한인 디아스포라 디지털 아카이브 연구

Korean Diaspora Digital Archive System

1. Korean Diaspora Digital Archive System (KoDDAS)의 개요

한인 디아스포라 디지털 아카이브(이하 KoDDAS)의 목적은 한국인 디아스포라 데이터를 조직 및 종합하여 관련 연구 활용 뿐 만 아니라 일반 대중들에게도 보급이 가능한 디지털 데이터 관리 인프라를 제공하는 것이다. 이러한 측면에서 KoDDAS 는 기존의 데이터 수집 및 보존 등의 기능을 확장하여, 정보의 보급 및 데이터 재사용에 중점을 기한다.

KoDDAS 는 우선 데이터 재사용의 촉진을 위하여, 미가공 데이터(즉, 인터뷰 대본, 사진, 동영상 등) 뿐만 아니라 역사적 혹은 지역적 상황에 적합한 정보와 객체관계를 활용한 구조적 메타데이터를 제공하는 통합 데이터베이스 시스템을 구현하여 한국인 디아스포라 데이터의 이해와 분석을 도모 할 것이다. 또한 정보 보급을 위하여, KoDDAS 데이터베이스에서 한국인 디아스포라 데이터의 탐색과 검색 및 저장이 가능한 한국인 디아스포라 디지털 아카이브 센터(이하 KoDDAC) 웹사이트를 구축 할 것이다.

