KISTI
ISSUE BRIEF

ISSN 2635-5728

과학기술인프라, 데이터로 세상을 바꾸다

https://kisti.re.kr/promote/post/issuebrief

2021. 12. 22. 제38호

오픈 사이언스 활성화를 위한 AI 기술 동향

이경하 · 설재욱 · 이종원 · 선충녕

오픈 사이언스는 과학지식과 데이터, 자료에 자유롭게 접근할 수 있게 함으로써 개방적인 정보 공유와 협력을 가능하게 하고, 나아가 사회 구성원들이 과학지식의 생산과 확산에 보다 적극적으로 참여하도록 하기 위한 움직임이다. 하지만 오픈 사이언스에 있어서 여러 제약 또한 존재한다. 이러한 오픈 사이언스를 효과적으로 지원하기 위해서는 기존 논문 출판 체계와 키워드 검색을 통한 자료 접근 이외에 좀 더 세분화된 지식의 구분과 이들의 연결, 접근 및 분석의 용이성을 강화할 필요가 있다. 최근의 AI 기술, 특히 AI 기반 자연어 처리 기술은 오픈 사이언스 활성화에 큰 역할을 할 수 있다. 이에 본 고에서는 오픈 사이언스를 위한 AI 기술 활용을 위해 현 상태를 점검하고 활용 방안을 논의하고자 한다.

CONTENTS

1. 들어가며
- 인공지능 시대의 도래
- 인공지능의 응용
- 오픈 사이언스의 정의와 배경
- 오픈 사이언스 구성 요소

2. 디지털 큐레이션과 AI
- 디지털 큐레이션의 정의와 범위
- 디지털 큐레이션을 위한 인공지능 기술 동향
- 디지털 큐레이션을 위한 인공지능 실현 방안

3. 오픈 콜라보레이션과 AI
- 오픈 콜라보레이션의 정의와 범위
- 오픈 콜라보레이션을 위한 인공지능 기술 동향
- 오픈 콜라보레이션을 위한 인공지능 실행 방안

4. 결 론
- 오픈 액세스 측면에서 한계점과 시사점
- 오픈 데이터 측면에서 한계점과 시사점
- 오픈 콜레보레이션 측면에서 한계점과 시사점
- 오픈 사이언스 활성화를 위한 AI 기술 개발 방향

KiSTi www.kisti.re.kr 한국과학기술정보연구원 Korea Institute of Science and Technology Information

1. 들어가며

인공지능 시대의 도래

- **(인공지능) 인간의 학습능력, 추론능력, 지각능력을 인공적으로 구현하려는 컴퓨터과학의 세부 분야임**
  - 1958년 Rosenblatt에 의해 초기 퍼셉트론 개념이 도입되었으나, Marvin Minsky의 비판에 의해 인공지능 분야는 1차 암흑기[1](1974~1980)에 빠짐
  - 1980년에 이르러 전문가 시스템(expert system)이라 불리는 인공지능 기술들이 보급되었으나 제한적 성능으로 다시 '87-'93년까지 2차 암흑기를 겪음
  - 1990년 중순부터 기초적인 뉴럴 네트워크들이 제시되고, LeCun 등에 의해 문자 인식 등 문제에 실제 적용 가능함에 따라 다시 각광
  - 뉴럴 네트워크는 계층이 깊어질수록 보다 나은 가용성(capacity)을 가지나, 역으로 학습이 잘 되지 않는 문제 발생. 이 문제를 G. Hinton이 2006년 해결가능함을 보임에 따라 보다 깊은 층의 뉴럴 네트워크를 구성, 여러 분야에서 압도적인 성능을 보임

인공지능의 응용

- **(이미지 및 영상 분야) 이미지 또는 영상을 입력받는 AI 응용으로 이미지 분류, 객체 인식, 얼굴 및 자세 인식 등이 대표적인 세부 분야임**
  - **(이미지 분류)** 숫자, 알파벳, 동식물, 음식, 사물 등에 대한 분류. 차량번호 인식 등이 대표적인 응용 사례임
  - **(객체 인식)** 이미지 내에 존재하는 여러 객체들이 무엇인지 인식하여 분류함
  - **(얼굴 인식)** 이미지 내의 사람 얼굴을 인식하거나 자세를 인식. 스마트폰의 얼굴 인식 잠금 해제 등 이미 광범위하게 상용화됨
  - **(모션 인식)** 영상에서 특징을 추출하여 사람의 행동을 인식함
  - **(자율 주행)** 차량 주행 중 차선 및 차량, 장애물 등 영상에서 다양한 객체와 영역을 분리, 인식하고 이에 따라 적절히 차량을 주행함
- **(음성 분야) 음성 데이터를 입력받는 AI 응용으로 음성 인식, 화자 인식, 감정 분석, 챗봇 등 여러 분야에 적용됨**
  - **(음성 인식)** 음성 인식은 사람의 또는 사람 간 대화, 동물 소리, 사물 소리, 기계적 잡음, 음악 등 여러 음성들에 대해 분류 및 의미 파악 등 다양한 분야에 적용

    ※ 단순히 대화 인식뿐만 아니라 음악의 분류, 잡음 인식을 통한 잠재적 기계 고장 파악 등에도 활용됨

1) 암흑기라 함은 인공지능 연구와 관련한 자금 지원이 심각할 정도로 중단되고, 연구가 더 이상 진전되지 못하였던 시기를 의미

- **(화자 인식)** 입력받은 음성을 통해 여러 화자들 중 화자를 식별하는 기술로 회의 발언의 의사록 자동 작성 등에 활용
- **(감정 분석)** 음성 감정 인식은 음성 데이터의 특성을 분석하여 인간의 감정을 인식
- **(음성 챗봇)** 사용자와의 대화를 통해 소통하는 기계를 의미하며, 음성을 통한 명령어를 인식하여 대답

