2018 한국인터넷거버넌스포럼(KrIGF) 워크숍 보고서

작성자 : 박다은(한성대학교)

세션명	인공지능과 오픈 데이터 (부제: 인공지능과 데이터는 만나야 합니다.)			
일시	2018.7.5. (목), 11:00~12:30		장소	서울창업허브 IR 미디어룸(2F)
참석자	사회	오원석(사단법인 코드, 이사)	발제	
	패널	이명진 ((주)리스트, 기술이사)		
		엄재홍(SKT, 매니저)		
플로어	약 50명 참여			

제안내용	CES2018에서도 가트너의 발표에서도 2018년에 그리고 앞으로의 미래에 가장 주목해야 할 기술이 인공지능이라는 것을 주저하지 않는다. 자율주행자동차에도 인공지능비서에도 그리고 IoT나 스마트도시에서도 인공지능이 차지하는 역할은 그 어떤 기술보다 높다고 할 수 있다. 그런 인공지능 기술을 튜닝해 효율적으로 성과를 내기 위해서 무엇보다 중요한 것이 데이터라는 것을 누구도 부인하지 못할 것이다. 최근 중국이 무섭게 인공지능 기술에 투자하고 있다. 바이두는 10만 양병설을 주장하며 인공지능 인해전술 전략을 내어 놓았으며, 센스타임은 세계 최고의 얼굴인식 기술을 선보이며 급부상하고 있다. 그 이면에는 중국 정부의 공공데이터 제공과 활용이 있었다고 한다. 우리도 그런 세계적인 추세에 뒤처지지 않기 위해서는 지금부터라도 준비해야 한다. 우리에게는 한글이라는 특수성이 있다. 인공지능 기술 지원을 위해 더 많은 공공데이터가 제공되고 가공 과정을 거쳐 더 많이 선순환되어야 세계적인 추세에 대응할 수 있다.
요약내용	인공지능 기술이 날로 발전해 가고 있다. 서비스를 구상하고 인공지능 알고리즘의 완성도를 높이기 위해서는 다량의 데이터를 통한 학습이 필요하다. 하지만 대부분의 개발자들이 필요한 기반 데이터를 얻는 데서부터 어려움에 직면한다. 양질의 데이터의 경우 구글, 아마존과 같은 대형 기술 플랫폼들이 자사 서비스를 기반으로 데이터를 얻고 있기에, 접근하기 어렵고 공개된 데이터의 경우, 가공되지 않거나 기계가 읽을 수 없는 형식으로 제공되어 데이터 정제에 시간과 수고가 많이 드는 어려움이 있다. 따라서 정제된 데이터의 경우 자산이라는 인식이 강하게 작용하여 공유가 소극적으로 이루어지고 있다. 이러한 상황을 타개하고자 국내의 경우, 정부차원에서 공공데이터 포털을 통해 공공기관에서 생산된 데이터를 제공하고 활용을 독려하고 있다. 국외에서도 다양한 오픈소스 데이터셋을 통해 알고리즘을 강화하기위한 노력이 이루어지고 있다. 이러한 움직임은 중소기업의 경쟁력을 올리고 공정한 AI 시장에서의 경쟁체계를 구축하는데 도움을 준다. 향후 깃허브와 같은 코드 공유 커뮤니티처럼 데이터 차원에서도 공유 생태계가 조성되어야 할 것이다.

논의 세부 내용

Ⅰ. 인공지능과 오픈데이터

1. 오픈데이터와 인공지능 (이명진 발제자)

(1) 오픈데이터와 공공데이터

- 오픈데이터: 저작권, 특허 같은 제약 사항 없이 모든 사람이 자유롭게 사용할 수 있는 개방된 데이터
- 공공데이터: 전자화된 파일 등 공공기관이 법령 등에서 정하는 목적을 위해 생성 또는 취득하여 전자적 방식으로 처리된 자료 또는 정보. open data와 government data의 교집합인 open government data.