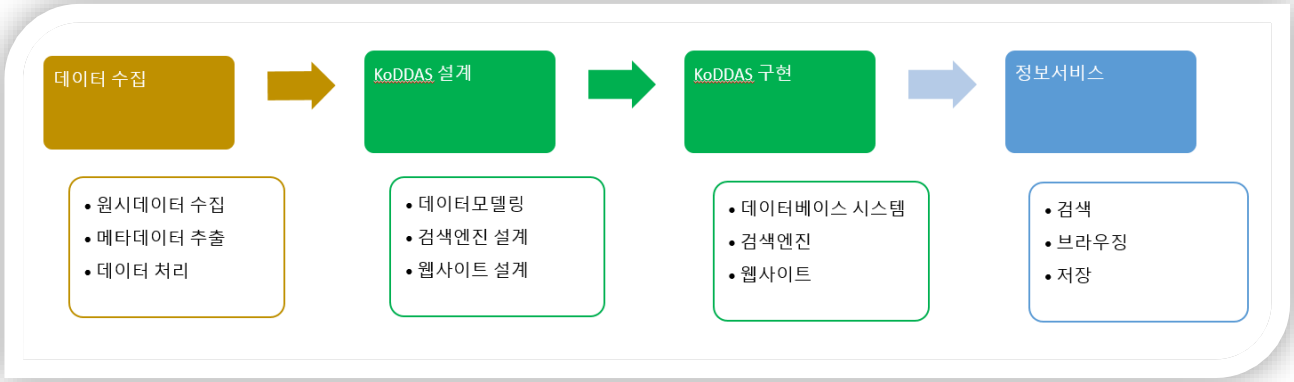


그림 1. KoDDAS 개요

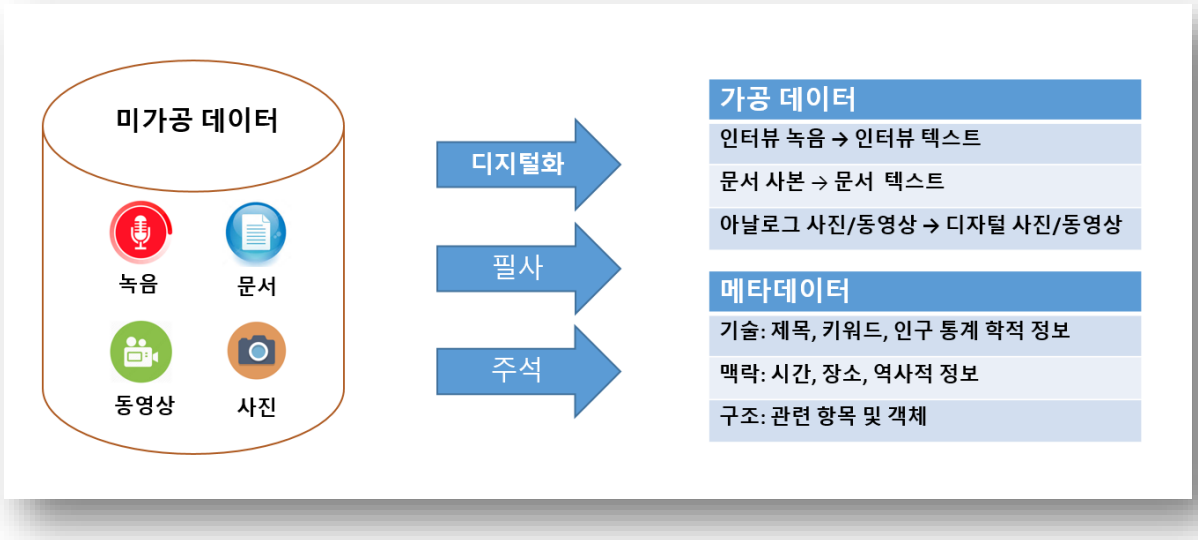
그림 1 은 KoDDAS 의 구현 개요로, 데이터의 수집 과 큐레이션 절차를 거쳐 한국 디아스포라 디지털 아카이브에서 웹 서비스로 제공 할 데이터베이스 시스템과 검색엔진의 설계 및 구현 까지의 절차를 보여준다. 첫번째 단계인 데이터 수집 에서는 디지털화, 데이터 보관과 키워드 및 메타데이터 색인 작업등을 수반한 데이터 처리 과정이 포함된다. 두번째 단계인 데이터 큐레이션은 인공지능(AI)도구를 활용한 "확장지능과정(augmented intelligence process)"를 포함한다. 이 과정은 KoDDAS 데이터베이스의 설계와 검색엔진 구현을 용이하게 하고, 데이터 마이닝과 정보 조직 결과의 향상을 도모할 것이다. 또한 세 번째 단계는 KoDDAS

데이터베이스와 검색엔진 및 웹사이트의 구현을 수반하며, 이 단계를 통하여 통합한 아카이브의 데이터 탐색과 검색 및 저장 등이 가능한 웹 기반 서비스가 제공될 것이다.

이렇게 구축된 KoDDAS 는 AI 기반의 큐레이션을 통하여 비정형 미가공데이터를 다각적 아카이브 데이터베이스 시스템으로 변환하는 로드맵을 제공함으로써, 한국 디지털 아카이브 센터의 효율적이고 효과적인 운영을 가능하게 할 뿐만 아니라, 타 디아스포라 디지털 아카이브 시스템 구축을 위한 기반이 될 것이다. 따라서 KoDDAS 는 현재의 디아스포라 연구자들이 필요한 귀중한 서비스를 제공함과 동시에 일반 대중들이 디아스포라 데이터에 보다 쉽게 접근할 수 있도록 하며, 또한 학술적 분석을 위한 문화 컨텐츠의 체계화 및 보존에 최첨단의 정보통신 기술을 적용한 사례로서 활용 될 수 있을 것이다.

2. 데이터 수집

그림 2 는 데이터 수집 과정을 도식화 한것으로, 초기 KoDDAS 데이터는 인터뷰 녹음, 사건의 녹화 동영상, 인터뷰 대상자가 제공한 사진이나 문서의 사본등으로 구성 되며, 오디오 녹음의



경우 디지털 인터뷰 문서로 변환 , 이 후 각 컨텐츠 요소에 따라 메타 데이터가 할당 된다. 정확한 메타데이터를 구축하려면 미가공 데이터의 심층적 분석과 KoDDAS 목표의 세심한 검토가 요구되나, 메타 데이터에는 데이터 항목을 설명하는 설명 정보, 위치 및 과거 배경, 구조 정보 등의 상황 정보가 포함된다.