  ※ 음성 인식을 통해 대화를 텍스트로 변환, 이를 이해하고, 답변 텍스트를 작성 후 다시 기계적으로 발화하는 과정을 통해 동작함

- **(텍스트 분야) 텍스트를 입력받아 처리하는 AI 응용으로 기계 독해, 기계 번역, 자연어 추론, 문장 생성 등의 다양한 활용 분야가 있음**
  - **(기계 독해)** 주어진 문서에서 특정 질의에 대한 응답을 제시하는 과제로 Q&A라고도 함
  - **(기계 번역)** 한 언어로 쓰여져 있는 텍스트를 이에 대응하는 다른 외국어로 번역하는 AI 응용 분야임
  - **(자연어 추론)** 문장들에서 서로 관련 있는 개체들을 인식하고 이를 통해 이들 간 관계나 결과를 추론. 비슷한 분야로 멀티 홉 질의 처리가 있음
  - **(문장 생성)** 주어진 입력에 따른 문장의 자동 생성, 전체 문서를 일부로 요약하는 문서 요약이나 특정 문체를 닮도록 또는 임의로 시나 소설 신문 기사를 생성하는 등의 응용이 존재함
  - **(문서 처리)** 대규모 문서 그룹에서 유사한 문서끼리 구분, 분류하거나 또는 문장 내 개체 식별, 관계 추출, 링킹 등 다양한 문서를 처리

오픈 사이언스의 정의와 배경

- **(오픈 사이언스) 과학지식과 데이터, 자료에 자유롭게 접근할 수 있게 하고 개방적인 정보 공유와 협력을 가능하게 하며, 나아가 사회 구성원이 과학지식의 생산과 확산에 적극적으로 참여할 수 있게 하는 움직임(유네스코한국위원회, 2020)**
  - **(새로운 개방)** 인터넷 보급과 디지털 기술의 발전에 따른 과학지식의 공유와 접근의 편의성 확대로 인해 오픈 사이언스가 보다 발달함
  - **(과학계 수요)** 디지털 기술을 통해 새로운 지식과 기술 해결책을 먼저 발견하여 연구와 혁신을 선도하려는 수요가 오픈 사이언스의 발달 촉진
  - **(신뢰성과 재현성 제고)** 연구 과정과 자료를 공유·공개함으로써 연구의 신뢰성과 재현성 제고

- **(출현 배경) R&D 패러다임의 변화, 연구성과물의 자유로운 접근 요구, 글로벌 아젠다로 급부상함**
  - **(R&D 패러다임의 변화)** 초대형 실험장비와 방대한 데이터 분석을 통한 과학연구 수행으로 패러다임이 전환

    ※ 데이터 분석 중심의 연구와 디지털 기술, 새로운 협업 도구를 이용한 연구 협력의 중요성 강조
  - **(연구성과물의 개방과 자유로운 접근)** 공적 자금으로 지원된 연구 결과물의 공유와 활용을 장려함

    ※ 데이터 처리, 분석 비용의 증가와 공유, 협업에 필요한 정보통신인프라의 필요성, 비싼 출판 및 구독 비용 등으로 인해 필요성 대두
  - **(글로벌 아젠다)** 오픈 사이언스 활성화가 세계적으로 중요한 의제로 채택되고 관련 정책[2]이 추진 중

오픈 사이언스의 구성 요소

- **(오픈 액세스, OA) 어떤 금전적, 법적, 기술적 장벽 없이 이용자가 합법적 목적 달성을 위해 자유롭게 연구결과물에의 접근, 복제 및 배포를 허용하는 것임**
  - **(종류)** 오픈 액세스(Open Access)는 크게 셀프 아카이빙과 OA 학술지로 구분
  - **(셀프 아카이빙)** 학술지에서 동료평가를 거친 논문을 저자가 자신의 홈페이지나 어딘가 집중된 OA 저장소(예: arXiv)에 올리는 것, Green OA라고도 함
  - **(OA 학술지)** 학술지 출판사가 논문들을 무료로 공개

    ※ 저자나 스폰서를 통해 논문 처리 비용(Article Processing Charge)을 지불하고 공개하는 Gold OA, 저자가 APC를 지불한 논문에 한해서만 공개하는 하이브리드 Gold OA 등의 유형 존재
- **(오픈 데이터, OD) 저작권, 특허 등의 기술적 법적 제약 없이 데이터에 접근, 활용, 재활용할 수 있도록 데이터를 개방하는 것임**
  - **(대상)** 출판 이후 뿐 아니라 연구 과정 중의 연구 자료, 관찰 및 실험 자료, 메타데이터 등 연구과정 전반에 있어서 생산되는 데이터
  - **(종류)** 오픈 연구 데이터(Open Research Data, ORD), 오픈 정부 데이터(Open Government Data, OGD), 오픈 공공 데이터(Open Public Data) 등 존재

2) 2004년 OECD 과학기술장관 회의에서 "공적자금이 투입된 연구데이터의 접근에 관한 선언" 채택. EU는 2014년 Horizon 2020 프로젝트를 중심으로 오픈 사이언스 촉진을 위한 프로그램 및 정책 추진. UNESCO 등 국제기구에서 오픈 사이언스 활성화를 위한 다양한 정책 개발

- **(오픈 콜라보레이션, OC)** 연구데이터, 연구 방법론 및 연구 인프라, 도구의 공개, 공유, 상호호환을 통한 연구 협업을 의미함
  - **(조건)** 공동의 결과물을 생산할 수 있도록 지원하고 협업의 진입 장벽을 낮추고 유연한 사회구조를 뒷받침할 수 있는 기술을 기반으로 한 협업 플랫폼 필요
  - **(종류)** 연구자 간 협업, 연구자와 기업 간 협업[3], 연구자와 시민[4] 간 협업 등이 존재

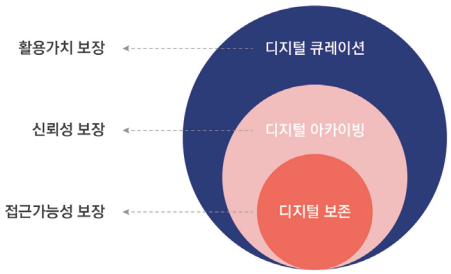
2. 디지털 큐레이션과 AI

디지털 큐레이션의 정의와 범위

- **(정의)** 디지털 정보자원의 장기보존, 접근성 증대, 공유, 재사용 정책을 통해 수요자에게 맞춤형 정보를 적시에 제공·공유하는 행위임
  - **(유사정보 연계)** 인터넷에 널린 정보 등을 주제별로, 혹은 관련된 연계성, 연관성을 지닌 무엇인가를 모아서 정돈하고 정리해서 스스로나 다른 사람에게 보여주고 공유하는 작업임(조병호, 2013)
  - **(빅데이터 큐레이션[5])** '데이터의 숨은 가치와 잠재력 발굴'을 추구하는 활동임(박성민 외, 2013)