(2) 오픈데이터의 중요성

- 대표사례: 서울버스 앱
- 버스 관련 정보 제공 웹에서 HTML 스크랩핑, 표출 - 사용자가 늘어나니 서버 부하, 경기도 버스 서비스에서 앱 차단 - 시민들의 항의로 오픈 API로 개방 - 서버를 운영해야 할 필요성 - 비용 필요 - 광고 게재 - 공공데이터의 상업적 이용에 대한 비판 - 네이버 운영 지원 - 카카오인수
- <공공데이터의 제공 및 이용 활성화에 관한 법률>(이하, 공공데이터법)의 제정으로 상업적으로도 제한없이 데이터 사용 가능
- 공공데이터를 이용해 다양한 서비스가 시민들에게 제공되고 있으며, 우리 삶에 미치는 영향력이 점점 커지고 있음

(3) 인공지능

- 기계로부터 만들어진 인공적인 지능
- 학습과 문제해결과 같이 인간의 인지 기능을 모방해 만들어진 기계의 지능
- symbolic, sub-symbolic 두 가지 인공지능 방법론이 있음

① Symbolic

- 지식베이스를 활용한 인공지능
- 1970년대 중반 ~ 1980년대 후반까지 AI 연구의 지배적인 패러다임
- 지식을 수집하고 이를 컴퓨터가 처리 및 이해할 수 있는 형태로 구축 및 저장한 후 추론, 검색 등과 같은 처리를 통해 사용자의 의사결정이나 질의응답 같은 지능형 처리 지원
- 대표적인 기술: 온톨로지
- 대표사례: IBM 왓슨
- 자연어 형식으로 된 질문들에 응답할 수 있는 질의응답 인공지능 시스템
- 2011년 Jeopardy! 퀴즈쇼 우승

② Sub-Symbolic

- 기계학습을 활용한 인공지능
- 1980년대부터 본격적인 연구 시작
- 지식의 표현 없이 기계학습과 같은 방법을 이용하여 학습, 패턴 인식과 같은 분야에 활용
- 컴퓨터에 명시적으로 프로그래밍을 하지 않고 통계기술을 사용하여 학습을 수행하는 인공지능으로 artificial neural networks, support vector machines, clustering 등 과 같은 다양한 알고리즘과 및 그 외 딥러닝 방법론 포함
- 대표적인 사례: 알파고
- 알고리즘: 심층신경망(DNN)과 몬테카를로 트리 탐색(MCTS) 알고리즘을 바탕으로 반복

적인 자가 대국으로 성능 개선하는 강화학습 수행
- 학습데이터: KGS 바둑 서버에 등록된 16만개 기보, 3천만개 착점 위치와 정보 패턴

(4) 데이터의 중요성

- 인공지능의 성능을 향상시키는데 데이터가 더 중요한지 알고리즘이 더 중요한지 정해진 답은 없음
- 한가지 확실한 것은, 어떤 인공지능 소프트웨어를 만들고 싶을 때, 가장 크게 한계에 부딪히는 것은 데이터라는 점
- "우리는 다른 사람들보나 더 나은 알고리즘을 가지고 있지 않다. 다만 우리는 다른 기업들보다 더 많은 데이터를 가지고 있을 뿐이다." - 구글 치프 사이언티스트 Peter Norvig
- 알고리즘도 알고리즘이지만 그만큼의 기술을 만들어 낼 수 있었던 건 결국 데이터 때문이라는 점을 암시

(5) 제한된 데이터 환경에서의 문제점

- 현재 대부분의 데이터가 인터넷에서 이미 우위를 선점하고 있는 대형 플랫폼에 종속되어 있는 실정
- 경쟁과 혁신의 감소로 신생기업이나 중소기업이 새로운 아이디어를 통해 문제를 해결할 수 있는 기회의 제한
- 전체적인 AI시장 및 AI로 얻을 수 있는 혜택 감소

(6) 인공지능을 위한 데이터 개방의 중요성

- 데이터 개방을 통한 알고리즘 개선 사례: SQUAD, IMAGENET
- 인공지능을 위한 데이터를 개방함으로서 중소기업의 경쟁력을 올리고 공정한 AI 시장에서의 경쟁체계를 구축할 수 있음을 알 수 있음

① SQUAD
- 위키피디아에 있는 문서 참여자들에 의해 만들어진 질의응답 관련 데이터 셋
- 500개 이상 문서로부터 본문 - 질문 - 답변 model 의 데이터셋 10만개 제공
- 매년 콘테스트를 개최 하는데, 한 사례의 경우, 2011년 67%에서 2018년 89% 까지 정확률이 올라감

② IAMGENET
- 이미지 프로세싱 데이터 공유 웹사이트
- 시각적 개체 인식 소프트웨어 연구를 위한 대규모 이미지 데이터베이스
- 2천개 이상의 분류로 1400만개 이상의 이미지 데이터 포암
- ILSVRC(ImageNet Large Scale Visual Recognition Challenge)
- 2010년 정확도 71.8% → 2017년 97.3%로 알고리즘 성능의 비약적 성장

