Clarivate™

Web of Science Data Citation Index (WoS DCI)

Clarivate | 2024.

Clarivate™

1. Research Data, 그 필요성과 현실

연구 데이터 (Research Data)란?

과학적 발견의 기반이자, 지식 생성 과정의 필수적인 자원

연구 데이터

과학 연구를 수행하는 과정에서
수집된 모든 형태의 정보

수집 과정

1차 데이터 수집: 실험, 관찰, 설문조사, 인터뷰, 시뮬레이션 등
2차 데이터 수집: 문헌 조사, 공공 데이터셋, 온라인 DB, Web search 등
공동 연구 및 협업을 통해 수집

데이터 형태

숫자 데이터 (통계값), 텍스트 (인터뷰 응답), 이미지 (사진 자료),
비디오 (의료 영상), 오디오 (공연 녹음), 코딩 된 데이터 등

구조화 정도

정형 데이터
비정형 데이터

예시

생물학 연구: 유전자 서열 데이터
심리학 연구: 설문조사 결과
사회학 연구: 인구 통계 자료
환경학 연구: 기후 변화 데이터
경제학 연구: 시장/소비자 행동 데이터 등

연구 데이터 (Research Data)의 필요성

연구 데이터의 재사용은 연구 자원의 접근성과 효율성을 높이는 데 큰 기여

연구 데이터

과학 연구를 수행하는 과정에서
수집된 모든 형태의 정보

1

재현 가능성

- 연구 결과 검증을 통한 신뢰성 담보
- 안정적인 후속 연구 진행을 위한 기반

2

데이터 재활용

- 과거의 데이터를 활용해 새로운 Insight 도출
- 다양한 데이터를 조합해 숨겨진 패턴과 관계를 발견

3

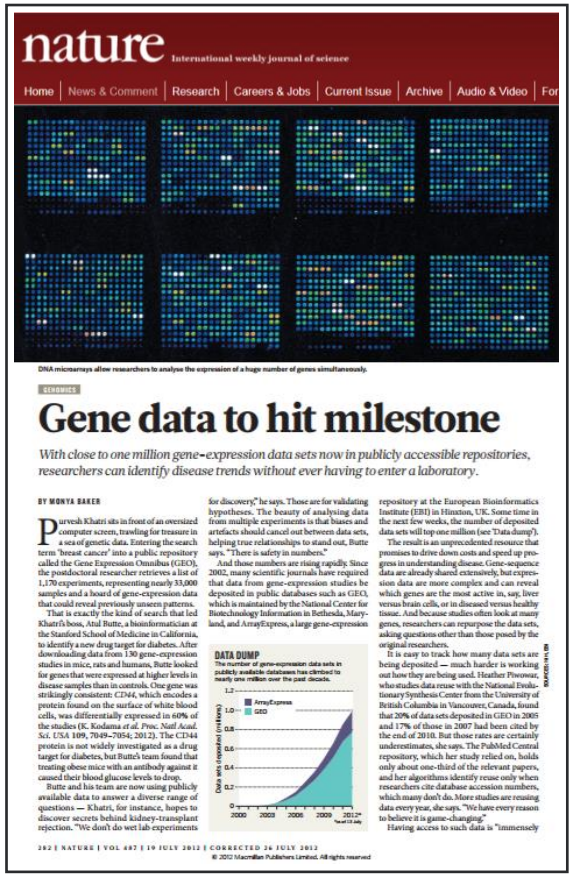
연구 네트워킹

- 유사 연구의 존재와 진행 상황 확인
- 연구 협력의 기회 발굴

연구 데이터 (Research Data)의 성공적인 활용 예시

다양한 방식으로 존재하는 데이터를 면밀하게 분석하는 것만으로 혁신적인 연구가 가능함

한 예로, 과거 Stanford School of Medicine에서는 공공 레파지토리인 Gene Expression Omnibus (GEO) 등에 존재하는 연구 데이터를 이용해 당뇨병을 위한 약제 표적을 확인하는 성과를 거둠



nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For

DNA microarrays allow researchers to analyse the expression of a huge number of genes simultaneously.

GENOMICS

Gene data to hit milestone

With close to one million gene-expression data sets now in publicly accessible repositories, researchers can identify disease trends without ever having to enter a laboratory.

BY MONYA BAKER

Purvesh Khatri sits in front of an oversized computer screen, trawling for treasure in a sea of genetic data. Entering the search term 'breast cancer' into a public repository called the Gene Expression Omnibus (GEO), the postdoctoral researcher retrieves a list of 1,170 experiments, representing nearly 33,000 samples and a hoard of gene-expression data that could reveal previously unseen patterns.

That is exactly the kind of search that led Khatri's boss, Atul Butte, a bioinformatician at the Stanford School of Medicine in California, to identify a new drug target for diabetes. After downloading data from 130 gene-expression studies in mice, rats and humans, Butte looked for genes that were expressed at higher levels in disease samples than in controls. One gene was strikingly consistent: *CD44*, which encodes a protein found on the surface of white blood cells, was differentially expressed in 60% of the studies (K. Kodama *et al. Proc. Natl Acad. Sci. USA* 109, 7049–7054; 2012). The CD44 protein is not widely investigated as a drug target for diabetes, but Butte's team found that treating obese mice with an antibody against it caused their blood glucose levels to drop.

Butte and his team are now using publicly available data to answer a diverse range of questions — Khatri, for instance, hopes to discover secrets behind kidney-transplant rejection. "We don't do wet lab experiments for discovery," he says. Those are for validating hypotheses. The beauty of analysing data from multiple experiments is that biases and artefacts should cancel out between data sets, helping true relationships to stand out, Butte says. "There is safety in numbers."

And those numbers are rising rapidly. Since 2002, many scientific journals have required that data from gene-expression studies be deposited in public databases such as GEO, which is maintained by the National Center for Biotechnology Information in Bethesda, Maryland, and ArrayExpress, a large gene-expression repository at the European Bioinformatics Institute (EBI) in Hinxton, UK. Some time in the next few weeks, the number of deposited data sets will top one million (see 'Data dump').

The result is an unprecedented resource that promises to drive down costs and speed up progress in understanding disease. Gene-sequence data are already shared extensively, but expression data are more complex and can reveal which genes are the most active in, say, liver versus brain cells, or in diseased versus healthy tissue. And because studies often look at many genes, researchers can repurpose the data sets, asking questions other than those posed by the original researchers.

It is easy to track how many data sets are being deposited — much harder is working out how they are being used. Heather Piwowar, who studies data reuse with the National Evolutionary Synthesis Center from the University of British Columbia in Vancouver, Canada, found that 20% of data sets deposited in GEO in 2005 and 17% of those in 2007 had been cited by the end of 2010. But those rates are certainly underestimates, she says. The PubMed Central repository, which her study relied on, holds only about one-third of the relevant papers, and her algorithms identify reuse only when researchers cite database accession numbers, which many don't do. More studies are reusing data every year, she says. "We have every reason to believe it is game-changing."

Having access to such data is "immensely

DATA DUMP

The number of gene-expression data sets in publicly available databases has climbed to nearly one million over the past decade.

282 | NATURE | VOL 487 | 19 JULY 2012 | CORRECTED 26 JULY 2012

© 2012 Macmillan Publishers Limited. All rights reserved

…describes how the publicly available Gene Expression Omnibus repository was used by Stanford investigators to lead them to identify a new drug target for diabetes.

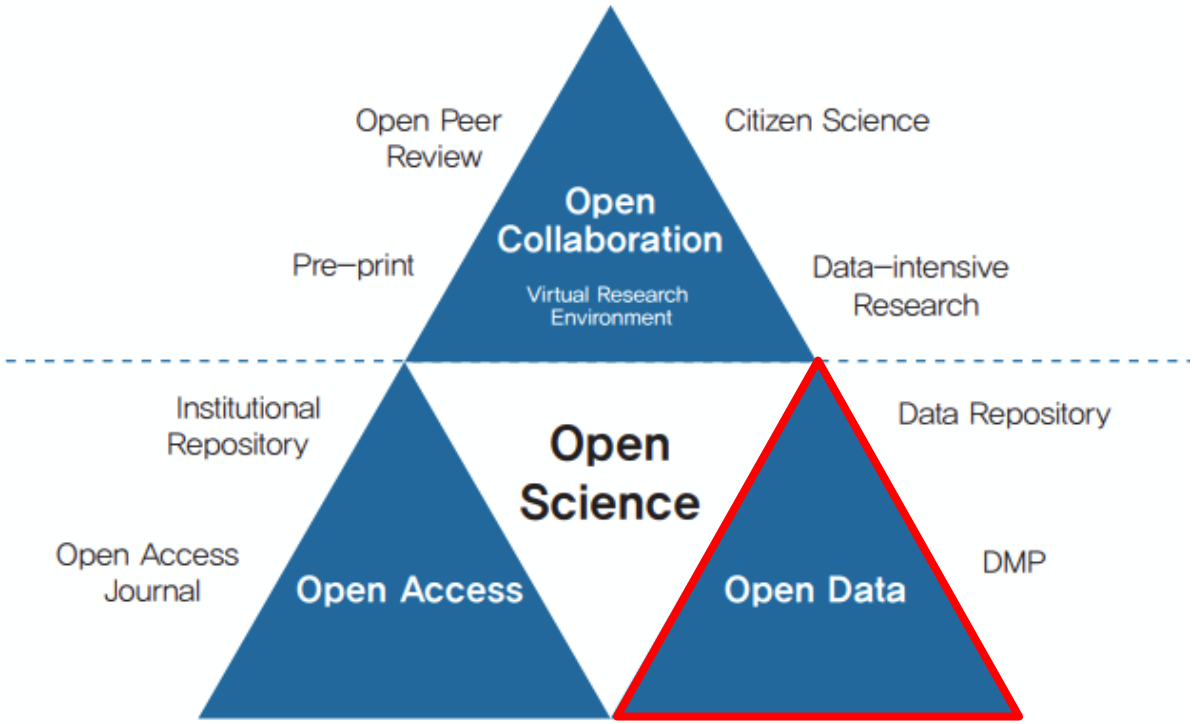
"There is enough info in existing databases to predict disease in samples that algorithms have never seen before,"

* Nature News, Nature Publishing Group, Jul 18, 2012.
Copyright © 2012, Rights Managed by Nature Publishing Group

http://www.nature.com/news/gene-data-to-hit-milestone-1.11019

연구 데이터 (Research Data) 공유 촉진을 위한 정책과 지침들

데이터 공유의 중요성과 당위성이 강조되는 경향과 함께,
연구의 데이터 개방을 요구하는 정책과 지침들도 지속적으로 강화되는 추세



Why is data important in the research lifecycle?

Accountability

Funding bodies such as the NIH, NSF, and ERC promote, and sometimes require, open data as a critical component necessary to translate research into applications and services with immediate social impact.