<그림 1> 양질 데이터 선별을 위한 큐레이션

<그림 2> 디지털 큐레이션의 범위



- **(범위)** 디지털 큐레이션은 개념화, 데이터의 수집, 평가, 입수, 보존, 저장, 이용, 변환하는 내용을 포함함(이혜림, 2020)
  - **(디지털 큐레이션 라이프 사이클)** 디지털 정보자원을 수집, 관리, 보존, 이용, 변환하는 과정을 지속적으로 수행하기 위한 프로세스를 의미함

3) Open innovation과 유사하나 오픈 이노베이션은 경제적인 가치 창출이라는 명확한 목적 하에 지식 유통을 활용하여 혁신의 내적 가속화와 시장 확산 도모
4) 크라우드 사이언스, 네트워크 사이언스, 시티즌 사이언스 등으로 지칭
5) 빅데이터 큐레이션, 콘텐츠 큐레이션은 디지털 큐레이션과 유사한 의미로 사용

<표 1> 디지털 큐레이션에서의 단위 업무

업 무	내 용
개념화	데이터 캡쳐 방법과 저장 필수 사항 등을 포함하여 데이터 생성과 수집에 관하여 구상하고 계획하는 것
수집	수집 정책에 따라 다른 조직이나 개인 데이터 생성자로부터 데이터 획득
평가	문서화된 정책, 기준, 법적 요구사항 등을 준수하면서 수집한 데이터를 평가
입수	데이터를 DB, 정보시스템, 디지털 아카이브에 추가하기 위해 준비하는 행위
보존	데이터 무결성, 진본성, 신뢰성, 이용가능성을 유지하면서 데이터의 장기 보존을 위해 실행하는 행위
저장	저장 관련 표준에 따라 안전한 방법으로 보관하는 것
이용	데이터가 이용 및 재이용 될 수 있도록 데이터를 접근가능하게 유지하는 것
변환	원본 데이터를 기반으로 새로운 포맷이나 서브데이터셋을 만드는 것

디지털 큐레이션을 위한 인공지능 기술 동향

- **(디지털 큐레이션 서비스 AI 접목) 이미지, 콘텐츠, 뉴스, 음악, 패션 등 다양한 분야의 디지털 큐레이션 서비스에서 이용자의 행태, 관심사, 로그파일 등을 분석하기 위해 인공지능 기술을 적용함**

  ※ 대부분의 디지털 큐레이션 서비스는 서비스 측면에서 인공지능을 접목하고 있음

<표 2> 국내외 디지털 큐레이션 서비스 동향

내 용	서비스 분야	내 용	인공지능 기술
핀터레스트 (Pinterest)	이미지	이미지 기반의 소셜 큐레이션 서비스이며 텍스트, 스크린샷, 카메라 사진으로 유사 이미지를 검색할 수 비주얼 서치 기능 제공	비주얼 임베딩 모델 (AI 유사 이미지 분류 모델)
플립보드 (Flipboard)	콘텐츠	RSS/SNS 계정을 연동하여 제공받은 스토리, 사진, 동영상을 매거진으로 전환하는 서비스	AI 콘텐츠 추천 모델
뉴썸 (Newsum)	뉴스	인공지능이 뉴스를 자동 편집하고 통계 기반으로 맞춤형 뉴스 추천	NAMI[6] (AI 뉴스 추천 인공지능 모델)
에어스 (AiRS)	뉴스	관심사가 같은 이용자 분석을 통한 유사 관심사의 뉴스 콘텐츠 추천	RNN 및 협업 필터링 알고리즘 (AI 뉴스 추천 인공지능 모델)
스포티파이 (Spotify)	음악	사용자에게 맞춤형 음악을 추천하고 사용자의 취향과 비슷한 청취자가 듣는 음원 추천	BaRT[7] (AI 음원 필터링 모델)
멜론 (Melon)	음악	곡 제목, 곡 정보, 음성 특징, 태그 등을 이용하여 장르, 주제, 컨셉별 음원 세분화를 통해 음원 추천	AI 음원 추천 모델
스티치픽스 (Stitch Fix)	패션	크기, 예산 및 선호 스타일에 따라 의류 품목을 추천해주는 개인 스타일링 서비스	AI 의류 추천 알고리즘

6) NAMI: News Assigning Machine Intelligence
7) BaRT: Bandits for Recommendations as Treatments

- **(인공지능 기술 동향) 큐레이션 서비스는 이용자 선호 정보 분석을 위해 고전적 추천시스템에서 베이지안 네트워크, 딥러닝을 활용한 인공지능 모델 기반 시스템으로 변모함(서봉원, 2016)**
  - **(유사 콘텐츠 추측 모델)** 사용자 선호 콘텐츠를 추측함으로써 여러 가지 항목 중 이용자에게 적합한 특정 콘텐츠를 추천, 제공하는 모델임

    ※ 콘텐츠 기반 필터링(Content-based Filtering): 이용자가 이용한 콘텐츠의 정보를 바탕으로 유사 항목을 추천하는 기술

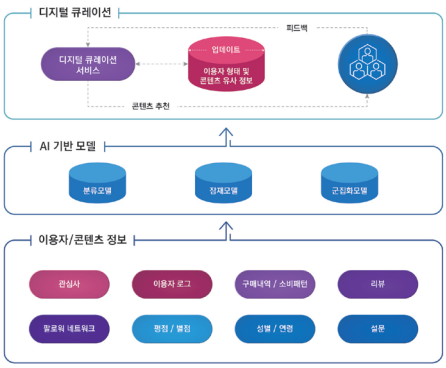
    ※ 협업 필터링(Collaborative Filtering, CF): 대규모의 기존 이용자 행동 정보를 분석하여 해당 이용자와 비슷한 성향의 이용자들이 기존에 좋아했던 항목을 추천하는 기술(서봉원, 2016)