③ 그 외 제공되는 개방형 데이터셋
- 지식베이스 기반 개방 데이터셋: Linked Open Data(LOD)
- 기계학습 기반 개방 데이터셋: Open ML

(7) 경험을 통한 교훈

- 회사에서 서비스를 제공하면서 실제 직면한 어려움과 교훈

① KEF (knowledge extension framework)
- 지식베이스 구축을 위해 비정형 데이터로부터 지식을 추출하기 위한 기계학습 기반의 플랫폼
- 학습데이터: Wikipedia의 Infobox를 이용해 학습 수행
- 120,000 개의 Wikipedia 문서를 학습하여 200개 이상 분류와 3,000개 이상 다양한 유형의 지식 추출
- 학습데이터를 만들기 위해 데이터를 만드는 과정에서, 자연어 태깅을 하여 해도 본문에 해당 값이 없어서 만들지 못하는 문제 발생
- 형태소 분석기를 쓰는 과정에서 값이 이상한 위치에서 잘리거나 이상한 값이 잘려나오는 경우 있는 등 자동화를 하면 어딘가에서 분명히 오류가 발생하기 마련이고 이는 인공지능의 성능에 영향을 미칠 수 밖에 없음
- **경험을 통한 교훈 #1:**
- 좋은 데이터를 확보하는 일이 쉽지 않다
- 좋은 데이터가 있다면 더 좋은 성능을 발휘할 수 있다
- 일반 문서를 가진 데이터는 어떻게? (ex. 뉴스)

② ALVIS
- 지식베이스를 기반으로 한 지능형 질의응답 플랫폼
- 지식베이스를 기반으로 사용자의 자연어 질의를 해석하고 그에 적합한 응답을 지식베이스로부터 찾아 변환하는 질의응답 플랫폼
- 한국사 관련 질의응답 플랫폼 구현을 위한 한국사 지식 베이스를 구현하기 위해 국사편찬위원회의 한국사 LOD, 한국학중앙연구원 한국민족문화대백과사전, 이뮤지엄 이용
- 지식베이스 구축 시 문제점: 간단한 수집방법의 부재(LOD, Open Api 등), 정규화 되어있지 않은 값들, 통일되어 있지 않은 데이터 형식, 동일 개체에 대한 서로 다른 아이디, 의미없이 구성된 데이터 구조 등 으로 가공에 시간이 오래 걸림
- **경험을 통한 교훈 #2:**
- 좋은 데이터 구축은 인간의 노동을 요함
- 웹 페이지 같이 가공된 데이터 보다는 날 것 그대로의 데이터 (raw data) 가 필요
- 데이터를 잘 만들면 부족할지언정 공유해야 함

(8) 더 나은 인공지능 시대를 위하여

- Better data is always better
- 인공지능에서 알고리즘의 성능을 높이기 위해 대량의 고품질 데이터가 중요
- 가능한 데이터를 공개적으로 사용할 수 있고 기계가 읽을 수 있는 형식의 데이터를 만드는 것이 중요
- **정부의 역할**
- 정부가 보유 및 생간(직접적이든 간접적이든)한 데이터는 엄청난 자산임을 인식
- 인공지능을 위해 데이터를 원천 데이터 형식으로 수집 및 제공
- 사용자가 보다 쉽게 데이터를 발견하고 이용할 수 있도록 다양한 형태의 노력 필요
- 데이터의 유통이 일어날수록 개인정보에 대한 철저한 관리와 규제 필요

2. 인공지능 비서와 데이터 (엄재홍 발제자)

(1) AI 서비스 등장 배경

- 새로운 인터페이스의 등장 및 대중화
- 스마트폰의 보급 → OS Infra, Device Infra, Sensor Infra 확보
- UX/UI를 통한 사용자 니즈 증대
- 딥러닝 등의 새로운 기술의 등장으로 인식률 향상, 관련 기술의 현실화
- **질의응답(QA) = DS +새로운 인터페이스의 적용**
- 보다 편리한 인터페이스를 이용한 DS(Decision Support) Task 사용

- 인공지능에 대한 외형적 기대 증가, 내적 품질 기대의 점진적 증가

- 대화 인터페이스

- 가장 인공지능적 모습으로 의사결정 보조/지원을 제공하는 방식
- 운전, 생활, 육아 등 core activity의 보조에 적합
- ex. 음성인식, 스마트스피커