Transparency

Sharing data enables scholars to validate others' findings, ensuring that academic research can be trusted as the reliable basis for future basic and applied research, as well as public policy.

Credit

Nearly 70% of researchers value a data citation as much as an article citation. 10% value a data citation *more* than an article citation.[5] It is important to track datasets alongside other reserach outputs.

연구 데이터 (Research Data) 공유 촉진을 위한 정책과 지침들

한국 또한 2023년 9월 과학기술정보통신부를 통해 국가연구데이터 관리 및 활용 촉진에 관한 법률 제정안을 입법 예고한 바 있음

국가연구데이터 관리 및 활용 촉진에 관한 법률 제정안 입법예고

공고번호 제2023-890호 | 법령종류 법률 | 입안유형 제정 | 예고기간 2023-09-27~2023-11-06

소관부처 과학기술정보통신부 | 담당부서 과학기술정보분석과 | 전화번호 044-202-6964 | 전자메일 jphong82@korea.kr

⊙과학기술정보통신부공고제2023-0890호

국가연구데이터 관리 및 활용 촉진에 관한 법률 제정안 입법예고를 하는데 있어, 그 이유와 주요내용을 국민에게 미리 알려 이에 대한 의견을 듣기위하여 "행정절차법" 제 41조에 따라 다음과 같이 공고합니다.

2023년 9월 27일
과학기술정보통신부장관

국가연구데이터 관리 및 활용 촉진에 관한 법률 제정안 입법예고

1. 제정이유

새로운 지식·가치 창출의 원천으로서 연구데이터의 가치가 높아짐에 따라 융합·공동연구 촉진을 위한 연구데이터 공유 체계의 필요성이 증대되고 있음. 미국을 비롯한 주요국에서는 일찍이 연구데이터의 체계적인 관리와 활용을 위한 정책을 추진해 왔으며, 과학기술 연구의 디지털화를 통해 연구데이터를 축적하고 인공지능 기술을 적용하여 새로운 성과를 창출하고 있음

최근 들어 우리나라도 공적자금이 투입된 연구데이터의 체계적인 관리와 활용을 위하여 데이터 관리계획의 도입, 국가연구데이터플랫폼의 구축 등 다양한 시도를 하고 있음. 하지만, 법·제도적 인프라의 미비 등으로 인하여 활성화에 한계를 보이고 있고, 현재 우리나라가 직면하고 있는 저출산·고령화, 기술유출 사건·사고의 증가 등을 고려할 때 연구데이터의 관리·활용체계 확립이 시급함

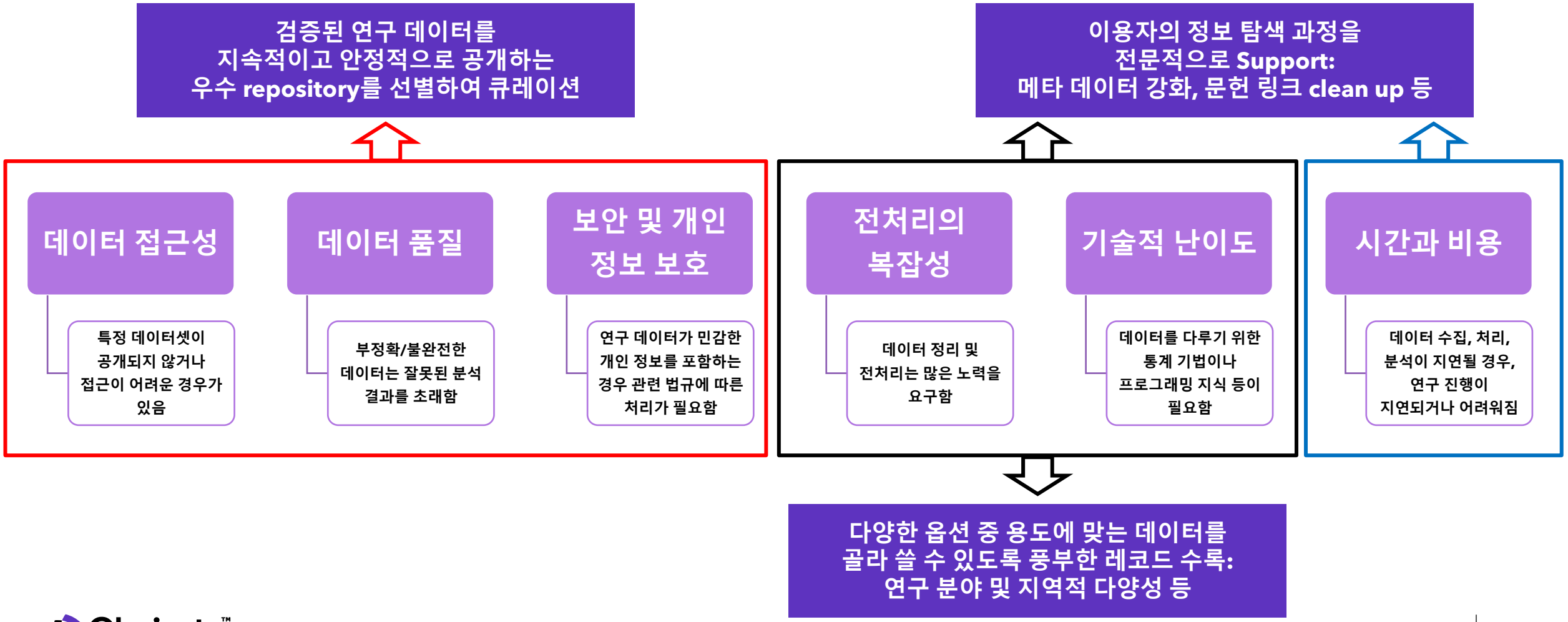
연구 데이터 (Research Data) 활용의 장애물

세계적으로 1700여종이 넘는 방대한 양의 레파지토리들이 연구 관련 정보를 제공하고 있으나,

레파지토리마다 서로 다른 플랫폼과 서로 다른 데이터 구조에 기반하고 있어 데이터의 통합적인 검색과 이용에 어려움이 있음

연구 데이터 (Research Data) 활용의 장애물

연구에 적합한 데이터를 확보하는 것부터 데이터 품질, 보안 및 개인 정보 보호 관련 법률 준수 등 다양한 장애물이 존재

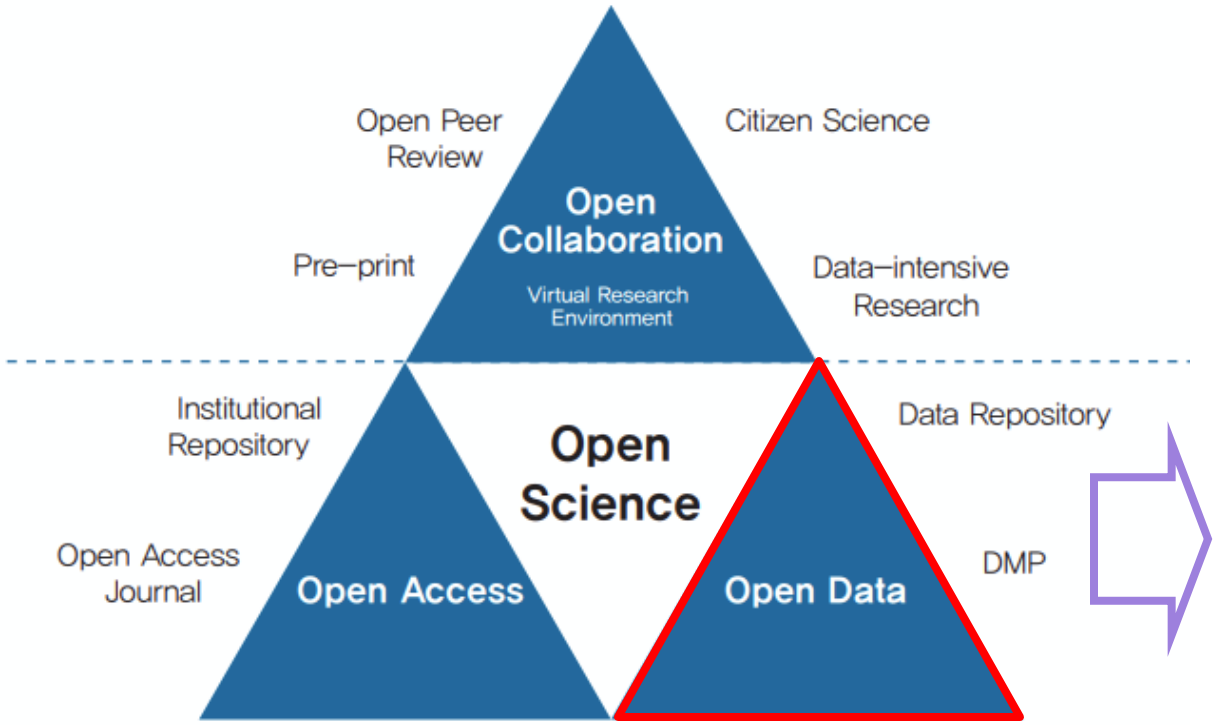


Clarivate™

2. WoS DCI (Data Citation Index)

Open Science와 WoS Data Citation Index (DCI)