그림 2. 데이터 수집 과정

3. 데이터 큐레이션

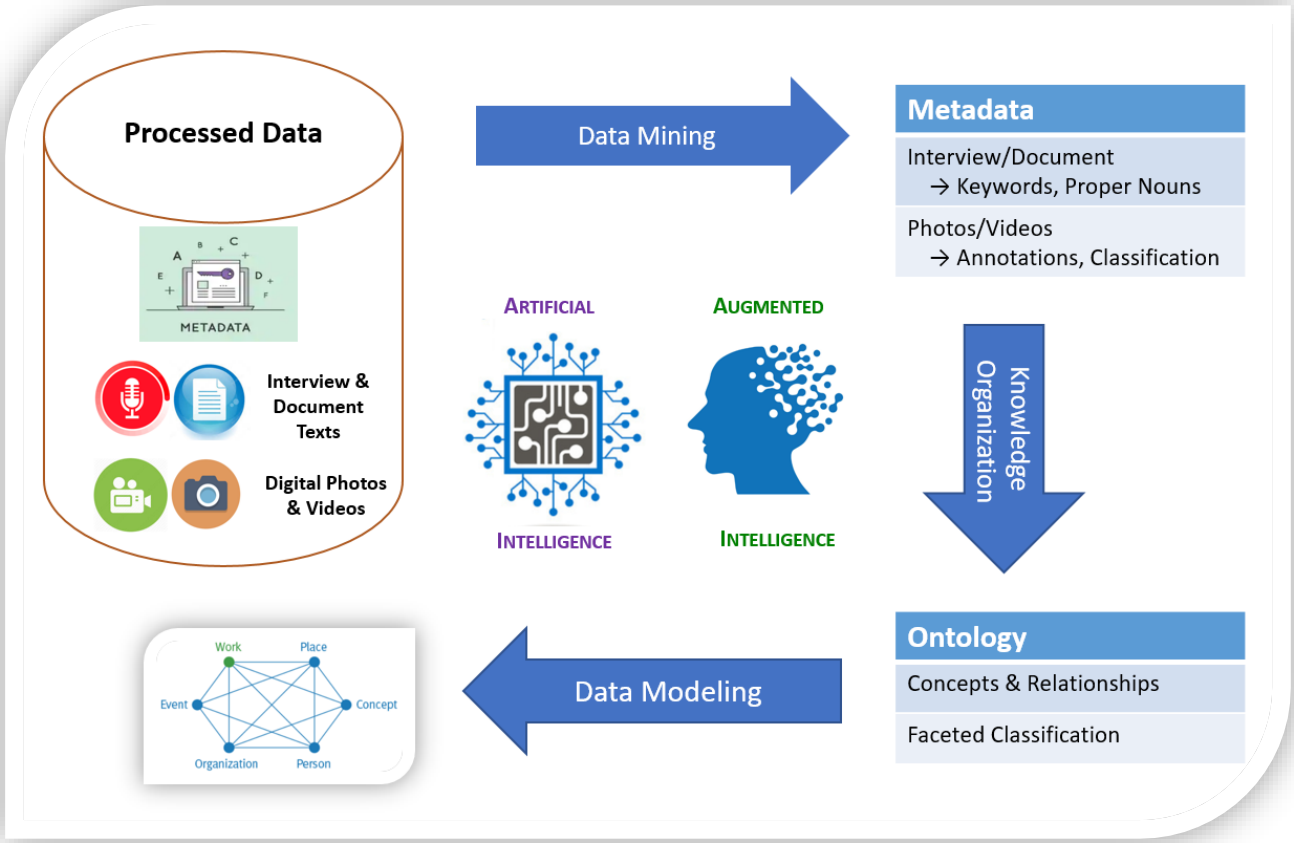


그림 3. 디지털 큐레이션 과정

데이터 큐레이션에서는 먼저 인터뷰 및 문서 텍스트에서 키워드와 고유명사를 추출하고, 래피드 마이너(Rapid Miner) 혹은 텐서플로(Tenor Flow) 등의 AI 기반 데이터 마이닝과 기계학습도구를 활용해 사진과 동영상자료 분류와 주석기입을 시행한다.

이과정을 통하여 생성된 메타데이터는 전문가 집단에 의해 활용성의 유무가 식별 되며, 이는 본 시스템의 패싯분류의 기반이 될 것이다. 또한 수집된 모든 텍스트 데이터와 메타데이터는개념-관계 지식맵의 구축을 위하여 수동 검증을 거친 다음, 준지도학습(semi-supervised learning)의 반복주기에 투입될것이다.

따라서 본 데이터 큐레이션 단계는 KoDDAS 데이터모델의 구축을 용이하게 하게 하고, 지식 생성과 검색을 가능케 하는 온톨로지와 시스템의 검색 능력을 향상 시키는 메타데이터를 구축한다.

