Technical Report

약용 생물자원 빅데이터 구축: 2019~2021

김상균 책임연구원*, 장호 선임연구원, 이명구 기술연구원,
김태홍 선임연구원, 예상준 책임연구원, 장윤지 기술연구원

한국한의학연구원

Big data for Medicinal Biological Resources from 2019 to 2021

Sang-Kyun Kim*, Ho Jang, Myung-ku Lee, Taehong Kim, Sang-Jun Yea, Yunji Jang

Korea Institute of Oriental Medicine

Abstract

This study aims to describe the big data for medicinal biological resources and to discuss how to use the data. The medicinal biological resources data was constructed by producing and processing the information related to medicinal materials in medicinal biological resources big data center of forest big data platform. The big data center has produced 54 million 168 GB data for 27 datasets of 9 types from 2019 to 2021 and has opened it to the public for free. This study also presents valuable use cases in two fields. The information on medicinal biological resources can be used in personal health management in daily life based on Korean medicine. In addition, it can be used to build a big data platform that analyzes biomedical data by combining the compounds of medicinal materials and existing bio big data in the biomedical fields.

Keywords: medicinal biological resources, forest big data, compounds of medicinal materials, health management, big data platform

Correspondence: 김상균(Sang-Kyun Kim)
1672 Yuseong-daero, Yuseong-gu, Daejeon, 34054, Rep. of Korea
Tel: +82-42-868-9526, E-mail: skkim@kiom.re.kr
Received 2021-10-12, revised 2021-10-25, accepted 2021-10-26, available online 2021-10-27
doi:10.22674/KHMI-9-2-6

서론

한국지능정보사회진흥원에서는 데이터 구축 및 분석을 활성화하고 데이터의 유통 및 거래 기반을 마련하기 위해서 국내의 주요 산업 분야별로 빅데이터 플랫폼 및 센터를 구축하는 사업을 2019년부터 진행해 왔다[1]. 특히, 산림 분야에서는 한국임업진흥원이 플랫폼 사업자로 선정되어 "산림빅데이터 플랫폼 및 센터 구축" 사업을 2019년부터 2021년까지 진행하고 있으며, 산림 플랫폼에 속해 있는 여러 센터들 중에서 한국한의학연구원에서는 약용 생물자원 빅데이터를 구축 및 가공하는 약용 생물자원 빅데이터 센터를 운영하고 있다[2].

약용 생물자원 빅데이터 센터는 2019년에 KT G-클라우드를 기반으로 빅데이터 센터 인프라를 구축하였으며, 2019년에는 일부 데이터를 센터의 하둡(Hadoop)과 CKAN[3]에 적재 및 발행하였다[4]. 또한, 산림 빅데이터 플랫폼에 데이터를 연계하고 적재하여 산림 빅데이터 거래소[5]를 통해서 데이터 유통 및 거래가 가능하도록 구축하였다.

이와 같이 2019년에 구축된 약용 생물자원 빅데이터 센터 인프라를 기반으로 2020년과 2021년에 약용 생물자원 데이터를 체계적으로 수집 및 구축하였다. 본 사업을 통해 약용 생물자원 빅데이터 센터에서는 최종적으로 9종의 27개의 데이터셋에 대해서 54백만 건 168GB의 데이터를 구축하고 개방하였으며, 이 데이터를 일반인들이 무료로 다운받아 사용할 수 있도록 하였다.

본 연구에서는 2019년부터 2021년까지 약용 생물자원 빅데이터 센터에서 구축한 데이터에 대해서 데이터의 생산 방법 및 개별 데이터의 특징을 기술하고, 약용 생물자원 데이터를 활용하는 시나리오에 대해서 논의하고자 한다.