WoS DCI는 Open Science 모델의 핵심 축인 Open Data를 지원



DISCOVERY AND WORKFLOW SOLUTIONS | WEB OF SCIENCE PLATFORM

Data Citation Index

Connecting data to the research it informs

453 repositories	**14M+** data sets
1.6M+ data studies	**405K+** software

WoS DCI: Repository Selection Process

클래리베이트 에디토리얼 팀은 1700개가 넘는 전 세계 레포지토리들을 정성/정량적 측면에서 검토

Publishing Standards

- 지속성과 안정성
- Funding statements
- 피어 리뷰
- 자료의 Age
- 메타데이터 및 문헌 링크

Editorial Content

- Emerging topics
- Active fields

International Diversity

- 저자, 에디터, 데이터 생산자, 데이터 등 다양한 요소를 고려

Data Citations

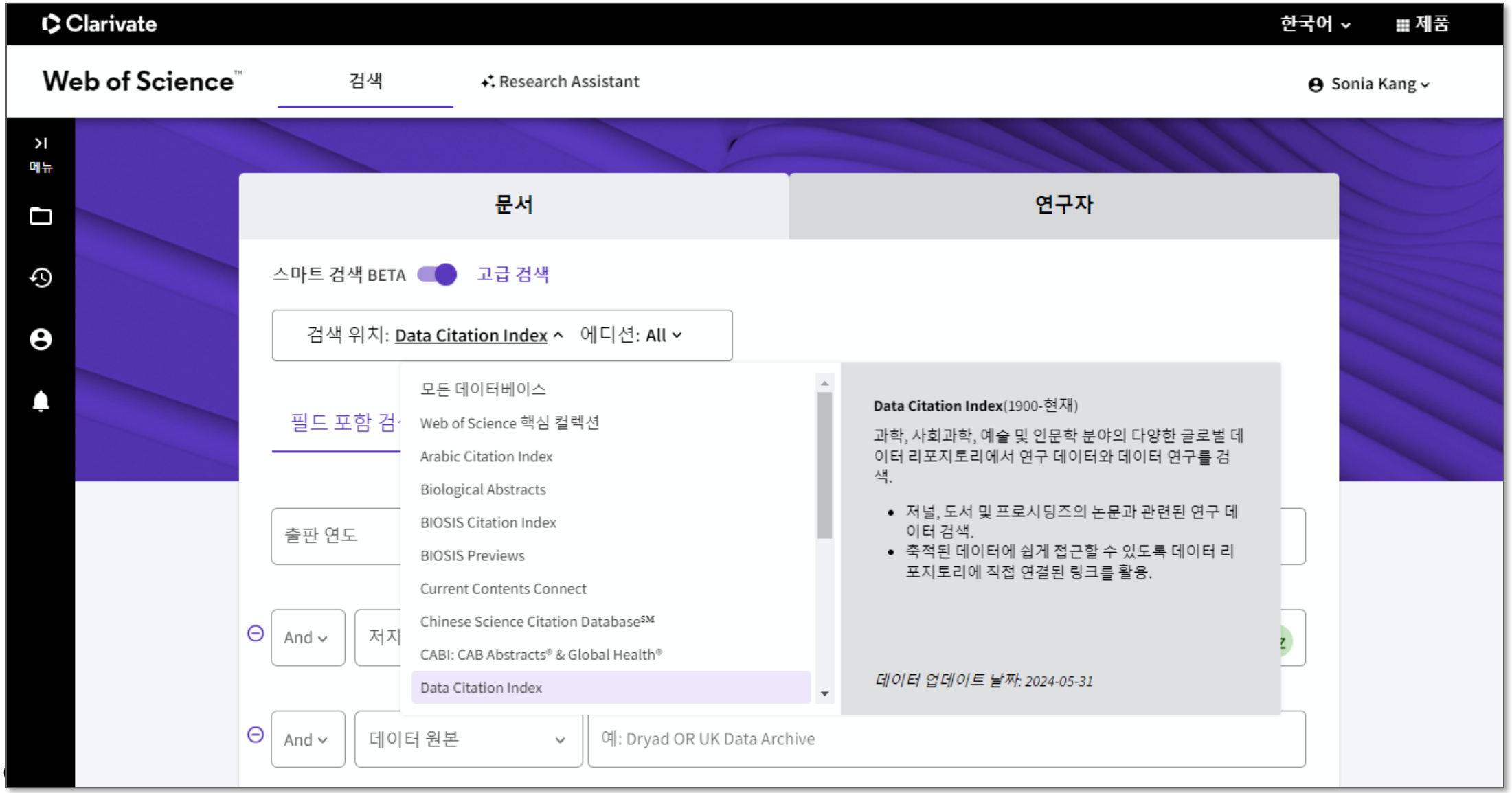
- 데이터를 인용한 참고문헌
- 데이터가 인용한 참고문헌

Repository Selection Process

https://clarivate.com/products/scientific-and-academic-research/research-discovery-and-workflow-solutions/repository-selection-process/