<표 3> 협업 필터링 알고리즘의 종류

종류	구현 기법
메모리 기반 CF	Neighbor-based CF (Cosine Similarity, Pearson-Correlation) User-based/Item-based CF (Top-N Recommendations)
모델 기반 CF	베이지안 (Bayesian belief nets) CF 클러스터링 CF 회귀 (Regression) 기반 CF 잠재 (Latent) CF 차원 축소 (Dimensionality Reduction) CF (SVD, PCA) 뉴럴 네트워크 CF
하이브리드 CF	Content-based CF Combining algorithm

출처) Su 외 (2009)

<그림 3> 디지털 큐레이션 서비스를 위한 인공지능 적용



  - **(사전학습 모델)** 대량의 데이터를 미리 학습하여 주어진 텍스트 데이터 내의 단어들에 대한 의미론적인 표현을 이해하는 모델이며, 2018년 BERT[8]가 공개된 이후 GLUE[9], SQuAD[10] 등 다양한 벤치마크에서 이전 모델들의 성능을 큰 격차로 능가함

8) BERT: Bidirectional Encoder Representations from Transformers
9) GLUE(General Language Understanding Evaluation): 다양한 자연어 이해 시스템을 교육, 평가 및 분석하기 위한 리소스 모음
10) SQuAD(Stanford Question Answering Dataset): 크라우드 소싱을 통해 구축한 위키피디아에 대한 질문-대답 데이터 셋

<표 4> 사전학습 모델 동향

모델 종류	내 용	사전학습 모델
자기회귀 모델 (Autoregressive Model)	- 고전적인 언어 모델링 작업에 대해 사전 훈련 - 이전 토큰을 모두 읽고 다음 토큰을 추측 - Transformer(Vaswani 외, 2017) 모델의 디코더 활용 - 적용 분야: 텍스트 생성	- GPT3 - GPT2 - CTRL - Transformer-XL - Reformer - XLNet
자동 인코딩 모델 (Autoencoding Model)	- 입력 토큰을 손상시켜 원래 문장을 재구성하는 방식으로 사전 훈련 - 전체 문장의 양방향 표현을 구축 - Transformer 모델의 인코더 활용 - 적용 분야: 문장 분류, 토큰 분류	- BERT - ALBERT - ROBERTa - ELECTRA - Longformer
시퀀스-시퀀스 모델 (Sequence-to-Sequence Model)	- 한 문장(시퀀스)을 다른 문장(시퀀스)으로 변환하는 모델 - Transformer 모델의 인코더, 디코더 활용 - 적용 분야: 번역, 요약, 질의응답	- MarianMT - T5 - BART
다중 모드 모델 (Multimodal Model)	- 서로 다른 유형의 값을 입력받아 task 수행(예: 텍스트와 이미지를 입력 받아 분류 task 수행)	- MMBT

출처) huggingface 내용 요약

디지털 큐레이션을 위한 인공지능 실현 방안

- **(학습데이터 신뢰성 제고) 인공지능 학습데이터 제작공정에서 공통적으로 준수해야할 신뢰 확보 검증지표 등의 표준 기준[11] 마련이 필요함(과기정통부, 2021)**
  - **(데이터 신뢰성의 중요성)** AI의 핵심 요소는 양질의 데이터이며, 대량의 데이터보다 목적에 맞게 잘 만들어진 데이터가 똑똑한 AI를 만들 수 있음
    ※ 데이터의 공정성과 신뢰성을 높이는 방안으로 '출처가 정확한 데이터 사용', '사용 목적에 맞는 데이터의 수집과 선택', '제한 사항과 가정의 정확한 언급', '데이터의 편향성 명시', '실제 환경에서의 적절한 테스트 이행' 등 5가지 기준을 제시함(김소영 외, 2021)
  - **(데이터 상호연계)** 메타데이터 연계규격을 바탕으로 데이터 연계·융합·식별하여 상호보완 기반 데이터 신뢰성 제고가 필요함

- **(인간과 AI의 협업) 섬세한 인간과 반복 처리 작업에 강한 기계의 공생 가능한 섬세한 설계가 필요함**
  - **(휴먼인더루프 AI)** 데이터 처리 과정이나 의사결정 시 사람과 기계 간 활발한 상호작용을 토대로 결과를 도출하는 방식을 도입하여 의사결정 시간 단축과 '확증 편향'[12] 현상을 극복할 수 있음
    ※ 플로(FLO)는 AI 추천 모델의 도움을 받아 콘셉트에 맞는 곡들을 크리에이터들이 선정하는 방식으로 '수작업 추천'과 'AI 추천 모델' 간 상호 보완 체계를 도입함

11) 학습데이터 활용 목적에 따라 신뢰 확보 요구사항을 세분화·구체화하고 검증지표, 측정방법 등 제시
12) 원래 가지고 있는 생각이나 신념을 확인하려는 경향으로, AI가 유사 콘텐츠만 추천하여 다른 콘텐츠는 틀리다고 판단하는 현상을 의미