- 다양한 관련 기술의 발전

① 음성인식(ASR):

- 사용자의 음성명령을 이해하고 기계가 처리할 수 있는 텍스트 형태로 변환
- 한영 변환 오류나 사용자의 특이 발화(ex. 사투리), 발화오류(ex. false awakening) 등을 수정하기 위한 언어모델 필요

② 자연어이해(NLU):

- 음성인식을 통해 인식되거나 사용자가 직접 입력한 명령을 분석해 의도 해석
- 문장의 구조 및 주요 성분 분석을 통해 문장을 해석하고 통계가 규칙을 이용해 문장 분석 수행
- 활용/변화가 큰 한국어에 대한 올바른 처리 위해 정확한 한국어 형태소 분석기 등 필요

③ 지식베이스(KB):

- 서비스를 위한 도메인(해당 분야) 데이터를 가공해 서비스에 적합한 형태로 제공
- 지식베이스 모델링 및 데이터 가공, 데이터 구축을 위한 기능적 pipe-line 필요
- 서비스를 위한 데이터의 주제범위(coverage) 및 수준(depth)에 대한 서비스 정책 필요

④ 텍스트 음성 변환, 음성합성 (TTS, Speech Synthesis):

- 사용자의 질문에 대한 응답 스크립트나 답변을 음성으로 변환해 제공
- 주요 단어 및 표현에 대한 사전 녹음 등의 작업이 필요하며 녹음 자료가 없는 경우 적절히 합성

- 음성인식 스피커 - 서비스 개요

- 고객, 외부 개발자와 인공지능 플랫폼 서비스의 진화 방향을 목표로 시작
- 서비스 시작: 음악, Smart Home, 일정, 알림, 날씨 서비스를 기본으로 시작
- 업그레이드: AI Cloud 기반 기능 업데이트로 서비스 기능 추가 (뉴스, 주문배달, 무드등 ...)

(2) 자연지능 vs. 인공지능

- 각각 장단점이 다르며 아직 한쪽이 어느 한쪽을 모두 대체하기에는 부족한 수준

- 자연지능의 장점

- 에너지 효율적
- 다양한 것을 다 할 수 있음(Universal)
- 멀티태스킹
- 복잡한 움직임 가능

- 인공지능의 장점

- 효율적인 처리속도
- 24/7/365 가동 가능
- 편향되어있지 않음
- 정확성
- 결론: 사람이 먼저 있고 인공지능이 보조적으로 사람의 효율을 높여주는 방향으로 가야 함

(3) AI Assitant

- 인공지능비서의 시장가치는 갈수록 높아지고 있음
- 정의: 과제지향적(task-oriented) 정의
- 차이점: 각기 다른 자동화 수준, 다양한 타겟 도메인, 다양한 타겟 과업 등
- 기기를 기반으로 한 플랫폼 전쟁이 계속되는 중
- 국내에서도 많은 기업들이 AI 스피커, 음성인식 AI 비서 시장에 뛰어들고 있음 (SK텔레콤, KT, LG U+, NAVER, 카카오, LG전자, 삼성전자 등)
- 국제적으로는 국내보다 먼저 진행되고 있었음
- 선두주자는 Amazon의 Alexa가 선두로 나서고 있음. 자사의 AI스피커인 Echo 및 Echo 시리즈를 통한 기기제어 및 쇼핑
- AI 비서 및 스피커를 통해 기기나 서비스 환경을 디자인 할 수 있게 됨
- 홈 허브를 목표로 함

(4) Data

- 실세계(real world)는 데이터의 세상
- 개념과 개념의 실체인 인스턴스 존재
- 개개의 개념들마다 실제 객체들이 존재하며
- 개념적 위계가 존재 (ex. 워드넷)
- 현재 우리는 어마어마한 데이터를 가지고 있지만 그 중 실제 활용하는 데이터는 일부분에 불과
- 이것은 마치 장님 코끼리 만지기와 같아서, 데이터는 골고루 만져봐야 대략적인 추세를 알 수 있지만 현실적으로 어려움
- 구글이 각종 서비스를 공짜로 제공하는 이유는 그 서비스를 통해 사람들의 데이터를 모아 알고리즘을 강화하기 위한 것