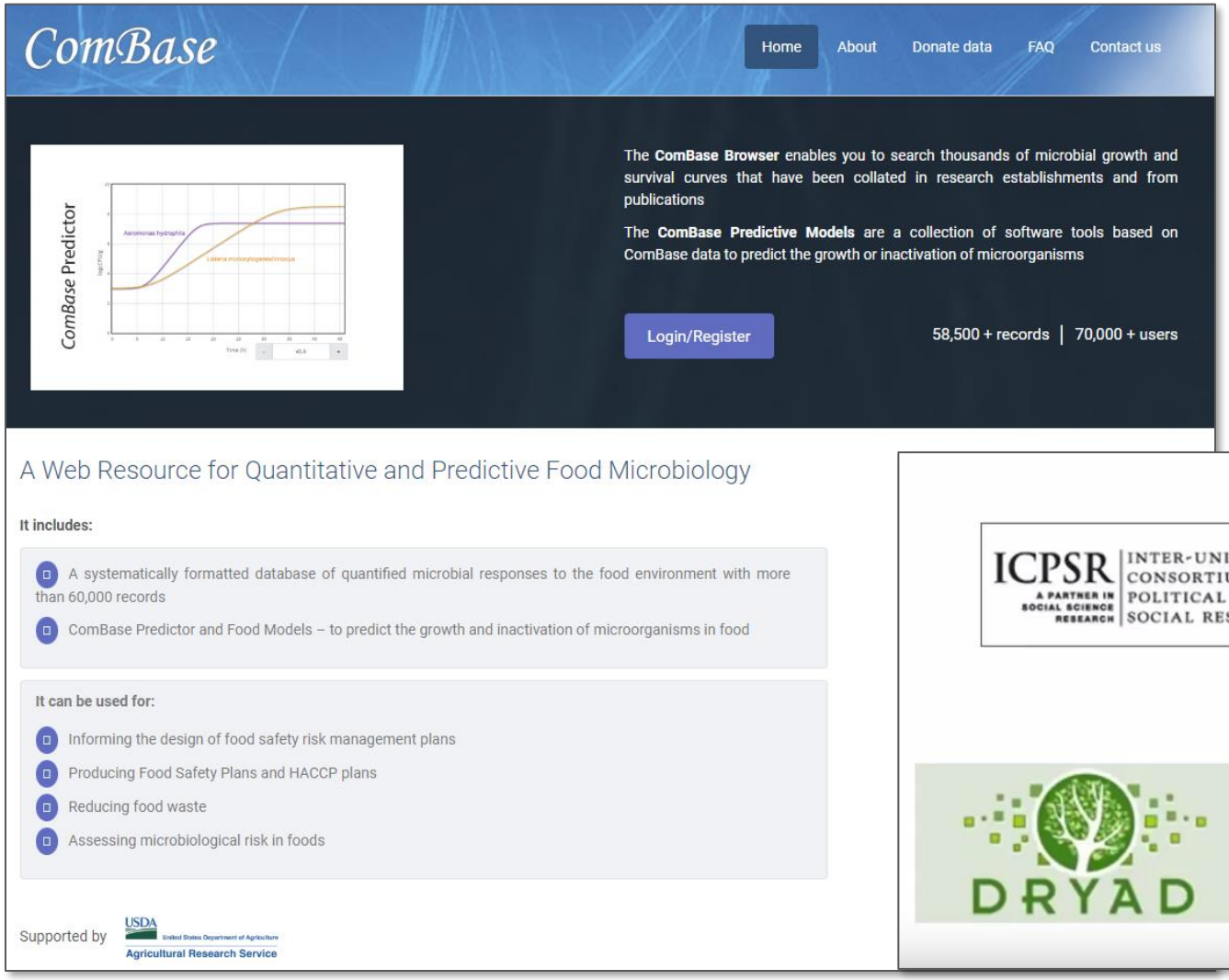
WoS DCI: WoS 전문 기술로 데이터 서비스

전 세계 주요 레포지토리의 데이터를 WoS의 전문 큐레이션과 함께 손쉽게 검색 및 분석



WoS DCI: Selected Data Repositories

에디토리얼 팀이 선별한 450여개 주요 글로벌 레포지토리의 데이터가 매 주 DCI에 업데이트 됨

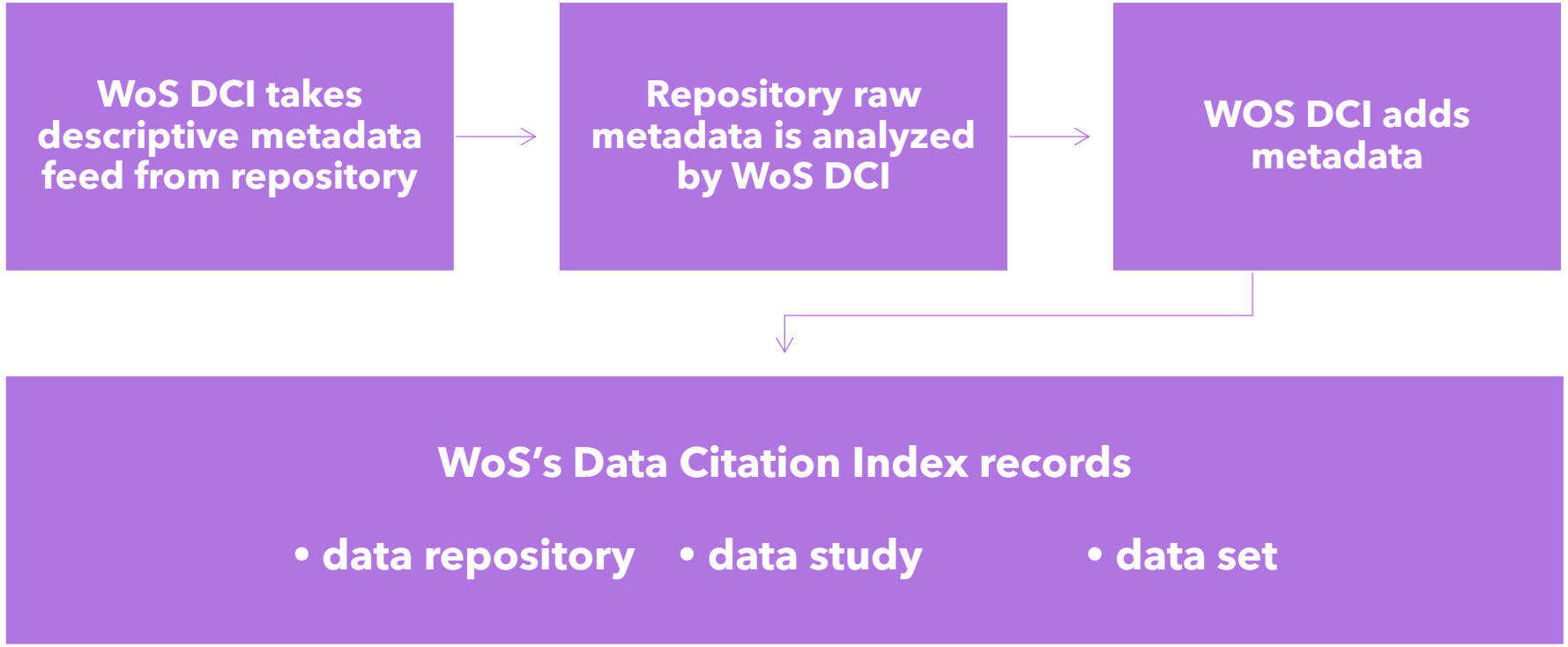


Master Data Repository List

https://clarivate.com/webofsciencegroup/master-data-repository-list/

WoS DCI: Data Restructuring

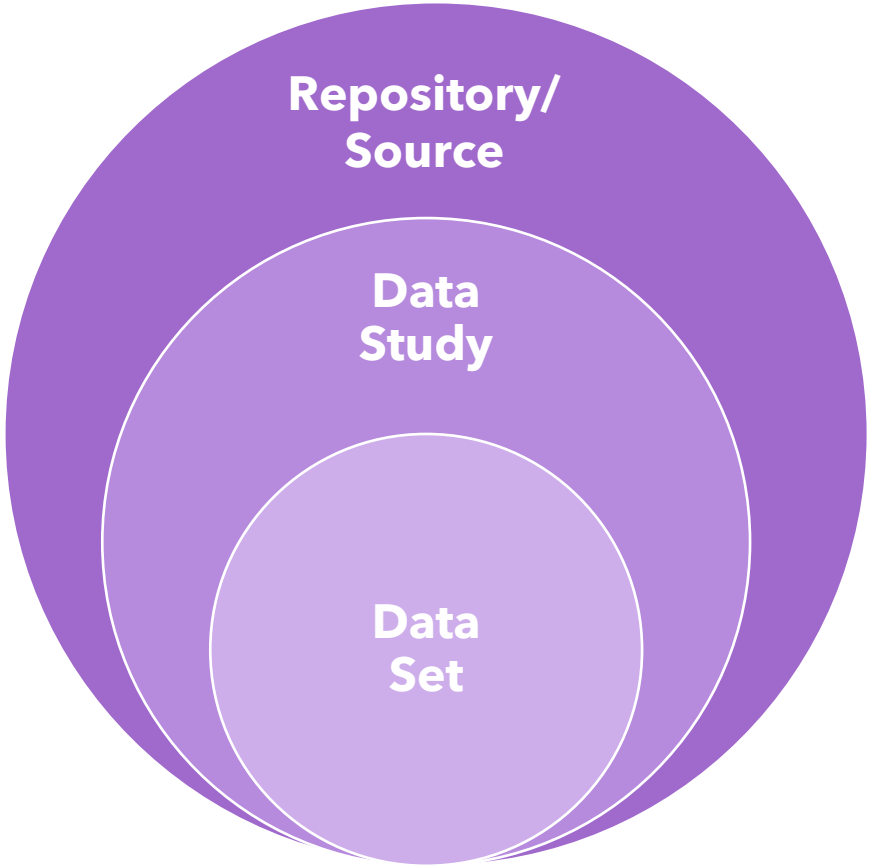
WoS DCI는 레포지토리의 메타데이터를 분석하고 WoS의 정보를 취합하여 데이터를 재구성



Data Citation Index records **link back to the source**,
to the repositories, data studies, and data sets themselves.

WoS DCI: 3 Level Contents Types

WoS DCI는 데이터를 3단계 콘텐츠 유형으로 재구성



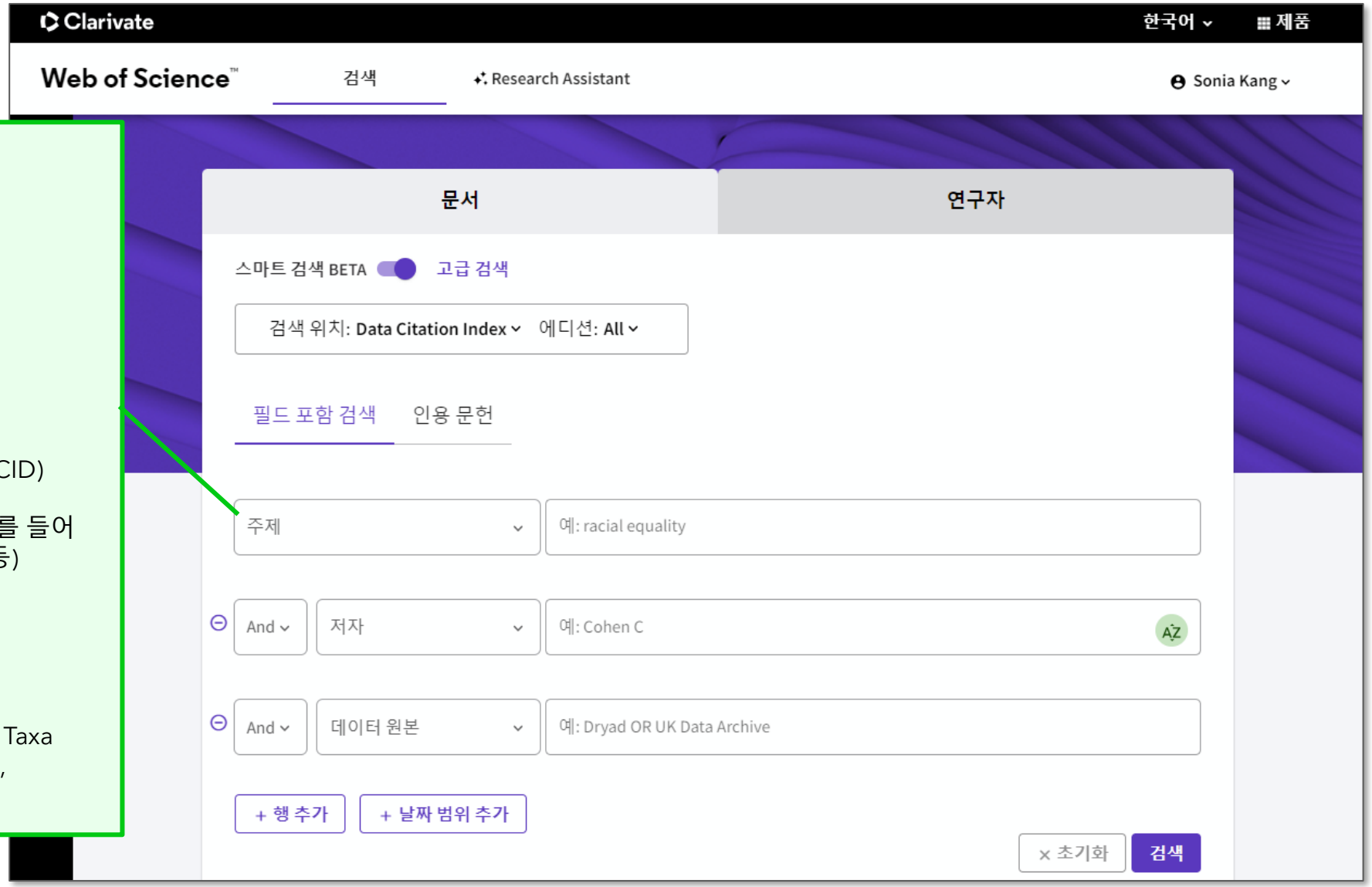
- **Repository/Source**: Comprises data studies, data sets, etc. Stores and provides access to the raw data.
- **Data Study**: Descriptions of studies or experiments with associated data which have been used in the data study.
- **Data Set**: A single or coherent set of data or a data file provided by the repository, as part of a collection, data study or experiment.
  (Data set에는 **Software 정보가 포함됨**: A computer program or package in source code or compiled form, which can be installed on another machine and used to support & analyse research)

WoS DCI: WoS UI를 통해 검색 및 분석 진행

데이터 검색을 위한 전문 필드 활용

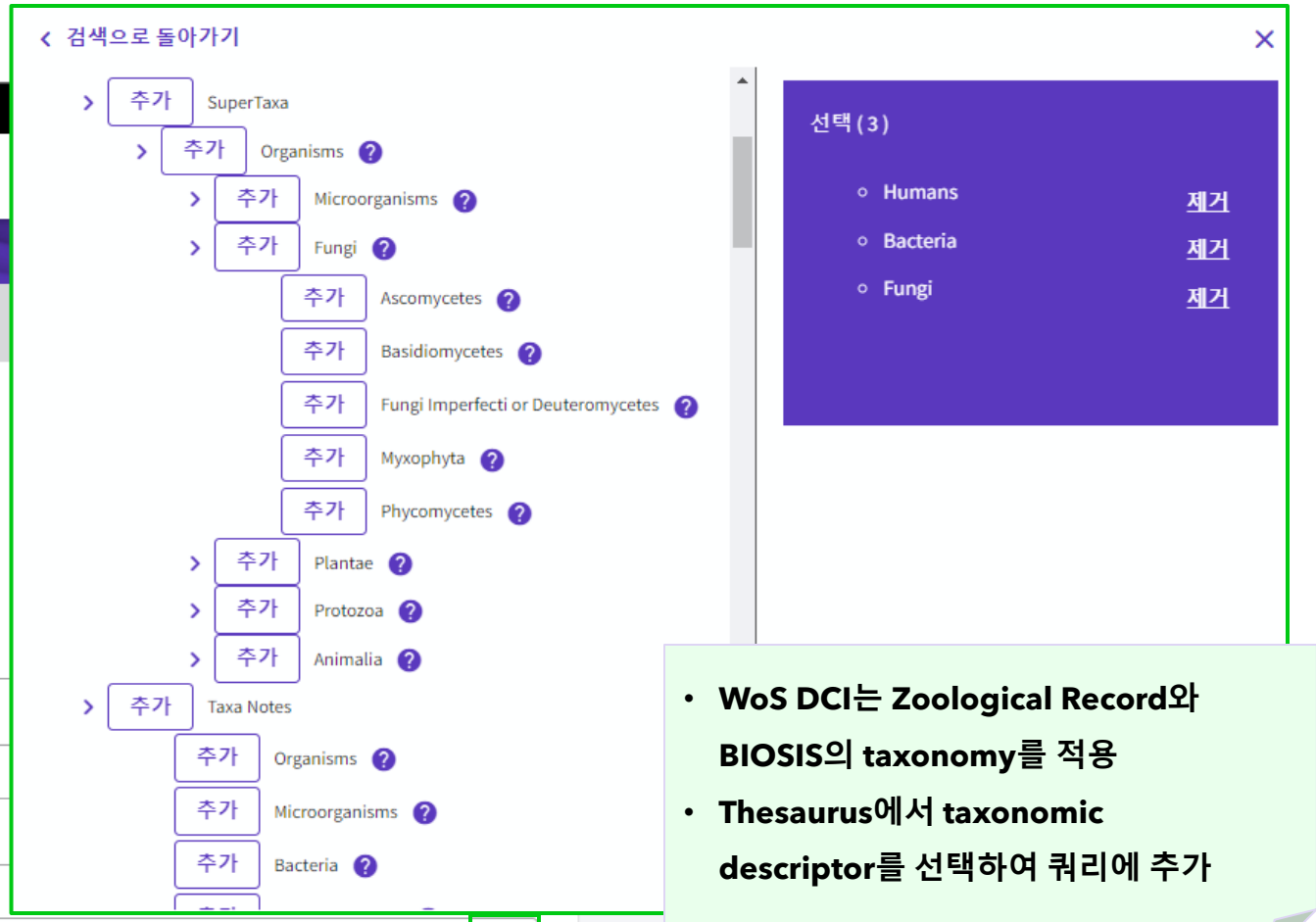
[검색 가능 필드 목록]