본론

1. 약용 생물자원 데이터 구축

약용 생물자원 빅데이터 센터에서 구축한 9종 27개의 데이터셋 목록과 건수 및 용량은 〈표 1〉과 같다. 산림빅데이터 플랫폼은 국내 자생 생물 데이터 구축에 초점을 맞추고 있기 때문에 약용 생물자원 빅데이터 센터의 약재 데이터는 「대한민국약전」[6]과 「대한민국약전외한약(생약)규격집」[7]에 수록된 약재 목록에 한정해서 구축되었다. 하지만, 고문헌 약용 생물자원 데이터의 경우 현대의 약재 목록 이외의 약재들도 포함하고 있으며, 산림 약용 생물자원 정보에도 약전의 약재 목록 이외에 국내 자생 약용생물 자원의 분포 정보를 포함하고 있다. 전체 약용 생물자원 데이터의 건수와 용량은 약 54백만 건 168GB이다. 건수와 용량은 데이터셋마다 다르지만, 구성성분 동의어, 단백질 유전체 목록, 성분-타겟 매핑 정보 3개의 데이터셋이 전체 건수와 용량의 대부분을 차지하고 있다. 다음은 9개의 각각의 종에 대한 설명이다.

표 1. 약용 생물자원 데이터종과 데이터셋 설명

데이터종	데이터셋	데이터 설명	건수(건)	용량(MB)
약용 생물자원 정보	약용 생물자원 목록	「대한민국약전」과 「대한민국약전외한약(생약) 규격집」 약재 리스트	551	0.149
	약용 생물자원 사진	약재 사진에 대한 URL	2,995	924
	약용 생물자원 효능 정보	본초학에 기술된 약재 효능과 일반인 대상으로 번역한 데이터	1,808	0.967
	약용 생물자원 치료 정보	본초학에 기술된 약재 주치와 일반인 대상으로 번역한 데이터	4,012	2.262
	약용 생물자원 주의사항	고문헌에 나오는 약재 복용시 주의사항 정보	599	0.325
고문헌 약용 생물자원 정보	고문헌 약용 생물자원 목록	고문헌에 기록된 약용 생물자원 식재료 목록	342	0.020
	고문헌 약용 생물자원 효능 정보	고문헌에 기록된 식재료 효능과 일반인 대상으로 번역한 데이터	610	0.131
	고문헌 약용 생물자원 치료 정보	고문헌에 기록된 식재료 주치와 일반인 대상으로 번역한 데이터	771	0.173
구성성분 정보	약용 생물자원 구성 성분 목록	PubMed 논문에서 추출한 약재의 구성성분 목록	32,805	2.345
	구성성분 구조식 그림	약재 구성 성분의 2D 구조식 이미지	21,547	891
	구성성분 동의어 정보	약재 구성 성분의 동의어 정보	14,199,951	44,477
	구성성분 구조 데이터	약재 구성 성분의 SDF 데이터 파일	21,547	64
	구성성분 SMILES 정보	약재 구성 성분의 SMILES 문자열	21,547	2.259
	구성성분 PubChem 매핑	약재 구성 성분과 PubChem 화합물과의 매핑 정보	16,641	1.061
	구성성분 물리 화학 특성	약재 구성 성분의 분자량 등의 물리화학적 특성 정보	150,829	1.222
	구성성분 신체 흡수도	약재 구성 성분의 구강 섭취 후 신체 내 흡수 정도 값	21,547	1.055
	구성성분 약물 유사도	약재 구성 성분과 기존 상용 약물들과의 유사한 정도 값	21,547	0.851
단백질 유전체 정보	단백질 유전체 목록	구성 성분과 연관된 STITCH DB 타겟 단백질 목록	12,814,683	2,262
	성분-타겟 매핑 정보	구성성분과 연관된 STITCH DB 타겟 단백질 매핑 정보	25,993,529	120,055
산림 약용 생물자원 정보	산림 약용 생물자원 분포 위치 정보	자생하는 약용 생물자원의 위치 정보	11,614	1.280
	산림 약용 생물자원 발견 시기 정보	자생하는 약용 생물자원의 위치를 기록한 시기 정보	11,614	0.721
약용 생물자원 활용 정보	약용 생물자원 생산 가공 정보	약재의 채취, 시기, 가공, 포제, 보관하는 방법에 대한 정보	435	0.607
	약용 생물자원 가격 정보	약재 소비자 판매 가격에 대한 정보	625,471	211
	웰빙 푸드 레시피	약용 생물자원을 활용한 레시피 및 링크 정보	1,000	1.131
약용 생물자원 증상 정보		약재의 주치 증상들에 대한 리스트와 분류 정보	2,806	0.772
약용 생물자원 처방 정보		처방과 구성약재 리스트 정보	52,738	17
약용 생물자원 논문 정보		약재의 구성성분에 대한 논문 메타데이터 정보	9,302	16
합 계			54,042,841	168,939