(5) AI&Data

- 4차 산업혁명은 자동화 혁명
- 다양한 데이터가 생성되면서 빅데이터를 분석하기 위한 알고리즘이 등장
- 머신러닝 전문가는 각자 알고리즘 도구를 이해하고 있을 뿐만 아니라, 어 떤 일에는 어떤 도구가 가장 효율적이고, 어떤 제약이 있는지 알고 있어야 함
- 이제는 기존 공장이 자동화 되는 것을 넘어 서비스(ex. 상담)까지도 자동화 되고 있는 상황
- 이를 나쁘게만 볼 것이 아니라 일자리의 품질이 좋아질 수 있다고 해석할 수 있음
- AI를 학습시키기 위해서는 데이터가 필요한데, 데이터의 특징은 휴먼리더블이 아닌 머신리더블 형태로 제공되어야 함
- 휴먼리더블 데이터: 상징, 이미지, 다이어그램 등
- 머신리더블 데이터: 숫자형 데이터, 벡터 등
- 많은 휴먼리더블 데이터가 정규화 되어있지 않아 활용되지 못하고 있음
- 머신리더블데이터로 가공하는데 시간과 비용이 많이 들다보니 기업에서 공개를 안 하는 악순환이 반복

(6) (AI) Assistant & Data

- 인공지능 비서 서비스를 위해 음악, VOD, 날씨, 시간, 주식정보 등 다양한 도메인의 데이터 필요

ex. 음악 데이터 : Melon, Bugs Music, Genie Music ...
VOD메타데이터 : B tv, Olleh tv ...
Podcast 메타데이터 : Podbbang ...

- 기존에 계약한 데이터를 다른 데에서도 활용하려면 다시 계약을 해야 하는 등 저작권 문제 발생.

- 저작권 문제를 해결하기 위해 위키피디아 같은 오픈데이터나 LOD를 활용하게 됨
- 그러나 한국 위키피디아는 영어위키피디아에 비해 데이터가 부족함.
- Amazon에서 Alexa 서비스가 나왔어도 국내 기업들이 따라잡지 못한 것은 공공데이터나 신뢰성 있는 데이터가 부족했기 때문에 자연히 노하우도 부족했기 때문임
- 데이터를 정제된 형태로 가공해 품질을 보증하는 역할 필요

(7) 결론

- 알고리즘에 대한 공유는 많이 일어나고 있지만 (ex. 깃허브) 데이터는 그러지 못한 상황
- 데이터도 오픈소스형 데이터 플랫폼이 생겨야 함
- 로우 데이터 - 전처리된 표준화된 데이터 - 상품으로서의 데이터 - 데이터 생태계를 위한 데이터 플랫폼
- 각각의 ICT 요소들에 대한 기술 노하우 + 데이터 - 서비스의 효과적인 개발 - 서비스 생태계로 확장될 수 있음

Ⅱ. 플로어

(플로어)

- 만약 샤오미 밴드에 10년 동안의 내 행동 히스토리를 저장해 분석하고 싶다면 어느 정도 수준으로 저장해야 하는가? 또한 음성비서나 IT 카메라도 데이터 저장 기능이 있는가.

(이명진 패널)

- 행위를 구분할 수 있는 데이터의 기준점과, 저장 시간 단위에 대한 기준을 명확히 수립해야 함
- AI는 결국 우리가 하고자 하는 목표가 아니라 하기 위한 도구이기 때문에, 활용하고자 하는 목표를 설정 한 뒤, 이를 위해 데이터가 어느 수준까지 필요한지 인식하고 그 기준을 세워야 함
- 만약 데이터가 나중에 어떻게 쓰일지 모르겠다면 무조건 자세히 저장할 것 같음

(엄재홍 패널)