- 주제 (제목, 초록, 저자 키워드)
- 제목
- 저자
- 데이터 원본 (레포지토리명)
- 출판 연도
- 초록
- 식별번호
- 주소
- 저자 식별자 (ResearcherID, ORCID)
- 저자 키워드
- 내용 유형 (데이터 분류 유형: 예를 들어 Repository, Dataset, Software 등)
- DOI
- 에디터
- 연구비 지원 텍스트
- 그룹 저자
- 언어
- 주제 설명어 (Taxa Notes, Super Taxa Term을 포함한 텍사노믹 데이터, Web of Science Category 등)



WoS DCI: WoS UI를 통해 검색 및 분석 진행

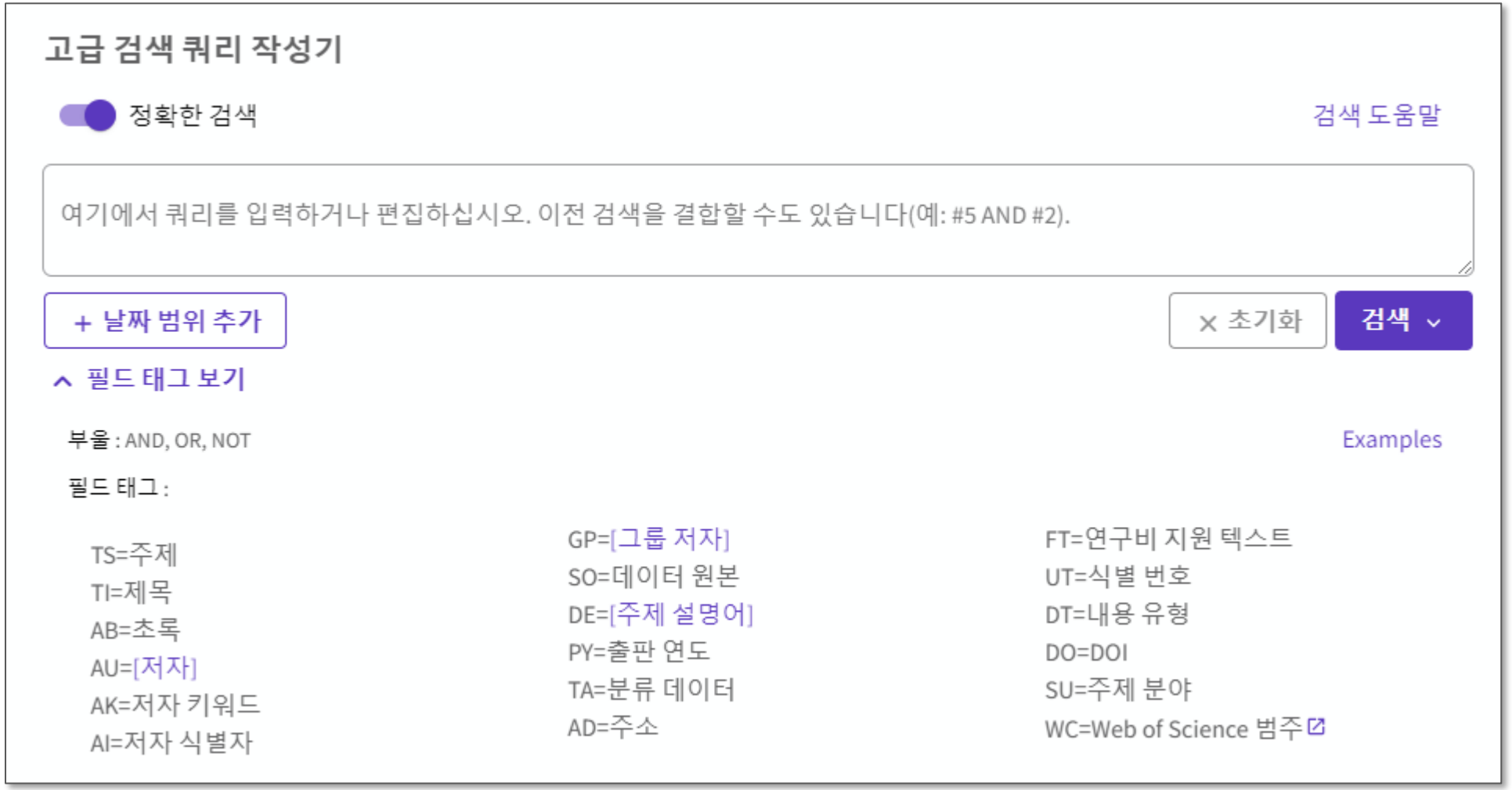
텍사노믹 데이터를 활용한 쿼리 작성



- **WoS DCI는 Zoological Record와 BIOSIS의 taxonomy를 적용**
- **Thesaurus에서 taxonomic descriptor를 선택하여 쿼리에 추가**