또한 이 단계는 인공지능(AI)이 지식을 발견하고 조직하는 인간 고유의 "지식 조직 영역"의 결과를 향상시키고 촉진하는 인텔리전스 어플리케이션의 실례가 될 것이다.

4. KoDDAS 의 설계 및 구현

그림 4 은 KoDDAS 시스템 구조모형으로, KoDDAS 는 클라이언트-서버 아키텍처로 구현되며, 인터넷을 활용한 데이터 관리 시스템(DBMS), 검색엔진, 웹서버와 데이터베이스 응용 프로그램으로 구성, 인터넷을 통하여 데이터 분석 및 탐색과 검색 서비스를 클라이언트에게 제공할 것이다.

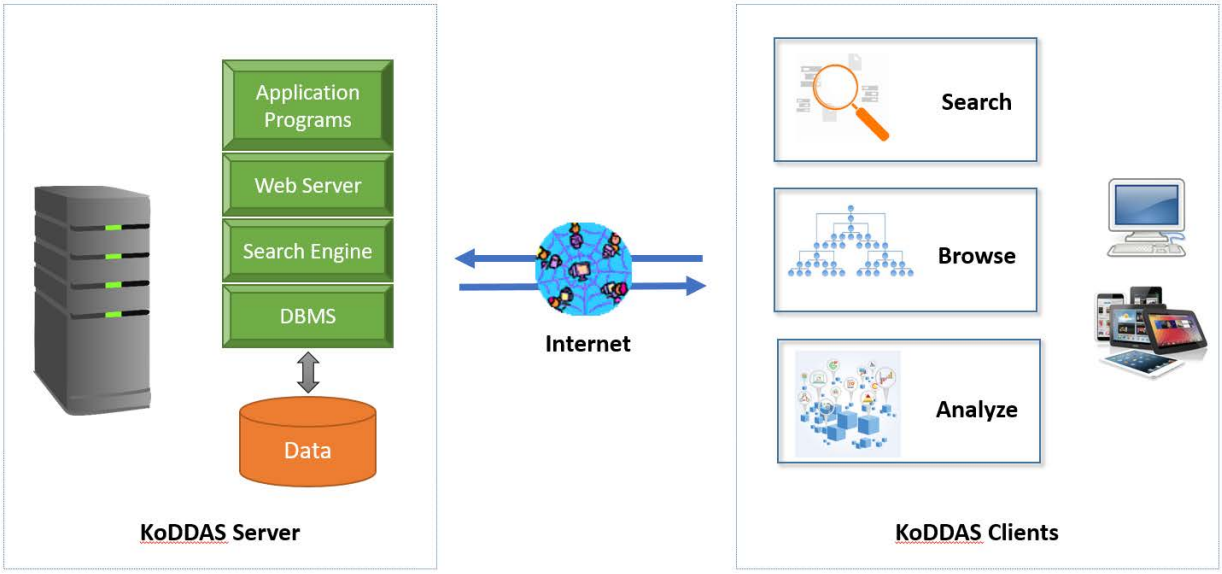


그림 4. KoDDAS 의 서버-클라이언트 아키텍처

또한 그림 5 에 제시된 KoDDAS 데이터 모델의 초안을 MySQL DBMS 로 구현하여 미가공 및 메타데이터를 구조화된 형식으로 캡슐화를 진행, 데이터 탐색과 검색 및 저장과 분석을 용이하게 한다. 특히 본 연구의 DBMS 는 기존 관계형데이터 모델(구조적 및 문맥적 메타데이터 관리)뿐만 아니라 패싯 분류 데이터와 개념-관계의 지식맵을 RDF(Resource Description Framework)포맷으로 제공한다. KoDDAS 에서 제공하는 검색엔진은 데이터 베이스 메타데이터 검색, 디지털 문서 어간 검색(語間, Free-text Search), 그리고 이 둘을 결합한 통합검색으로 구성된다. 데이터베이스 검색은 DBMS 에 질의를 위한 응용프로그램 환경에서 실행되는 반면, 어간검색(語間, free-text search) 검색은 전통적인 벡터 공간 모델 프레임 워크에서 인터뷰 내용을 색인화 하고 검색한다. 또한 데이터베이스 검색 외에도, 지역별 데이터 통계 표시와 같은 KoDDAS 웹 사이트의 상호 작용 구성 요소뿐 만 아니라 LOD(Liked Open Data)포맷의 구조적 패싯분류와 지식맵 KoDDAS 데이터 검색 인터페이스를 제공하기 위하여 다른 응용 프로그램(예: PHP/MySQL, Node.js)도 구축될 것이다.

