 시점 기준으로 히스토리를 제공할 수 있습니다.

2.3 메타데이터 및 스코어링

- **품질 및 신뢰도**: 기존 QG/FC 기준을 유형에 맞게 조정합니다. 예를 들어 표준 문서는 본문 길이가 짧아도 높은 신뢰도로 간주하고, 커뮤니티 글은 본문 길이·추천 수·출처 등으로 신뢰도를 평가합니다.
- **가치 평가**: 키워드 뉴스에서 중요한 것은 다양성과 깊이입니다. 같은 키워드라도 공식 뉴스, 표준 문서, 학술 논문, 커뮤니티 토론 등 서로 다른 유형을 균형 있게 수집하도록 점수화합니다.
- **중복 감지**: 동일 링크나 거의 유사한 제목의 자료는 중복으로 판별하여 하나만 저장합니다.

---

3. 선정 및 발행 워크플로우

1. **자동 수집**: 키워드가 입력되면 N개의 모듈이 병렬로 데이터를 수집합니다. 수집 주기는 키워드 뉴스 요청 시 또는 배치 작업으로 지정할 수 있습니다.
2. **자동 분류/요약**:
   - 유형별 분류는 수집 시점에 결정됩니다.
   - 영문 자료는 필터링 후 중요 문장 추출과 GPT 기반 요약을 수행합니다 【1002_2퀄리뉴스의 기술적목표.txt†L39-L47】.

편집자 대시보드:

- 편집자는 수집된 자료를 메타데이터와 함께 검토하며, 포함 여부를 선택합니다.
- 자료 유형별 필터, 날짜 필터, 신뢰도 필터 등을 제공해 효율적으로 검토할 수 있게 합니다.

발행 형식:

- 키워드별 페이지를 생성하여, 연대기 순/유형별/관련성 순 등 다양한 섹션을 제공합니다.
- 각 항목에는 제목, 발행일, 간단 요약, 출처, 신뢰도 지표를 표시합니다.
- 독자는 키워드의 역사 흐름과 최근 동향을 한눈에 볼 수 있게 됩니다.

---

4. 기술적 뒷받침 제안

1. **모듈화 및 확장성**: 앞서 제안한 모듈 구조를 기반으로, 새로운 유형의 자료가 필요할 때 모듈을 추가하는 방식으로 확장합니다.
2. **검색 API와 법적 고려**: 논문·표준·정책 문서의 경우 크롤링 허용 범위를 확인하고, 각 서비스의 API 약관을 준수합니다.
3. **검색 및 분석 도구**: Elasticsearch(또는 OpenSearch)를 활용해 키워드 검색과 시간순 정렬을 빠르게 수행하도록 합니다.
4. **데이터 파이프라인 자동화**: Airflow나 Prefect 같은 워크플로우 관리 도구로 수집→정제→저장 작업을 자동화하면 다양한 키워드의 지속적인 업데이트를 관리할 수 있습니다.
5. **학습 데이터 구축**: 장기적으로는 수집된 역사 데이터를 AI 모델 학습에 활용해, 키워드와 관련된 트렌드를 예측하거나, 키워드 추천 기능을 제공하는 것도 가능합니다.

5. 결론

“편집장이 하루 한 키워드”를 중심으로 하는 역사 큐레이션형 퀄리뉴스를 구축하려면, 키워드를 단위로 다양한 출처에서 오랜 기간의 데이터를 체계적으로 수집해 저장하고, 유형별·시간별로 분류·스코어링하는 데이터 파이프라인이 필요합니다. 기존 퀄리뉴스의 구조를 모듈화하여 뉴스·논문·표준·커뮤니티 등 여러 모듈을 추가하고, 편집자가 이를 선별할 수 있는 UI와 자동화된 발행 시스템을 만들면, 한 키워드에 대한 과거와 현재, 다양한 관점의 정보를 독자에게 전달하는 새로운 뉴스 포맷을 실현할 수 있습니다.