1) 약용 생물자원 정보

약용 생물자원 빅데이터 센터에서 포함하고 있는 약재 목록은 총 551개이다. 약재는 라틴명(생약명)을 식별자(Identifier)로 사용하여 구별하며 모든 다른 데이터 셋들과의 정보 연결도 라틴명을 통해 매핑된다. 약재의 사진 데이터 셋은 약재 사진에 대한 URL을 포함하고 있어 해당 URL로 가면 사진을 확인할 수 있다. 각각의 약재에 대해서 본초학[8] 교과서에 나오는 효능과 주치 데이터를 제공하며, 효능과 주치 용어를 일반인들이 이해할 수 있도록 번역하고 요약한 데이터도 제공한다. 또한, 약용 생물자원 주의사항 데이터에는 고문헌에 나오는 복용시 주의사항 정보를 포함하고 있다.

2) 고문헌 약용 생물자원 정보

고문헌 약용 생물자원 정보는 《동의보감》의 단방(單方), 《향약집성방》, 《의방유취》의 고문헌에서 약재를 포함한 342개의 식재료에 대해서 효능과 주치 데이터를 추출한 정보이다[9]. 모든 데이터에 대해서 출전이 명시되어 있으며, 일반인들이 이해할 수 있는 수준의 번역 및 가공 데이터도 제공한다.

3) 구성성분 정보

구성성분 정보는 PubMed 데이터베이스에서 약용 생물자원의 크로마토그래피 논문을 검색한 후에 논문의 초록과 본문을 보고 약재의 구성성분을 추출한 정보이다[10]. 추출된 구성성분 목록에 대해서 ChemDraw Professional v19[11]을 이용해서 구조식 그림을 그렸으며, SMILES (Simplified molecular-input line-entry system), SDF (structure-data file), 물리화학특성 데이터셋 데이터[12]는 ChemAxon의 Calculators and Predictors v19[13]를 이용해서 계산하였다. 구성성분 동의어와 PubChem 매핑 데이터는 PubChem Compound 데이터베이스[14]에서 제공하는 정보를 이용하여 구축하였으며, 개별 구성성분에 대해서 신체 흡수도[15]와 약물 유사도[16]를 계산하였다.

4) 단백질 유전체 정보

단백질 유전체 정보는 STITCH v5 데이터베이스[17]에서 제공하는 단백질 목록과 성분-단백질 매핑 정보를 구축한 것이다. 이 정보에는 약용 생물자원과 연관된 데이터뿐만 아니라 다른 생물종에 대한 단백질 유전체 정보도 포함하고 있다.

5) 산림 약용 생물자원 정보

산림 약용 생물자원 정보는 국내에서 실제 자생하고 있는 생물자원을 탐사해서 발견한 시간과 GPS 위치를 기록한 정보이다[18]. 하지만, 약용 생물자원의 보호를 위해서 정확한 GPS 정보는 공개하지는 않으며, 위도와 경도 정보를 소수점 셋째 자리까지 잘라서 공개하였다.

6) 약용 생물자원 활용 정보

약용 생물자원 활용 정보는 약재의 생산 및 가공 정보, 가격 정보, 음식 레시피 정보 등을 포함한다. 약재의 생산 가공 정보는 오아시스의 약재 정보 중 생산가공 데이터[19]를 활용하였으며, 가격 정보는 국내 주요 약재 판매 웹사이트에서 가격 정보를 크롤링하였다. 또한, 음식 레시피 정보는 국내 여러 음식 레시피 관련 포털에서 약재 관련 음식 레시피를 크롤링하였는데 모든 개별 레시피 데이터에 대해서 출처를 명시하였다.