- 우리나라는 화이트리스트제도와 블랙리스트제도라는 미국과 다른 정책이 있음. 스타트업 할 때도 문제가 많이 생기고, 핀테크 어플리케이션을 만들기도 어려움
- 음성비서의 경우에도 제한이 있음. 통신사(LG)이기 때문에 개인정보 보호법의 영향을 많이 받음
- 데이터를 잘 저장하는 방법에 대한 연구가 많이 있었음
- 라이브로깅 프로젝트라는게 있는데, 이는 굉장히 롱스팬이 필요함. 1년 365일을 도는 사이클이 여러 번 쌓여야 뭔가 가져갈 수 있음. 일단 10년을 아무것도 못하고 데이터만 저장해야 함.
- 모든 데이터를 다 저장한 뒤 식별할 건지 아니면 필요한 데이터만 저장할 건지에 대한 설계가 필요하고, 이런 방식에 대한 디자인이 필요
- 데이터 수집에 대해 개인정보와 같은 여러 문제가 있고, 관련 법규에도 문제가 있다보니 회사에서는 필요한 데이터만 익명화 시켜서 활용
- 데이터 수집과 관련한 제약이 많기 때문에 이를 해결하기 위한 서비스를 디자인하면 중요한 사업을 할 수 있을 것
- 애플은 경우 센서데이터 많이 로깅을 하는데, 애플이 개인정보법 법규를 피하기 위해 중간에 회사를 두어 모든 센서데이터를 그 회사에서 처리함. 그 회사에서 개인을 식별할 수 없게 익명화를 하고, 데이터가 애플에 넘어가는 식으로 법을 우회함. 삼성에서도 비슷한 방식 사용.
- 잘 저장하는 것도 중요하지만 수집에 따르는 문제점을 살펴보는 게 먼저일 것 같음

(플로어)

- 데이터를 팔 때 품질을 보증하는 단체가 존재하는가

(엄재홍 패널)

- 데이터의 품질보증을 위한 객관적인 단체는 없음
- Amazon의 Alexa같은 경우 위키피디아 처리를 하는데, 위키데이터를 가공해서 Amazon에 제공하는 회사가 따로 있음. 그곳에서 에러 컨트롤 등 all about wiki를 다 처리
- SKT에서도 상원군 이라는 역사적 인물과 관련된 에러가 있었다. 상원군 하면, 상원이란 아들이나 친구를 말함에도 상원군이라는 역사적 인물로 인식하는 에러였음. 이런 문제를 방지하기 위해서는 데이터 품질 자체에 대한, 에러 가능성에 대한 컨트롤 할 수 있는 태그를 달거나 상원군이라고 하면 앞에 위키에는 상원군 '왕, 왕조' 이렇게 써 있지만 앞에 '역사 인물'이런 것 붙여줘서 혼돈이 안 되게 컨트롤 해 줘야 함
- 이를 위한 가이드가 아직 없고 서비스를 만드는 사람들이 각자의 디자인 정책에 맞게 시행착오를 반복하면서 만들고 있기 때문에 사실은 그게 틈새시장이라고 볼 수 있음
- 오픈데이터, 링크드데이터 등 되게 많은 데이터가 존재하지만 오류가 생기면 안 되기 때문에 데이터를 중간에서 무결성있게 관리해주는 사업만 해도 굉장히 수요가 높음. 모두가 품질에 대한 고민을 하고 있음.

(오원석 진행자)

- 데이터는 정답이 없기 때문에, 있어서도 안 된다고 생각함

	(플로어) - 미국에서는 구글이나 위키피디아를 통해 많은 데이터를 얻을 수 있지만 국내에서는 그런 데이터를 솔직히 얻을 수 없는 실정이다. 국내에서 데이터를 사고파는 것에 대한 기업생태계 인식이 어떤지, 차후에 한국사 데이터를 공유하신다면 어떤 방식으로 공유할 건지 궁금하다. **(이명진 패널)** - 한국은 데이터는 누구나 가져다 쓸 수 있다는 인식이 너무나도 넓게 퍼져 있어 데이터를 사고파는 것에 대한 인식이 좋지 않음 - 국가차원에서 데이터 거래소등을 만들기 위한 프로젝트를 진행 중이지만 잘 안 되고 있음 - 한국사데이터는 데이터를 잘 만들어서 공유하겠다는 것이 아니라 더 이상해봐야 더 많은 노력이 들어간다는 판단 때문에 여기까지 하고 말아야지라는 생각 때문에 기준을 정한 것 - 데이터를 활용하는 목적이 질의응답(QA)를 위한 것이라 지식데이터를 위한 형태로 만들었고, 전체데이터를 덤프 받거나 혹은 오픈API를 통해서 역사인물 목록을 가져가거나 인물 개개인에 대한 정보를 가져갈 수 있는 형태로 제공을 하려고 계획 중 **(오원석 사회자)** - 로우한 데이터가 더 많이 나와야 함 - 이미 오픈데이터라고 하는 생태계가 있고 공공데이터가 로우하게 더 많이 생산된다면, 로우데이터를 정제하는 회사도 생기고, 지식베이스를 만드는 회사도 생기고 그걸 유통하는 회사도 많이 생겨나는 등 데이터 거래가 충분히 일어날 것