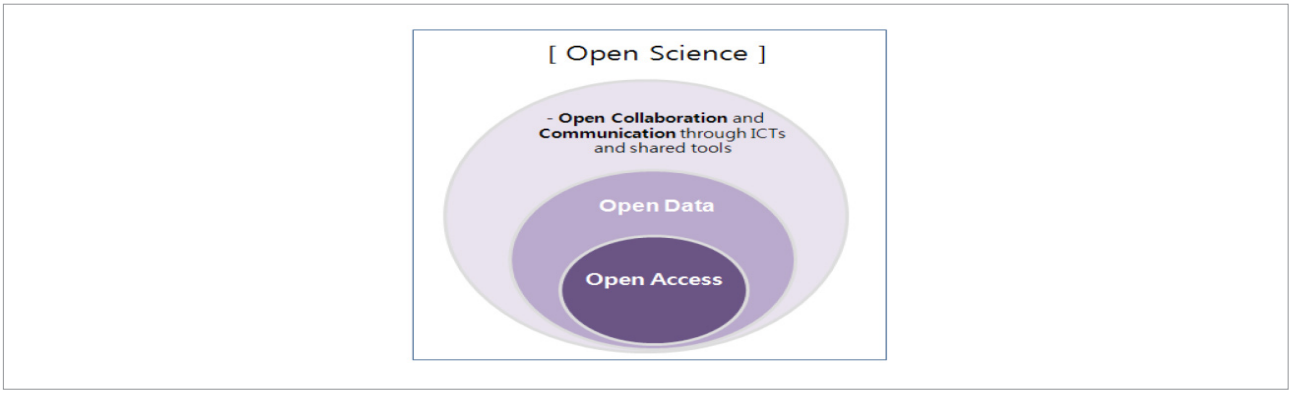
- **(미래지향적 법제도 정립)** AI 시대 기본이념과 원칙, 역기능 방지 시책 등 법제도 마련이 필요함(관계부처 합동, 2019)
  - **(AI의 안정성 확보)** AI 로봇의 공격, 자율주행차 사고 등 AI의 안정성 침해 사례를 통해 AI 사고나 부작용을 방지할 수 있는 법제도 마련이 필요함
  - **(AI 규제 완화)** AI 생태계 및 기술 발전을 위한 인공지능 관련 규제 샌드박스 활용 활성화가 필요함

3. 오픈 콜라보레이션과 AI

오픈 콜라보레이션의 정의와 범위

- **(정의)** 연구데이터, 연구 방법론 및 연구 인프라, 도구의 공개, 공유, 상호호환을 통한 연구 협업임
  - 연구 협력과 소통 강화를 통한 개방형 연구 문화 형성
- **(범위)** 오픈 액세스, 오픈 데이터를 기반으로 연구 협력 및 소통을 개방할 수 있는 수단과 정책, 교육, 문화를 포함함

<그림 4> 오픈 콜라보레이션의 범주



출처) 오픈 사이언스 정책의 도입 및 추진방안(2017)

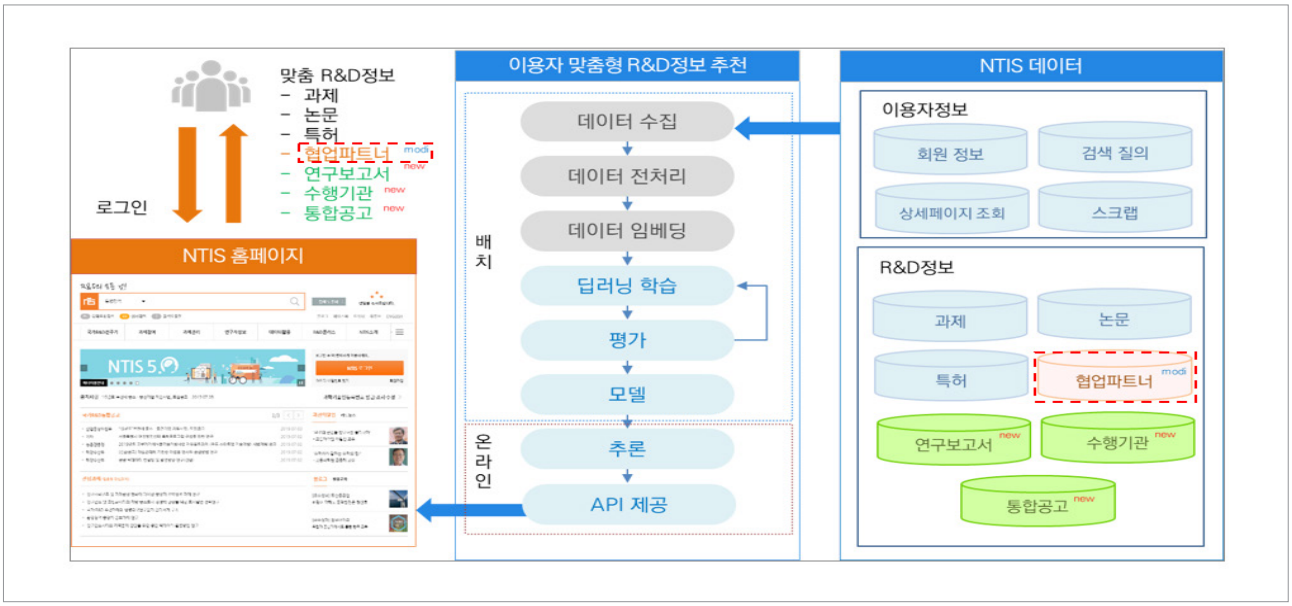
오픈 콜라보레이션을 위한 인공지능 기술 동향

- **(개방형 연구협력의 확대)** 개방형 연구 협력을 통한 다양한 기술 개발 사례가 등장함
  - 공공데이터 또는 연구데이터를 활용한 분석 및 의사 결정 사례
  - 기계학습 데이터 및 벤치마크 데이터셋의 공개 및 리더보드를 통한 AI 모델들의 정확도 경쟁
  - GitHub 등 공개 SW 리포지토리를 이용한 코드 공유 및 협력 개발
  - 논문 서지 정보 분석을 통한 관련 연구자 추천 및 신규 주제 발굴 지원 사례 등

오픈 콜라보레이션을 위한 인공지능 실행 방안

- **(연구자 추천) 논문과 논문 이용자 성향 분석을 통해 해당 주제 관련 협업연구자 정보를 추천하여 연구 협력을 지원함**
  - 인공지능 기술을 활용한 논문 이용자 패턴 분석 및 이를 기반으로 이용자에게 맞춤형 정보[13] 제공
- **(연구 정보 추천) 연구자의 연구 이력을 분석한 관련 연구 정보(논문, 보고서 및 과제 RFP)의 추천**

<그림 5> NTIS 이용자 맞춤정보 추천 서비스 사례



출처) NTIS(2021)

- **(협업 공간 제공) 프로파일이 유사한 연구자들이 소통할 수 있는 개방형 장소를 제공함**
  - 논문, 보고서, 특허 등 과학기술정보와 사용자 관심사 분석, 추천을 통한 연구자 간 네트워크 구축
  - 과학기술정보와 연구데이터의 연계, 연구자 간 관계 등 네트워크 식별 및 분석을 통한 과학기술 주체, 정보, 데이터의 연계
  - 온라인 상에서 협업이 쉽게 이뤄질 수 있도록 개방형 커뮤니티 및 협업 체계 구축