KoDDAS 의 서버는 64 기가바이트의 램과 9 테라바이트의 하드디스크 공간을 갖춘 10-core 리눅스 워크스테이션에 설치될 것이며, 스마트폰이나 태블릿등의 클라이언트의 개인용 장치에서 IE 혹은 크롬과 같은 웹브라우저 상의 HTML/PHP 인터페이스를 제공하기 위하여, MySQL(DBMS)와 및 Apache(웹서버)를 실행할 것이다.

또한 KoDDAS 는 관계형데이터베이스의 형태로 중앙서버에 데이터를 집계함으로써, 한국인 디아스포라 정보의 활용과 보존을 보장하며, 문화적 혹은 사회경제적 연구와 같은 메타데이터가 풍부하면서도 구조화되고 고품질의 정보를 필요로 하는 디아스포라 연구자들에게 포괄적인 정보원이 될 것이다.

한편, 특정 기록, 역사적 사실, 인물의 내용을 수록한 인터뷰 나 문서에서의 어간검색이 가능하도록 설계된 검색엔진은 데이터베이스의 정형적 검색을 보완하여 시간이나 장소, 시간별 추이, 인물 등의 패싯에 따라 검색결과를 도출 가능케 한다. 또한 KoDDAC 웹사이트에서는 일반 대중들에게 인구별 분포, 이용자 정의에 따른 기술적 통계 와 같은 전반적인 통계 데이터와 개요 뿐만 아니라 LOD 포맷의 지식 맵과 패싯 브라우징이 가능한 한국인 디아스포라 데이터 인터페이스를 제공할 것이다.