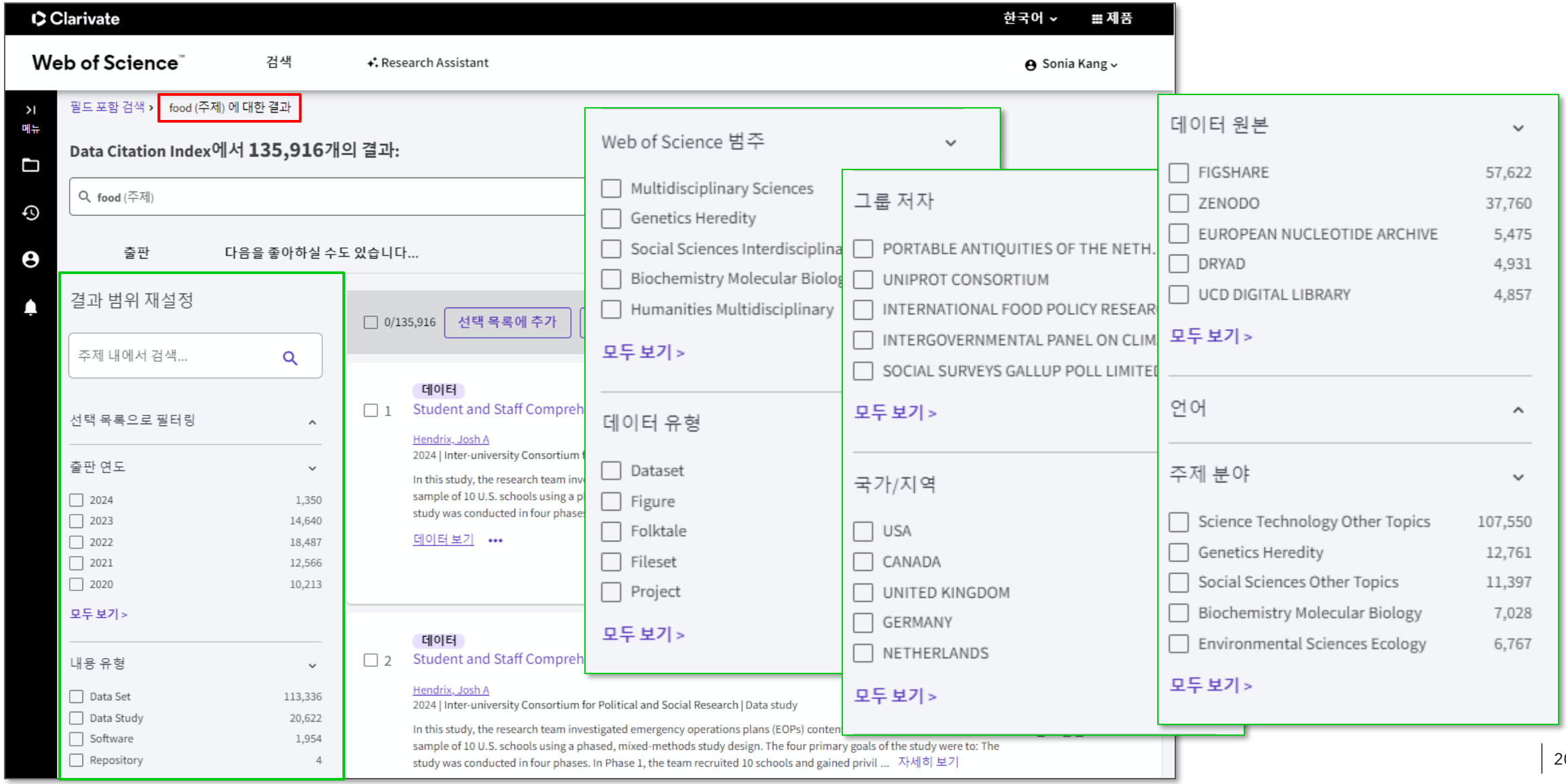
WoS DCI: WoS UI를 통해 검색 및 분석 진행

필드 태그를 사용하면 보다 정교한 검색 쿼리 작성 가능



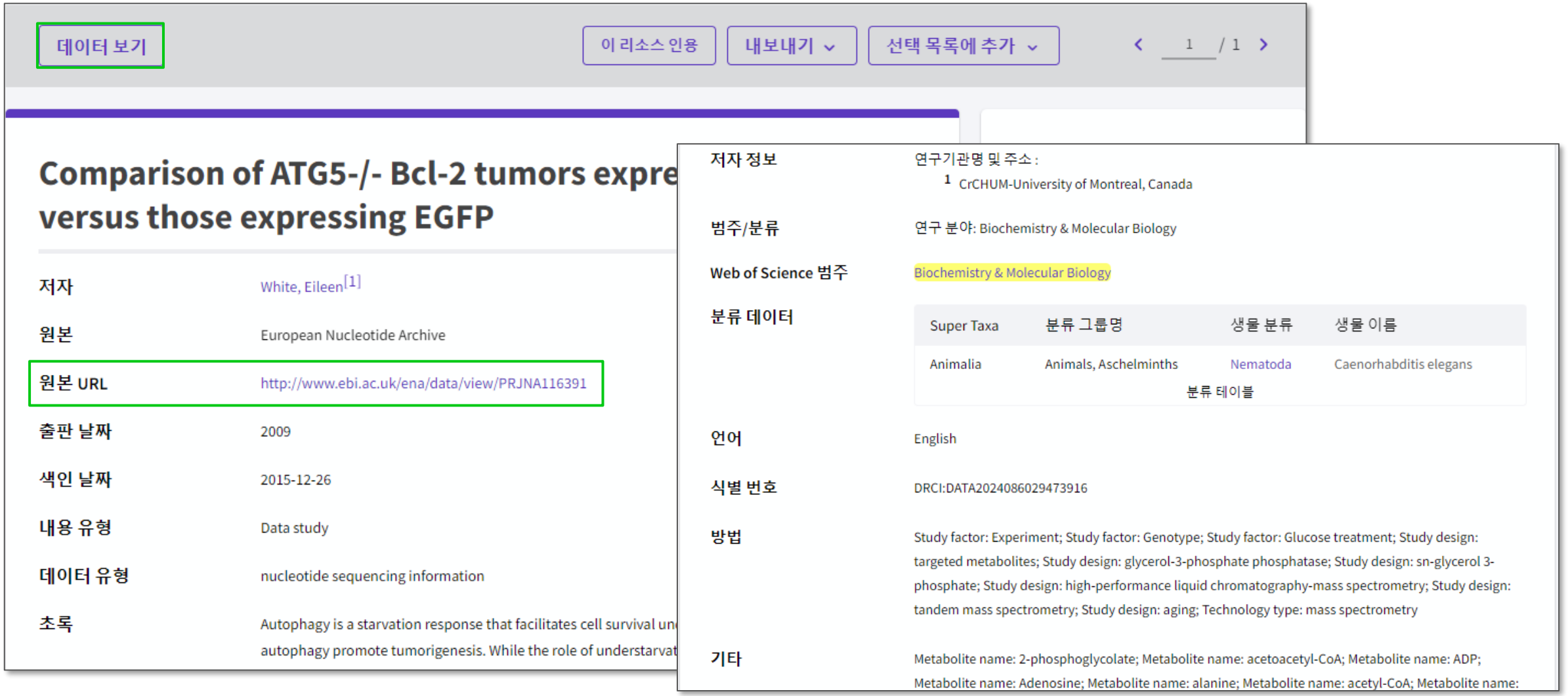
WoS DCI: WoS UI를 통해 검색 및 분석 진행

검색 결과 범위를 재설정할 수 있는 다양한 전문 필터 활용



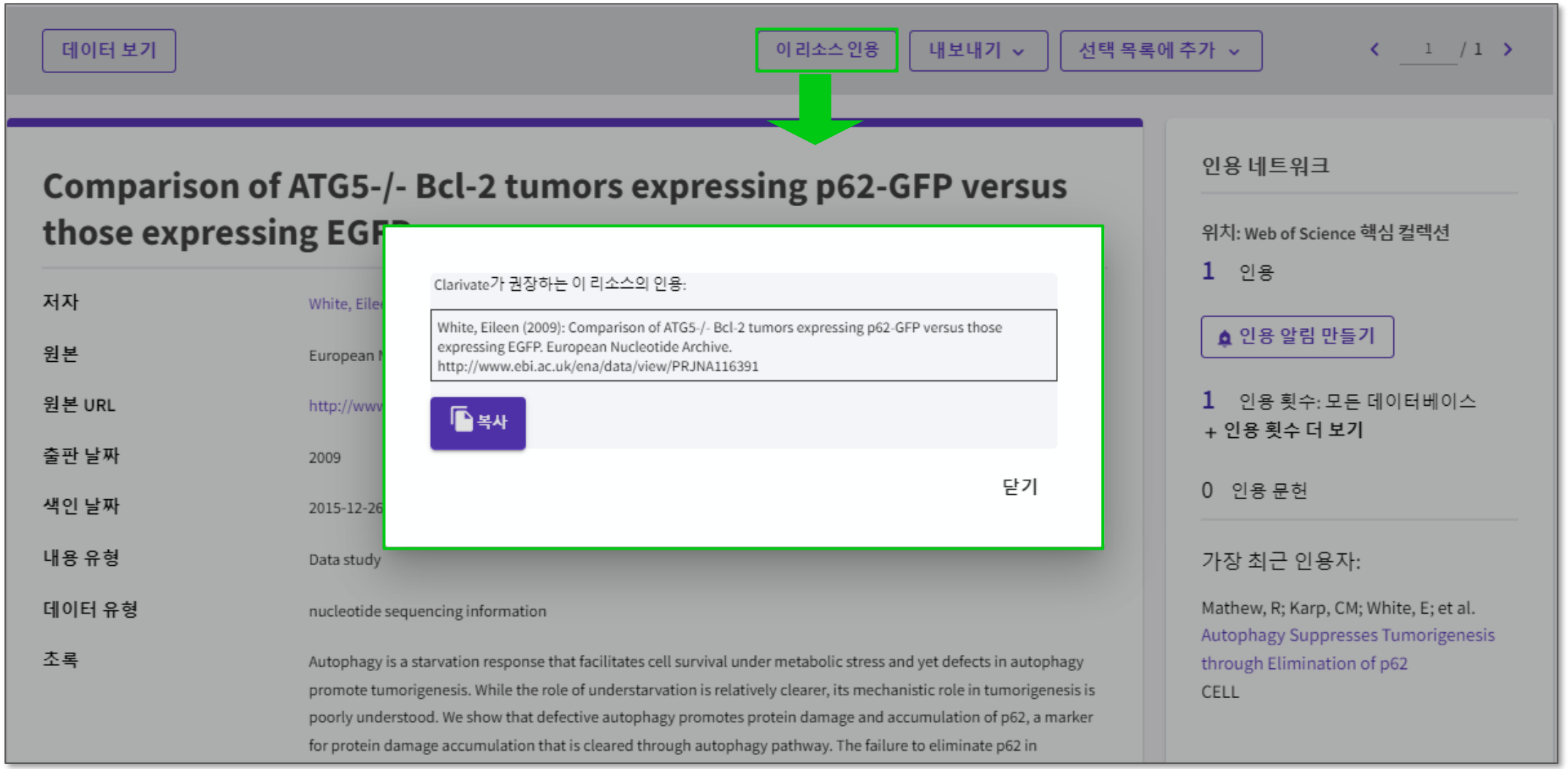
WoS DCI: WoS UI를 통해 검색 및 분석 진행

원본 데이터 바로가기 링크, 데이터 제공자 및 관련 키워드 등의 메타데이터 확인 가능



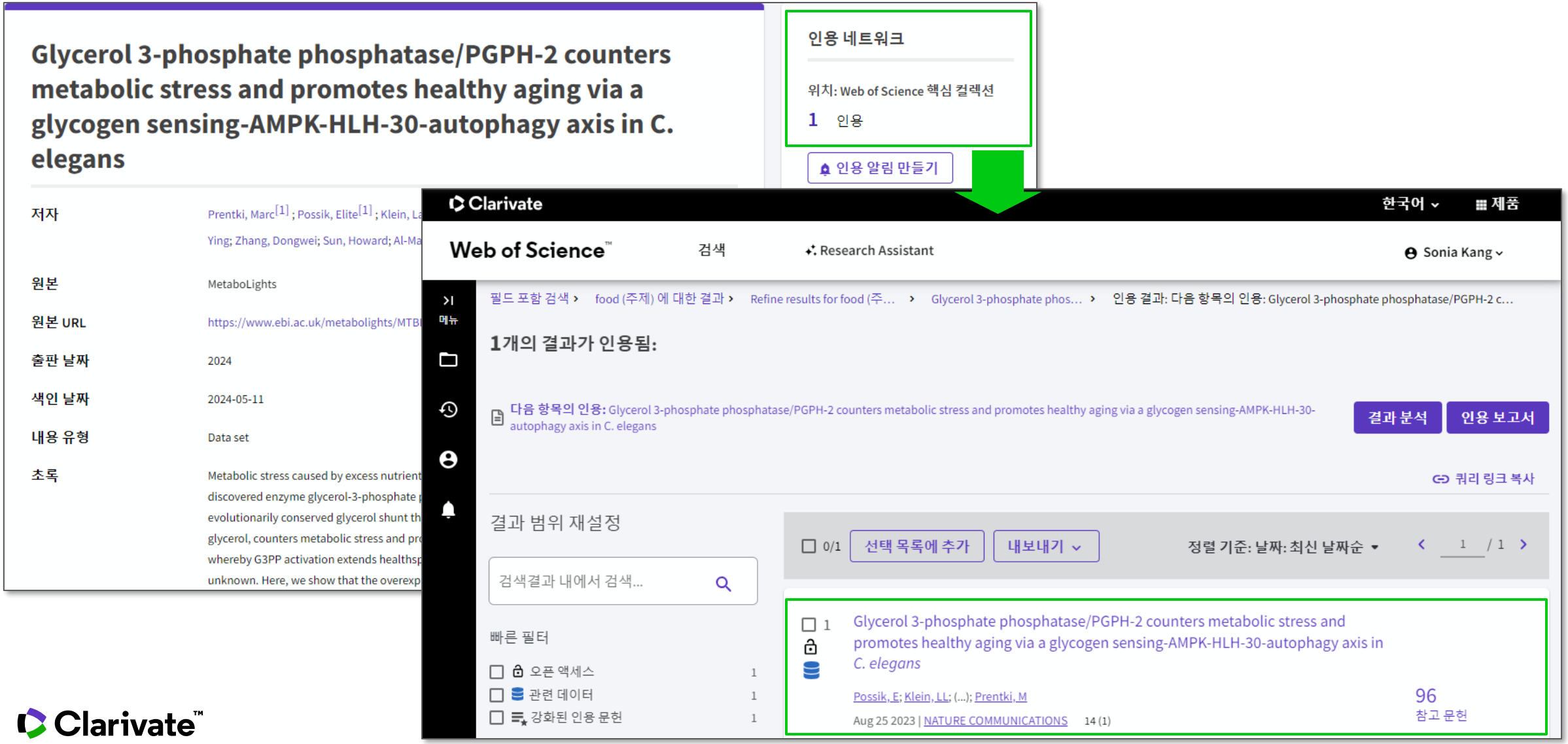
WoS DCI: WoS UI를 통해 검색 및 분석 진행

DataCite의 인용 가이드라인을 기반으로 하는 인용 포맷 제공



WoS DCI: WoS UI를 통해 검색 및 분석 진행