7) 약용 생물자원 증상, 처방, 논문 정보

약용 생물자원의 증상 정보, 약용 생물자원 처방 정보, 약용 생물자원 논문 정보는 2021년에 추가된 데이터종으로써, 기존 약용 생물자원 정보와 연관되는 정보들이다. 증상 정보는 약용 생물자원의 효능과 주치 정보에서 나오는 증상 용어들을 추출해 정리한 데이터이며, 처방 정보는 한의 온톨로지[20]에 나오는 처방 리스트이다. 그리고 논문 정보는 약용 생물자원 구성성분이 추출된 논문의 메타데이터 정보이다.

2. 약용 생물자원 데이터 활용

1) 산림빅데이터 거래소

모든 약용 생물자원 데이터는 산림빅데이터 거래소에서 검색 및 다운로드가 가능하다. 산림빅데이터 거래소에는 약용 생물자원 빅데이터 센터뿐만 아니라 산림빅데이터 플랫폼에 속하는 여러 다른 센터들의 데이터들도 검색할 수가 있다. 산림 빅데이터 거래소는 데이터를 상품으로 관리하며 유료 데이터의 경우 상품을 결재 후 다운로드를 하도록 되어 있다. 하지만 약용 생물자원 데이터는 모두 무료이기 때문에 회원가입 후 바로 다운로드가 가능하다. 단, 약용 생물자원 사진, 구성성분 구조식 그림, 구성성분 구조 데이터는 거래소에서 URL 목록만 다운받을 수 있으며, 실제 파일들은 별도의 서버에 존재하고 해당 URL에 접속하면 데이터를 개별적으로 다운받을 수 있다.

2) 데이터 활용 시나리오

약용 생물자원 데이터는 모두 약재와 연관된 데이터들이기 때문에 약재 정보를 필요로 하는 여러 분야에서 활용이 가능하다. 특히, 본 연구에서는 다음과 같은 유용한 두 가지의 대표적인 시나리오에 대해서 논의한다.

첫째, 문헌에 기재된 올바른 정보에 기반해서 일상 건강 관리를 하는데 약용 생물자원 정보를 활용하는 것이다. 인터넷에는 약재의 효능과 관련해서 수많은 정보를 검색할 수 있다. 하지만, 전문 의료 용어를 그대로 사용해 일반인들이 이해하기 어려운 것도 있으며, 고문헌에서 어떤 식재료가 고혈압과 암 등의 질병에 좋다고 기록되었다는 다소 과장된 광고들도 존재한다. 약용 생물자원 데이터에서 약재의 효능과 주치 정보는 본초학 교과서에서 가져온 후에 일반인들이 이해할 수 있는 수준으로 번역하고 요약해서 제공하였다. 그리고 복용 주의사항 데이터와 고문헌의 효능 및 치료 데이터는 데이터를 추출한 출전을 명시하였다. 〈그림 1〉은 인삼에 대해서 구축된 데이터들의 일부이다.

약용 생물자원 데이터는 이러한 약재 정보뿐만 아니라 고문헌에 언급된 식재료의 효능과 주치 정보, 식재료를 이용한 웰빙 푸드 레시피 등을 포함하고 있으며, 정보 시스템에서 증상을 검색하기 쉽도록 효능과 주치에서 증상 용어들도 제공하고 있다. 이와 같이 문헌에 기반한 약재 정보, 식재료 정보, 음식 정보 등을 증상 정보와 연관시키면 개인이 일상에서 호소하는 증상들에 대해서 식재료, 음식, 약재 등을 추천받을 수 있기 때문에 건강 관리를 하는 데 도움을 줄 수 있을 것이다.