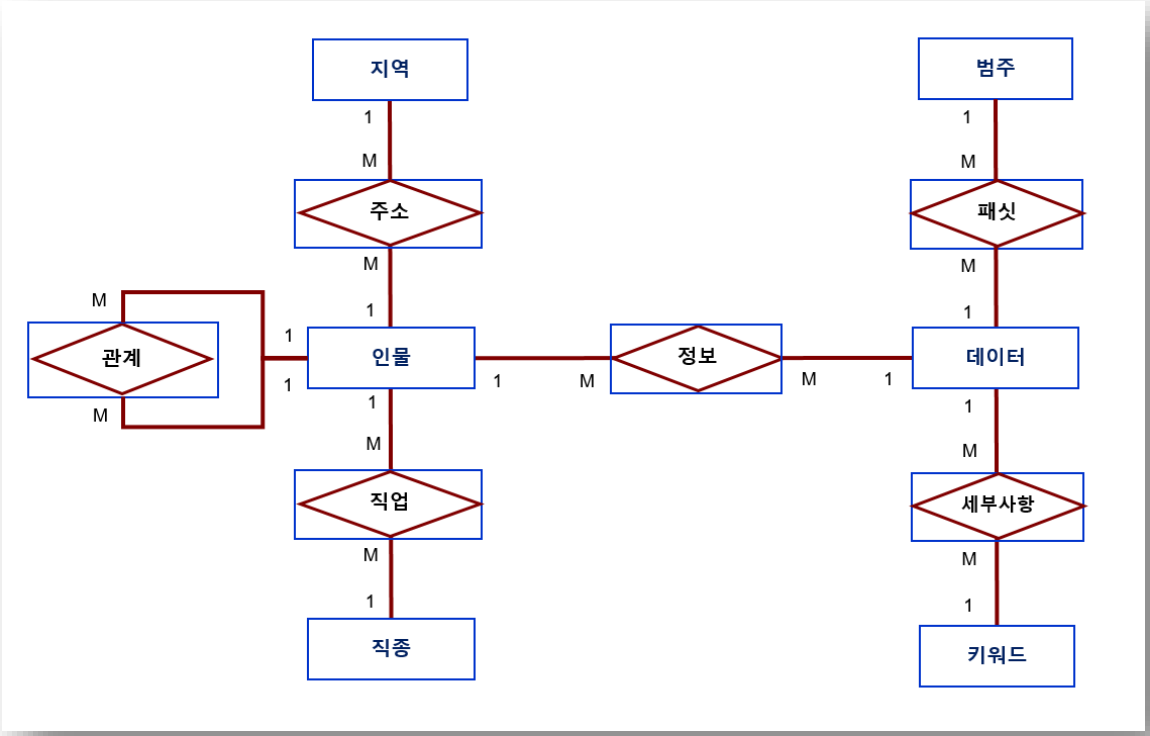


그림 5. KoDDAS 의 초기 데이터모델

그림 5 는 KoDDAS 의 초기 데이터모델로 , 본 데이터 모델의 핵심 개체는 “PERSON”과 “DATA” 이다. 이중 핵심 개체 “PERSON”의 연결개체인 “JOB” 과 “ADDRESS”은 시간경과에 따른 각 대상의 직업과 주소 정보를 수록하고, “RELATION”은 대상의 가족 및 사회 구성원 간의 관계를 수록한다. 또한 “DATA” 는 수년간 수집 한 사진, 비디오, 인터뷰 원고 등의 핵심 데이터의 저장을 위한 개체로서, “FACET” 개체를 통해 패싯 방식으로 범주를 분류하고, “DETAIL” 개체를 통하여 통제어를 할당한다. 그리고 “PERSON”과 “DATA” 개체를 연결하는 “INFO”개체는 누가 · 언제 · 어떠한 데이터를 수집하였는지를 기록한다. 또한 제안된 본 데이터 모델은 키워드 및 카테고리 구성 뿐만 아니라 쿼리를 활용하여 데이터의 분석 · 검색 · 현황 · 통계 등도 가능케 한다.

이처럼 본 연구의 초기 데이터 모델은 한국인 디아스포라 데이터의 핵심 요소(예 : 데이터, 인물, 직업, 장소)와 이의 메타데이터(예: 카테고리, 키워드) 그리고 데이터간의 관계 (예: 관계, 정보, 패싯)를 나타내고 있으며, 구상화된 KoDDAS 데이터베이스의 기본적 토대로써 캡슐화 한다.

적절한 데이터모델이 존재하지 않는 데이터 아카이브는 데이터의 통계 및 재활용등을 저해하는 무질서의 데이터 저장소의 가치로만 인식될 뿐이며, 이는 추후 효율적인 한국인 디아스포라 데이터 접근과 유의미한 통계분석 등에 곤란하기에, 본연구에서 제안하는 것과 같이 체계적이고 완성도 있는 데이터모델의 구축이 절실 하다.

이후 본 연구는 위와 같은 KoDDAS 데이터 모델의 프로토타입을 초기 구축한 다음, KoDDAS 데이터베이스의 각 구성 요소의 검증화 과정을 거쳐 능률성을 증가 시킬 것 이며, PHP/HTML 및 기본 SQL 쿼리를 이용하여 검색 및 분석이 가능한 인터페이스 구현 등, KoDDAS 의 점진적 정교화를 위하여 프로토타입의 세분화를 다회 반복 시행할 것이다. 더불어 KoDDAS 는 XML 등과 같은 표준 포맷에서 내·외부데이터를 입력하거나 백업 및 배포 등이 설계된 메카니즘을 기반으로 하는 데이터 입·출력 및 업데이트 인터페이스 역시 포함 할 것이다.

5. KoDDAS 서비스

KoDDAS 의 주요 서비스(그림 6 참조) 는 데이터 보존서비스, 데이터 보급서비스, 정보검색 그리고 데이터 분석 서비스이다.

데이터 보존 서비스는 데이터 집계 및 구성 보존을 포함하며, 물리적 형태의 원본 데이터와 DBMS 와 같은 응용프로그램의 디지털화 된 미가공데이터와 메타데이터와 한국 디아스포라 아카이브 센터의 유지 관리를 위한 보존형식(eg XML,PDF)도 저장한다.

데이터 보급 서비스는 사용자가 요구 시에 오픈 링크드 형식으로 데이터 저장이 가능하도록 시간, 지역, 키워드등의 카테고리별로 분류한 KoDDAS 데이터들의 검색을 제공한다.