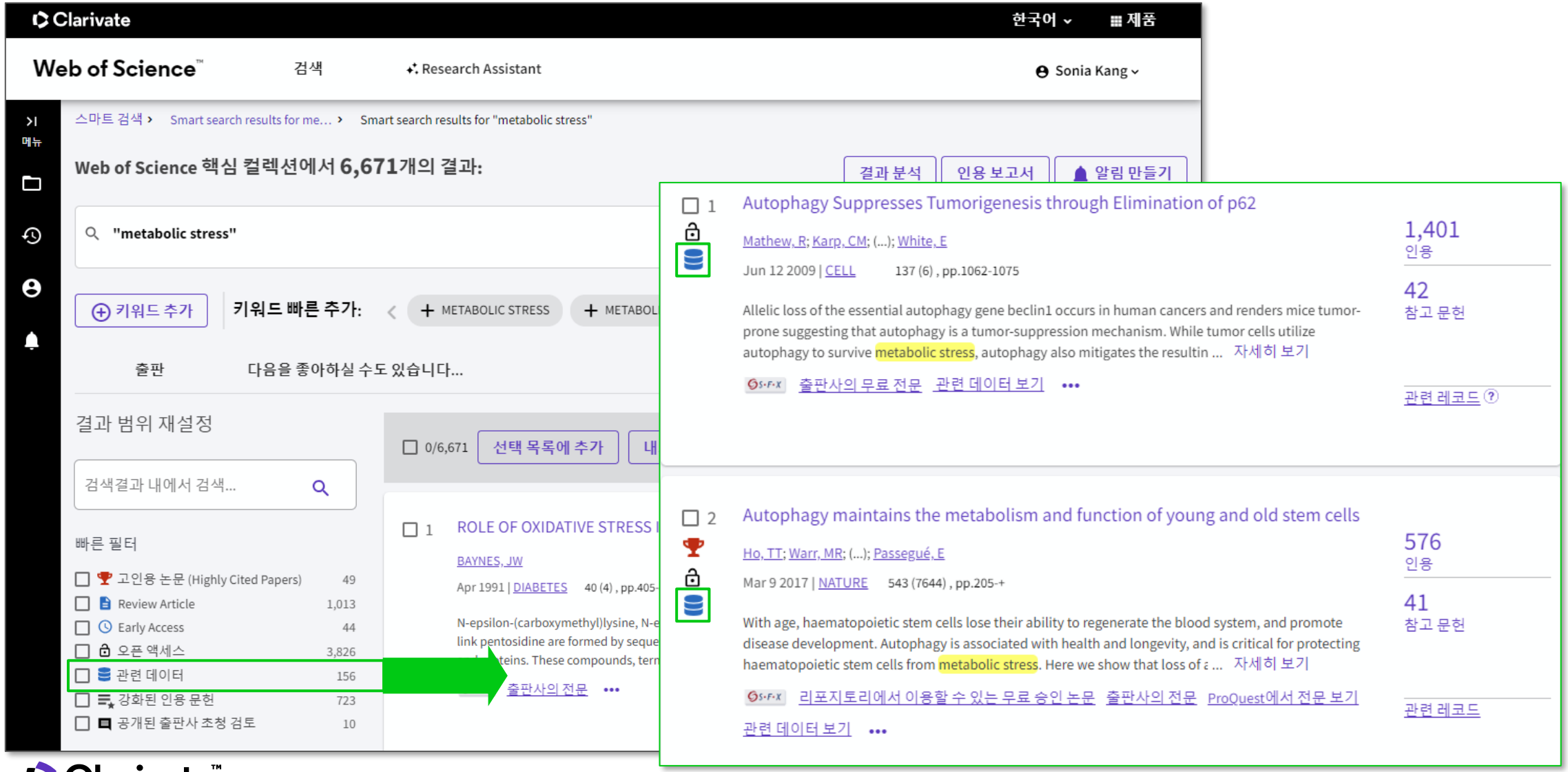
데이터가 WoS 논문에 인용된 경우, 해당 논문을 바로 확인할 수 있음



Clarivate™

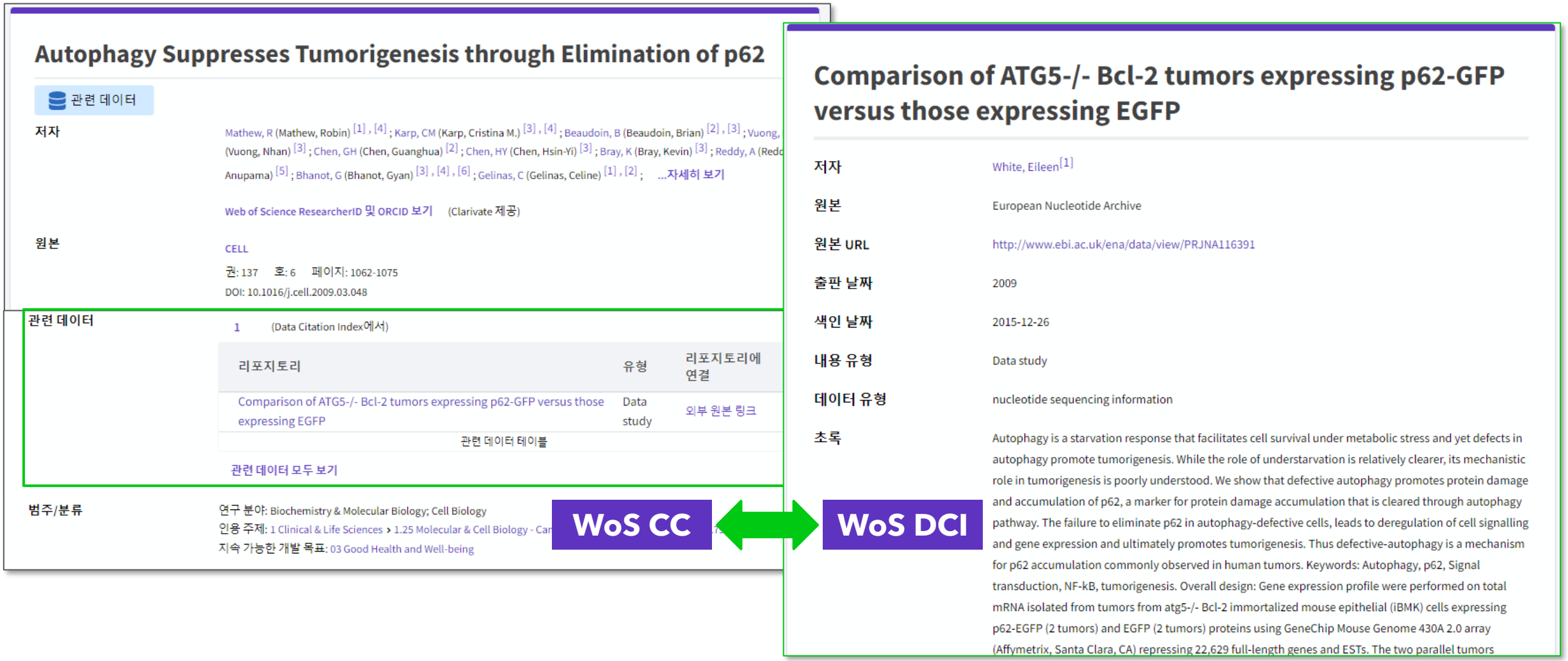
WoS DCI: Access from other WoS Indexes

반대로, WoS 검색 중에는 "관련 데이터" 필터를 통해 WoS DCI record로 이동 가능



WoS DCI: Access from other WoS Indexes

WoS 이용 전반에서 DCI에 수록된 연구 데이터를 확인 및 엑세스 가능

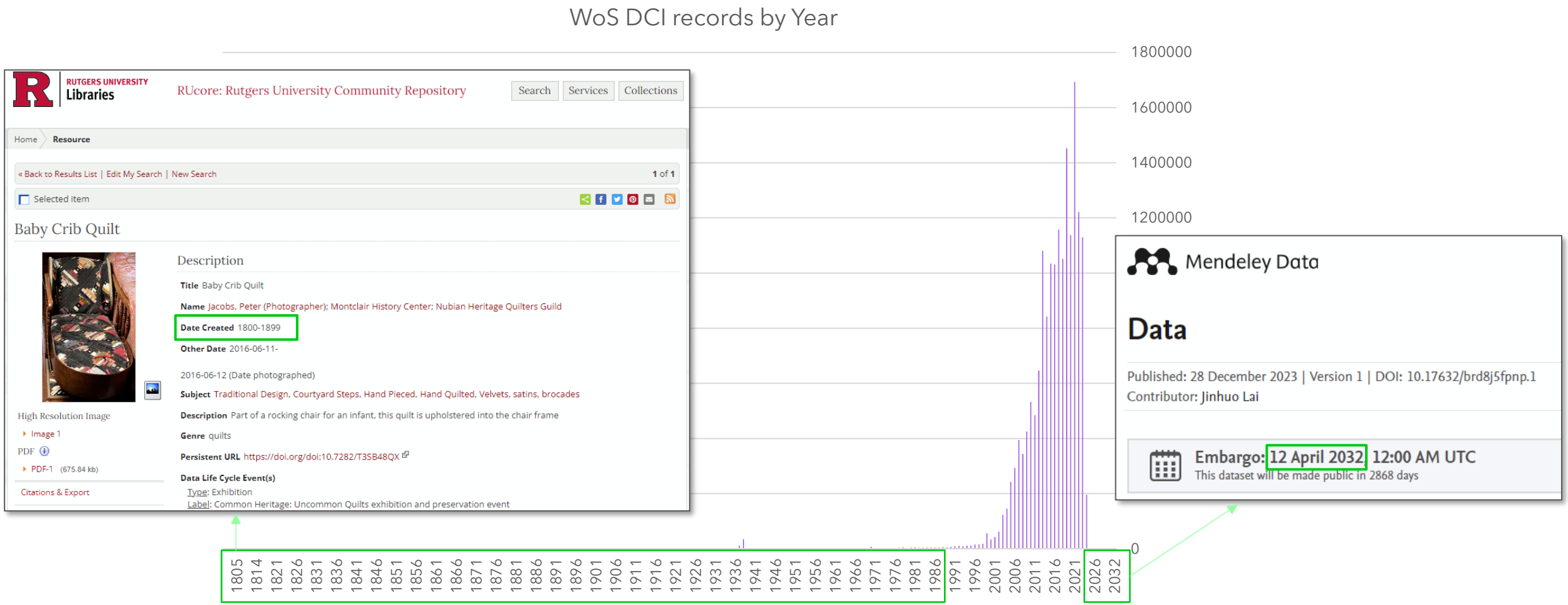


Clarivate™

3. WoS DCI (Data Citation Index)의 데이터 구성

WoS DCI: Records by Year

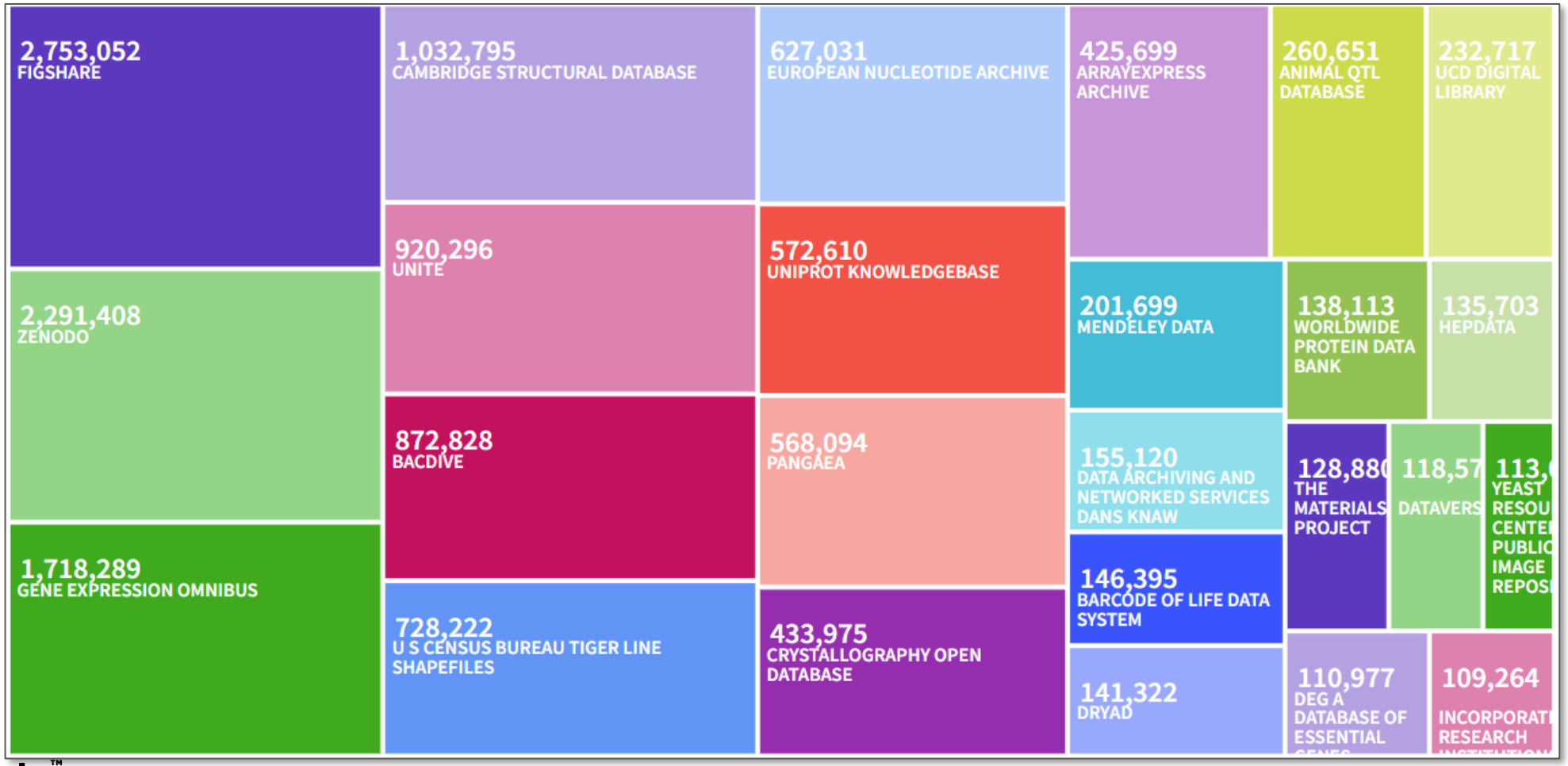