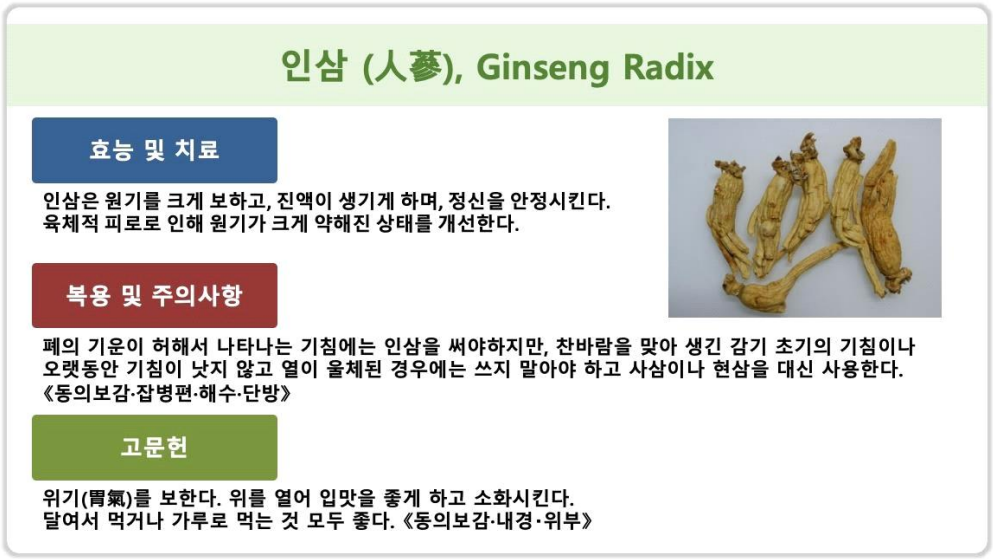

그림 1. 인삼의 효능 및 주치, 복용 주의사항, 고문헌 주치 예제

둘째, 바이오 의료 분야에서 약재의 구성성분 데이터를 활용하는 것이다. 본 연구에서 구축한 구성성분 데이터는 데이터마이닝을 이용해서 자동으로 구축한 것이 아니라 PubMed 데이터베이스에 있는 연구 논문들을 읽고 직접 구축한 것이다. 따라서 기존에 문헌에서 잘 알려진 성분 이외에도 최신 실험 연구를 통해 밝혀진 성분들도 포함되어 있다. 특히, 구성성분 정보는 구성성분명 이외에도 성분 구조식, SMILES 문자열, 물리화학적 특성 정보 등 개별 성분의 상세 정보들도 제공함으로써, 인실리코 기반의 신약, 화장품, 기능성 제품 개발에 활용될 수 있다. 뿐만 아니라 인공지능 기반의 빅데이터 플랫폼에 추가되어 약재의 효능을 예측하는 알고리즘을 만드는 데도 활용이 가능할 것이다. 현재, 약용 생물자원 구성성분 정보는 3BIGS에서 구축한 3X-KBank[21] 플랫폼에서 활용되고 있다. 3X-KBank는 유전자, 약물, 질병 등 다양한 바이오 데이터를 통합한 바이오 데이터 통합 플랫폼이다. 약용 생물자원의 구성성분 데이터는 기존에 이 플랫폼에 존재하는 약재 및 유전자 정보들과 통합되었으며 신약 후보군의 발굴 및 다양한 신소재 발굴에 대한 연구에서 활용될 예정이다.

결론

본 연구에서는 2019년에 구축한 약용 생물자원 빅데이터 센터 인프라를 기반으로 3년 동안 구축한 약용 생물자원 데이터에 대해서 기술하였다. 현재까지 약 54백만 건 168GB의 데이터를 생산하였으며, 산림빅데이터 거래소를 통해서 데이터를 개방 완료하였다. 약용 생물자원 데이터는 약재의 효능과 주치를 포함해서 구성성분과 단백질 등의 9종 데이터들을 포함하고 있다. 이와 같이 다양한 종류의 데이터는 약재를 활용하는 다양한 분야에서 활용이 가능하지만 본 연구에서는 두 가지 분야에서 유용한 활용 사례를 제시하였다. 하나는 일상 개인 건강 관리로써 본 연구에서 구축한 데이터를 활용하면 한의 기반의 건강 관리 솔루션들을 개발할 수 있을 것이다. 다른 하나는 바이오 의료 분야에서의 구성성분 데이터 활용으로 약재의 구성성분과 기존의 바이오 빅데이터를 결합하여 바이

오 의료 데이터를 분석할 수 있는 빅데이터 플랫폼 구축에 활용될 것으로 기대된다. 향후에는 현재까지 구축한 데이터를 산림빅데이터 거래소를 통해 지속적으로 관리하고 유지할 계획이다. 또한, 약재의 구성성분 데이터에 대해서는 PubMed에서 최신 연구가 발표되면 데이터를 추가로 구축할 계획이다.