13) 협업파트너(이용자와 연관성이 높은 다른 이용자) 정보와 과제정보, 논문정보, 특허정보 제공

4. 결론

오픈 액세스 측면에서 한계점과 시사점

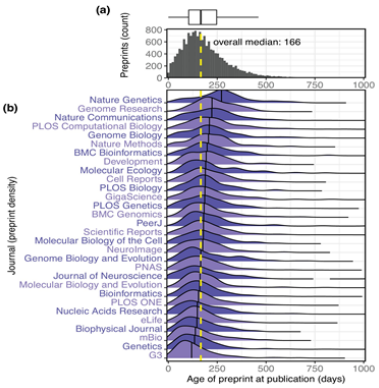
- **(한계점) 출판 논문 수 증가로 논문 발표가 연구 진행 속도를 따라가지 못하는 등의 문제가 발생하고, 오픈액세스 운동을 상업적으로 이용하는 부실학회 등의 출현으로 인한 어려움이 존재함**
  - **(출판급증)** 논문 수가 폭발적으로 증가해 1년에 1백만 편 이상 출판되고 있으며, 연구자가 각자의 분야에서 따라잡기 불가능한 수준임[14)]
  - **(연구속도)** 기존 성과 발표 체계 속도보다 빠르게 연구가 진행되어 이미 사전출판부터 논문이 인용되는 경우가 발생하고[15)], 검증 부족이나 잘못된 결과를 담은 논문 공개로 인한 혼선 등 문제점이 존재함[16)]
  - **(부실학회)** 부실 학술지 조기 검출을 위한 정밀한 학술지, 논문의 품질 분석, 평가 방법이 요구됨
- **(시사점) 연구 자료 접근성을 향상시키기 위해, 논문과 함께 다양한 출처의 자료들을 연결하고 키워드 검색을 넘어 AI 기반 탐색 지원이 필요함**
  - 특정 키워드에 의존하는 검색보다 개략적인 주제만으로도 학술정보를 찾을 수 있는 탐색기반 접근 지원이 필요함
  - 다양한 연구 결과물과 논문과의 연계를 위해 문헌 단위 접근보다 논문의 구조나 역할로 재구성된 정보들이 특성의 형태로 부착되어야 함
  - 단순 표절 검사를 넘어 의미, 구조를 이용한 유사성 검사와 재현 가능성을 측정하는 방법이 필요함

오픈 데이터 측면에서 한계점과 시사점

- **(한계점) DMP를 행정적 활동으로 간주[17)]하고, 연구데이터는 개인 소유라는 지배적 인식과 보상체계 미비로 인해 데이터 공유가 활발하지 못함**
  - **(연구자 인식)** 연구데이터 공유 필요성에 비해 경험이 부족하고 개인 소유 혹은 공동 소유 인식이 강하기 때문에, 오픈 데이터 활동에 소극적임

14) A. Extance, “AI tames the scientific literature,” September 2018, Vol.561, Nature
15) CVPR 2017 참관기, https://www.kakaobrain.com/blog/34
16) [오철우의 과학풍경] ‘사전출판 과학논문’ 어떻게 다룰까?https://www.hani.co.kr/arti/opinion/column/1018620.html
17) Miksa, Tomasz; Simms, Stephanie; Mietchen, Daniel; Jones, Sarah (28 March 2019). "Ten principles for machine-actionable data management plans". PLOS Computational Biology. 15 (3): e1006750. doi:10.1371/journal.pcbi.1006750. PMID 30921316. S2CID 85563774.

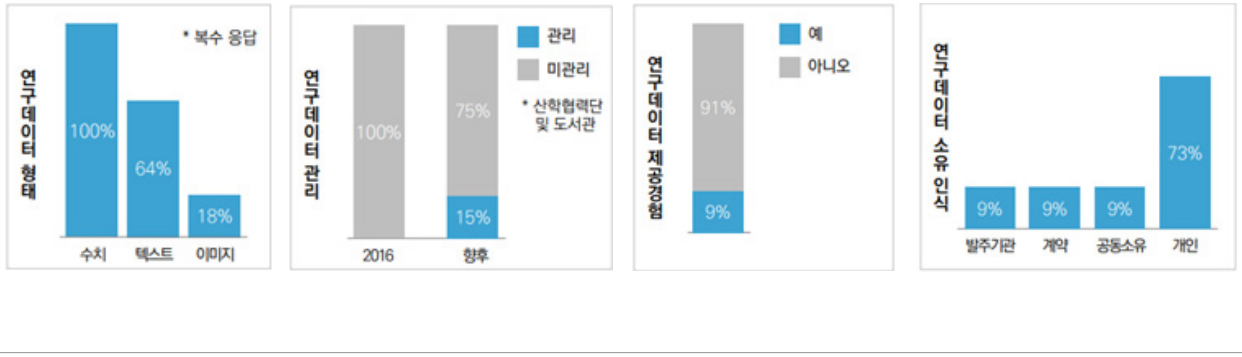
<그림 6> 사전출판(bioRxiv) 게시 날짜와 다른 곳에 처음 출판된 날짜 사이 간격 (중앙값 166일 또는 5.5개월)



- **(정보접근성 문제)** 데이터는 DMP를 통해 메타데이터를 확보하는 수준이며, 해외에서도 DMP를 공유하는 등의 문제가 있음[18)]

- **(시사점) DMP 작성 및 갱신 상황에서 연구자 부담의 경감 방법이 필요하며, 학술 출판 시 데이터 인용을 돕고 데이터 접근성 향상 방법이 요구됨**
  - DMP 작성의 부담을 줄이고 연구 결과의 재현성 확인을 위해 DMP가 갱신되는 상황에서 연구논문과 함께 제출하면 필요한 정보를 추출하여 DMP 내용을 수정/보완해주는 기술이 요구됨
  - 연구데이터를 논문 수준으로 인용한다는 인식 정착이 필요하며, 이를 위해 논문 작성 시 데이터 인용 정보 검색과 논문 삽입을 돕는 도구가 필요함
  - 논문에서 데이터 활용 정보를 수집하고 구축된 데이터와 연계하여 사용자들의 데이터 접근성을 향상시킬 방법이 필요함