또한 데이터베이스 검색과 자유텍스트 검색이 구성된 정보검색 서비스는 인터뷰 내용 뿐 아니라 KoDDAS 에서의 필드별 검색도 가능할 것이다.

마지막으로 데이터 분석 서비스는 기존 KoDDAS 데이터베이스가 계산한 전반적인 기술 통계와 이용자들의 요구에 의한 임의 쿼리에서 컴파일된 맞춤형 통계를 제공한다.

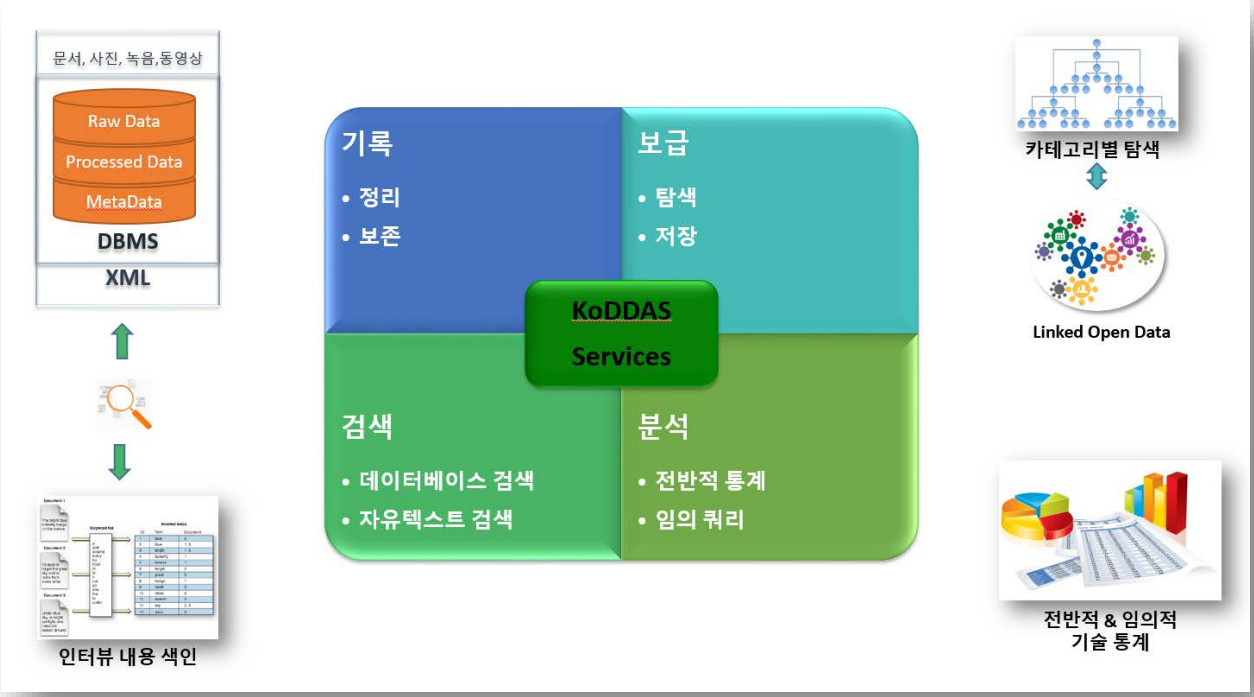


그림 6. KoDDAS 의 주요 서비스

이와 같은 KoDDAS 의 구축은 한국인 디아스포라 데이터를 집계, 조직, 보존 하고 온라인 데이터를 제공함으로써, 데이터 분석을 용이하게 하고, 데이터 재사용을 촉진시키고자 하는 한국 디아스포라 디지털 아카이브 센터의 성공적인 운영에 기여 할 것이다.

KoDDAS 의 중요성은 바로 기존 아카이브센터들의 서비스를 훨씬 능가하여 다양한 서비스를 가능케 하는 “정보 기술 어플리케이션의 혁신적인 통합”에 있다. 이러한 관점에서 제안된 이 연구는 다양한 사회과학 주제분야의 전문지식을 수집하여 이용자에게 광범위한 서비스를 제공하는 다중 분야의 사용자 중심 연구의 최첨단 비전을 구현 할 것이다.