감사의 글

이 논문은 한국한의학연구원 주요사업 “AI 한의사 개발을 위한 임상 빅데이터 수집 및 서비스 플랫폼 구축(KSN2012110)”과 한국지능정보사회진흥원의 산림빅데이터 플랫폼 및 센터 구축 사업의 지원을 받아 수행되었습니다.

참고문헌

1. 과학기술정보통신부. 빅데이터 플랫폼 및 센터 구축 사업 선정 결과. 대한민국정책브리핑. 2019-05-13. Available from: https://www.korea.kr/news/pressReleaseView.do?newsId=156331134 (accessed 2020-10-05)
2. 한국한의학연구원. 약용 생물자원 빅데이터 센터 구축. 빅데이터 센터 5 개 분야 구축 사업 결과보고서. 2020.
3. CKAN. Available from: https://ckan.org (accessed 2020-10-05)
4. 장윤지, 김태홍, 이명구, 장호, 예상준, 김상균. 약용 생물자원 빅데이터 센터 구축. 한국지식정보기술학회 논문지. 2020;15(5):721-30.
5. 한국임업진흥원. 산림빅데이터 거래소. Available from: https://www.bigdata-forest.kr (accessed 2020-10-05)
6. 식품의약품안전처(대한민국). 대한민국약전. 식품의약품안전처고시 제 2019-102 호, 공포 2019-11-06. 전부개정. 시행 2020-02-07.
7. 식품의약품안전처(대한민국). 대한민국약전외한약(생약)규격집. 식품의약품안전처고시 제 2020-12 호, 공포 2020-02-25. 일부개정. 시행 2020-02-25.
8. 한의과대학 본초학 편집위원회. 본초학. 서울:영림사. 2008.
9. 김상현, 남보령, 김상균. 일반인을 위한 한의학 지식 구축. 한국지식정보기술학회 논문지. 2018;13(2):231-40.
10. KIM SK, Nam SJ, Jang HC, Kim AN, Lee JJ. TM-MC: a database of medicinal materials and chemical compounds in Northeast Asian traditional medicine. BMC Complementary Alternative Medicine. 2015;15:218.
11. PerkinElmer ChemDraw Professional. Available from: https://perkinelmerinformatics.com/products/research/chemdraw (accessed 2020-10-05)
12. Kim SK, Lee SH, Lee MK, Lee SH. A systems pharmacology approach to investigate the mechanism of Oryeong-san formula for the treatment of hypertension. Journal of Ethnopharmacology. 2019;244.
13. ChemAxon Calculators and Predictors. Available from: https://chemaxon.com/pro

ducts/calculators-and-predictors (accessed 2020-10-05)
14. Wang Y, Xiao J, Suzek TO, Zhang J, Wang J, Bryant SH, 2009. PubChem: a public information system for analyzing bioactivities of small molecules. Nucleic Acids Research. 2009;37:W623-33.
15. Veber DF, Johnson SR, Cheng HY, Smith BR, Ward KW, Kopple KD. Molecular properties that influence the oral bioavailability of drug candidates. Journal of Medicinal Chemistry. 2002;45(12):2615-23.
16. Bickerton GR, Paolini GV, Besnard J, Muresan S, Hopkins AL. Quantifying the chemical beauty of drugs. Nature Chemistry. 2012;4(2):90-8.
17. Szklarczyk D, Santos A, von Mering C, Jensen LJ, Bork P, Kuhn M. STITCH5: augmenting protein-chemical interaction networks with tissue and affinity data. Nucleic Acids Research. 2016;44(D1):D380-4.
18. 한국한의학연구원. 국가생약자원의 수집조사 연구. 식품의약품안전처 연구보고서. 2020.
19. 한국한의학연구원. OASIS 전통의학 정보포털. Available from: https://oasis.kiom.re.kr (accessed 2020-10-05)
20. Kim SK, Park DH, Kim AN, Oh YT, Kim JY, Yea SJ, Kim C, Jang HC. 한의 온톨로지 기반 시맨틱 검색 시스템. 한국콘텐츠학회논문지. 2012;12(12):533-43.
21. 3BIGS. 3X-KBank. Available from: https://3bigs.com/3x-kbank (accessed 2020-10-05)