<그림 7> 국내 연구자의 연구데이터 인식 조사 현황 (20개 대학, 2016년 KISTI 조사 결과)

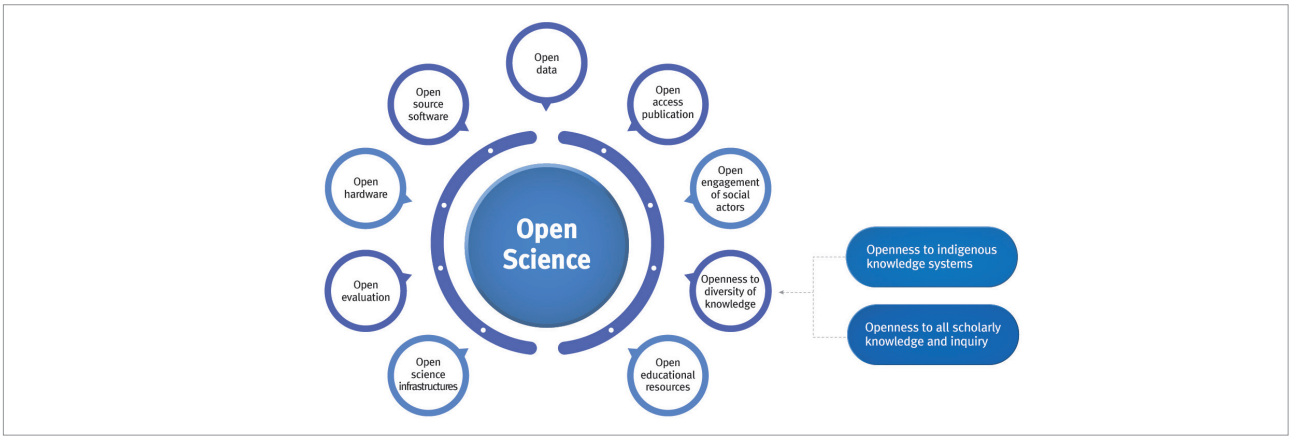


18) NSF DMP content analysis: What are researchers saying? https://asistdl.onlinelibrary.wiley.com/doi/10.1002/bult.2012.1720390113

오픈 콜라보레이션 측면에서 한계점과 시사점

- **(한계점) 콜라보레이션의 개념은 사회적 행위자까지 포함하는 열린 참여의 개념으로 확장되고 있으며, 과학자들이 주가 되던 협력 관계를 넘어 사회 구성원이 포괄적으로 접근 가능하도록 도구와 절차를 개방하는 단계로 변모 중임**
  - **(유네스코 권고)** 궁극적으로 과학에 사회의 다양한 이해관계자가 적극적으로 참여할 수 있도록 하는 사회적 행위자의 열린 참여(Open engagement of societal actors) 개념으로 확장됨[19]

<그림 8> 2021년 2월 17일 유네스코 발전에 기반한 개방형 과학 요소 (위키피디아 2021.11.25. 재인용)



  - **(공유/협업 성공사례)** 오픈백과(위키피디아), 오픈소프트웨어(github)와 같이 대중의 참여로 성공한 공유/협업의 사례들의 특징 분석을 통해 연구 과정에서의 적용 방법에 대한 고민이 필요함

- **(시사점) 협력을 넘어 참여 유도를 위해서는 쉽게 접근하고 적은 노력으로 기여할 수 있는 환경 조성이 필요함**
  - 기존 출판 위주의 연구 결과 공유보다 연구 과정 및 결과가 생산, 소비, 유통될 수 있도록 디지털 플랫폼 기반으로 변화가 필요함[20]
  - 오픈 소프트웨어의 주요 성공 요인은 모듈화된 개발 방법론임[21]
  - 기존 연구는 소수 연구자나 연구그룹에 의해 진행되어 연구 단계 구분이 명확하지 않고 결과물인 논문 작성 및 공유에만 집중함
  - 결과 공유에서 과정의 참여로 전환하기 위해서는 연구 단계나 절차마다 분리하여 다수의 참여자가 공동 작업을 진행하더라도 기여를 명확하게 판별하고 충돌 여지를 줄이는 것이 필요함

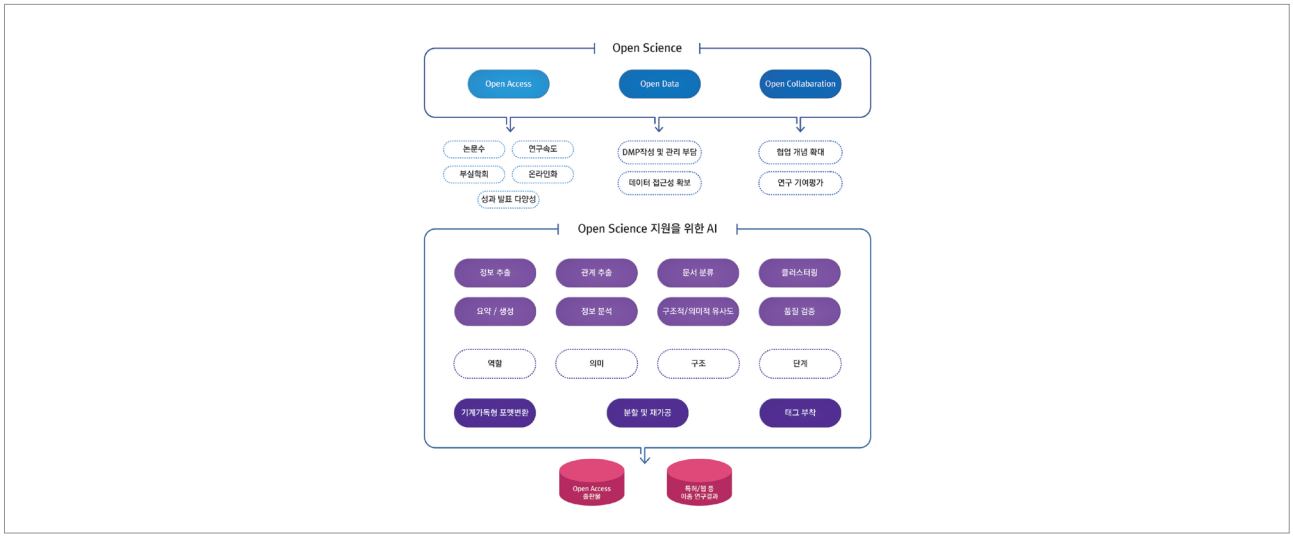
19) 위키피디아 open science, https://en.wikipedia.org/wiki/Open_science
20) 네이버 사전: 디지털 플랫폼 - 정보·통신 온라인에서 생산·소비·유통이 이루어지는 장. 생산자와 소비자가 유기적으로 유통할 수 있으므로, 업계에서는 이것을 점유하기 위한 경쟁이 치열하게 벌어지기도 함
21) 전체의 문제를 서로 독립적인 모듈로 분해하고 각각의 모듈에 대해 다수의 기여자가 동시에 참여할 수 있는 구조를 가지고 있음. 이를 통해 충돌이나 재작업의 위험을 낮추고 오류 발생 시에도 빠르게 원인 분석과 해결이 가능함