WoS DCI는 1900년 이후의 데이터를 주로 다루지만,
선정된 레포지토리가 보유한 데이터는 연도 하한선 없이 수록하고 있음



WoS DCI: Records by Data Source

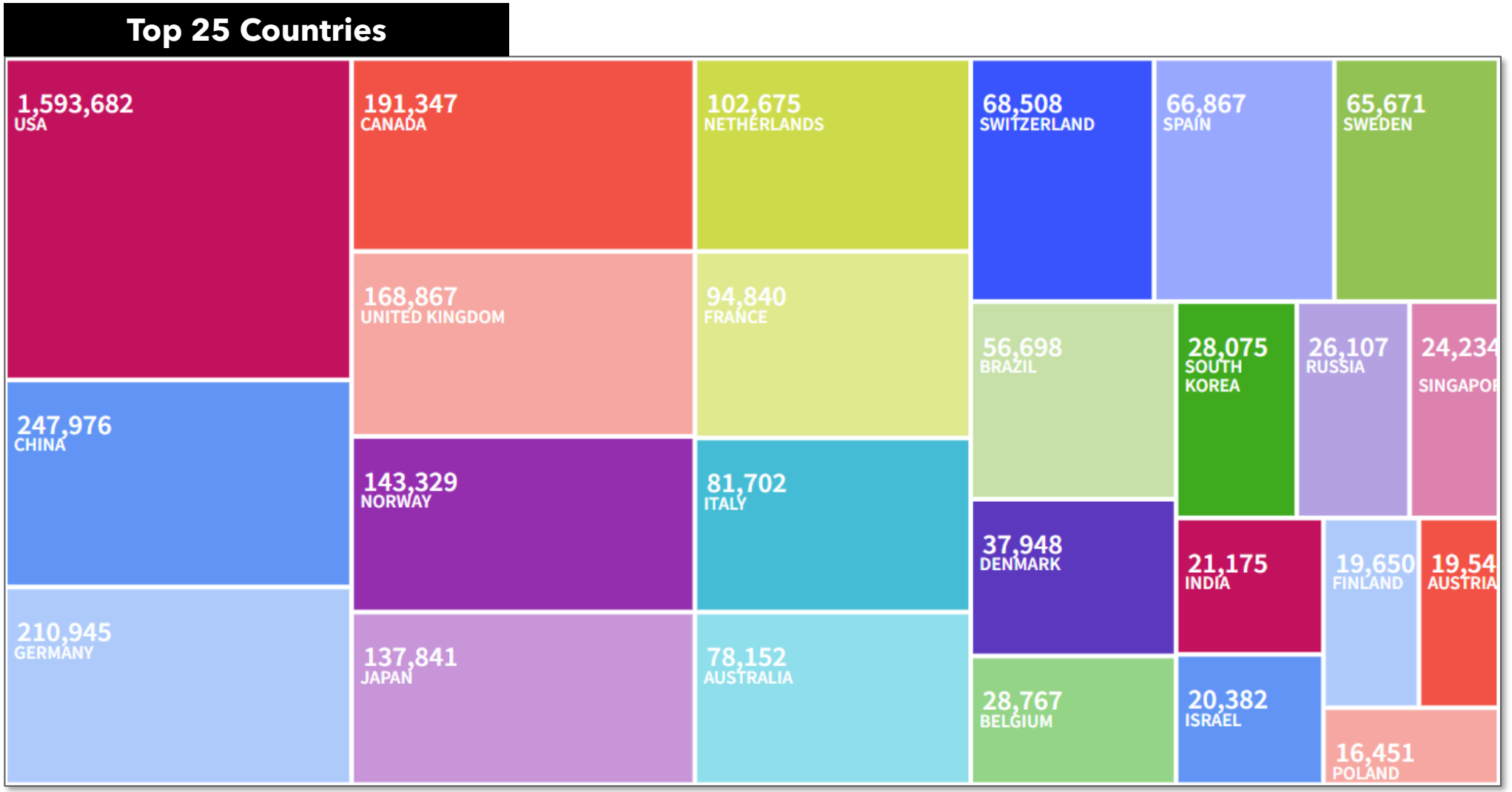
WoS DCI는 450개 이상의 레포지토리 정보를 인덱스하고 있으며,
이 중 95% 이상이 컨텐츠에 대한 무료 액세스를 제공함

Top 25 Data Sources



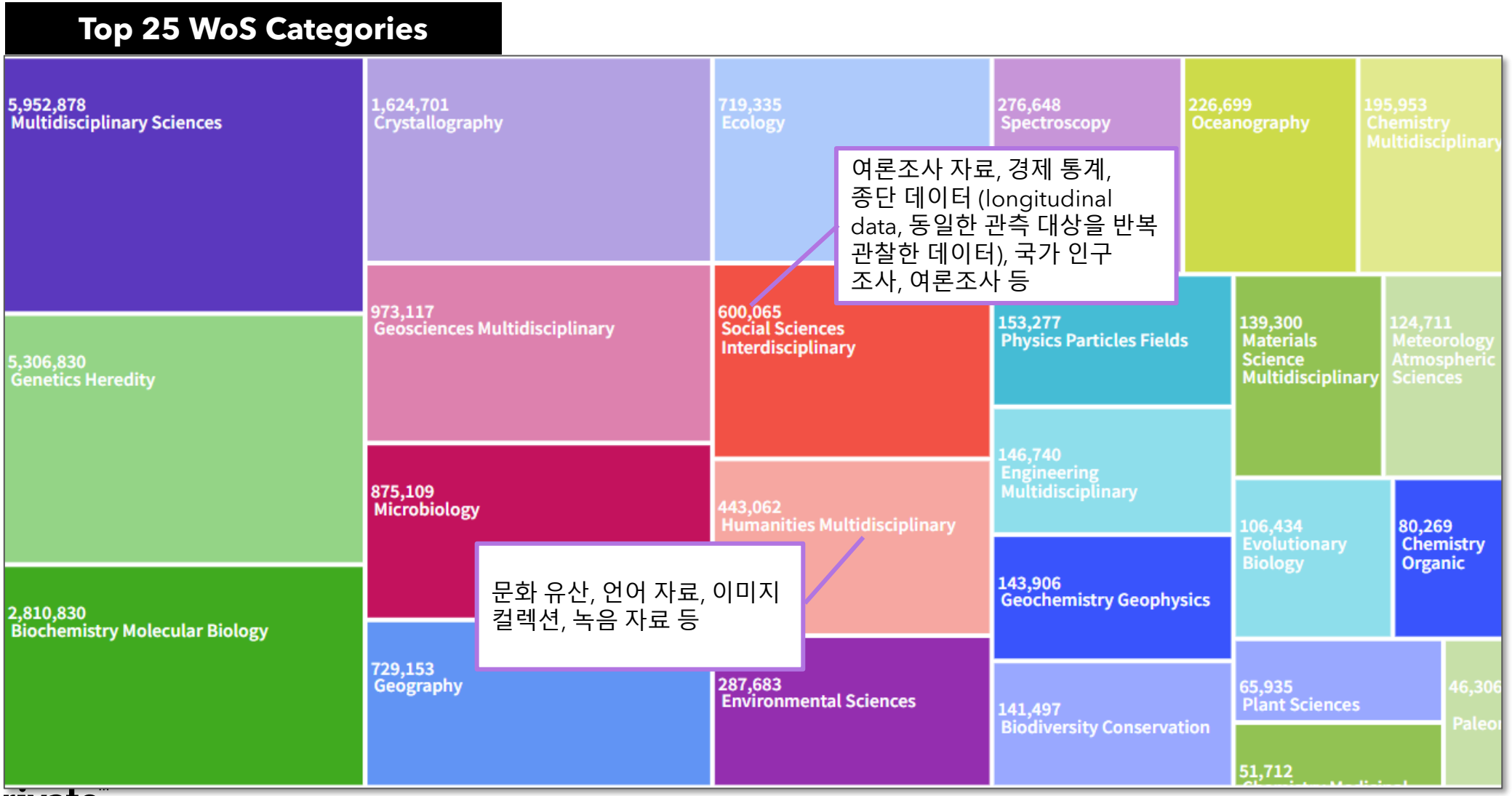
WoS DCI: Records by Country

WoS DCI는 270여국의 데이터를 광범위하게 수록함



WoS DCI: Records by WoS Category

Genetics Heredity를 포함한 160여개 WoS Category 관련 데이터를 수록



Clarivate™

감사합니다

Clarivate | 2024.