- 출판된 논문들의 경우도 연구 구조별로 분석함으로써 부분별로 재사용하거나 접근할 수 있도록 가공하고 구축하는 것이 필요함

오픈 사이언스 활성화를 위한 AI 기술 개발 방향

- **(메타에서 원문) 기존 학술연구 결과에 대한 접근은 키워드를 이용하여 메타를 검색하는 방향으로 진행됨**
  - 원문 분석을 통해 논문에 대한 다양한 관점(문제, 데이터, 방법, 결과 등)을 제공하고 이 관점을 통해 최근 연구부터 영향력 있는 연구까지 접근할 수 있어야 함
  - PDF 기반 종래 논문을 기계 가독적인 형태로 변환하고 내용을 구조적, 의미적 관점에서 분석하는 AI 기반 접근을 통해 원문의 접근성 향상이 필요함
- **(논문에서 연결기반) 기존 출판된 논문 위주로의 연구 결과 접근에서 사전 출판이나 웹문서를 통해 선공개되는 방향으로 전환됨**
  - 정보 접근성을 위해 논문 위주의 정보서비스가 다양한 출처의 자료들을 식별된 속성 기반으로 연결하여 서비스가 제공되어야 함
  - 기존 식별체계 이외 내용 중심으로 이종 정보 간 연결망 구축을 위한 식별 및 태깅 기술 개발이 요구됨
- **(데이터와 논문의 연결) DMP 등 데이터 관련 정보들과 학술논문과의 연결이 필요함**
  - 수치나 이미지 형태의 데이터에 대한 접근성 확보가 필요함
- **(기존 연구 모듈별 라이브러리 구축) 연구 단계의 구조적, 의미적 모듈화 및 연결을 통해 참여자들의 기여 유도 기반 마련이 필요함**
  - 기존 연구들을 모듈화하고 각 단계에 대응하도록 분할, 가공할 수 있는 기술이 필요함

<그림 9> 오픈사이언스 활성을 위한 AI 기술 개요



참고문헌

- 과기정통부 (2021), 「신뢰할 수 있는 인공지능 실현전략」, 2021.5.
- 관계부처 합동 (2019), 「인공지능 국가전략」, 2019.12.
- 김소영 외 (2021), 「사회를 위한 보건의료 분야 인공지능 활용 가이드」, 2021.8.
- 박성민, 박성배, 채승병, 김영도, 김지환 (2013), 「기업의 新경쟁력, 빅데이터 큐레이션」.
- 서봉원 (2016), 「콘텐츠 추천 알고리즘의 진화」.
- 신용우, 정준화 (2021), 「'이루다'를 통해 살펴본 인공지능 활용의 쟁점과 과제」, 국회입법조사처, 이슈와 논점, 제1799호.
- 신은정, 안형준 외 5인 (2017), 「오픈사이언스정책의 도입 및 추진 방안」, 과학기술정책연구원, 2017.12.
- 우상근, 김예지 (2021), 「주목받는 인공지능(AI) 9대 핵심 기술 분석 및 주요 시사점」, 한국지능정보사회진흥원, 2021.1.
- 유네스코한국위원회 (2020) 유네스코 오픈사이언스 권고를 향하여 –오픈사이언스 권고마련의 배경과 경과, 향후 전망.
- 이혜림 (2020), 「디지털 큐레이션 가이드라인과 체크리스트」, 한국과학기술정보연구원.
- 조병호 (2013), 「디지털 큐레이션 서비스 동향」, 정보통신산업진흥원 주간기술동향.
- 한국정보화진흥원 (2018). 「지역경제, 공유경제로 풀다」, Hot Issue Report 2018-1.
- Su, X., & Khoshgoftaar, T. M. (2009). "A Survey of Collaborative Filtering Techniques", Advances in Artificial Intelligence, 2009(12).
- Vaswani, Ashish, et al (2017). "Attention is all you need" Advances in neural information processing systems. pp.5998-6008.

저 자

이 경 하
KISTI 국가과학기술데이터본부
국가과학기술데이터본부전략팀
선임연구원
T. 042-869-1642
E. kyongha@kisti.re.kr

설 재 욱
KISTI 국가과학기술데이터본부
디지털큐레이션센터 선임기술원
T. 042-869-1742
E. wodnr754@kisti.re.kr

이 종 원
KISTI 국가과학기술데이터본부
NTIS센터 선임연구원
T. 042-869-0746
E. jwon1991@kisti.re.kr

선 충 녕
KISTI 국가과학기술데이터본부
융합서비스센터 선임연구원
T. 02-3299-6221
E. wilowisp@kisti.re.kr

KISTI ISSUE BRIEF 제38호

발 행 일 2021. 12. 22.

발 행 인 김재수

편 집 위 원 조민수, 최희석, 이준, 정한민, 함재균, 이준영, 이상환, 정도범

발 행 처 34141 대전광역시 유성구 대학로 245
한국과학기술정보연구원 정책연구실
https://www.kisti.re.kr

I S S N 2635-5728