국가 연구데이터플랫폼과 바이오 연구데이터플랫폼의 메타데이터 상호운용성에 관한 연구*

A Study on Metadata Interoperability between the National Research Data Platform and the Bio Research Data Platform

박성은 (Seong-Eun Park)**
고영만 (Young Man Ko)***

초 록

'국가 연구데이터플랫폼'과 '바이오 연구데이터플랫폼'은 비교적 최근 구축되어 활발하게 각각의 생태계를 만들어 가고 있다. 따라서 다른 메타데이터 표준을 기반으로 독립적으로 구축되어 향후 상호운용성의 문제가 발생할 수 있다. 본 연구의 목적은 각 플랫폼의 메타데이터 요소를 매핑하고, 이를 검증하여 상호운용성을 확보하기 위한 기반을 제안하는 것이다. 이를 위해 각 플랫폼의 메타데이터 표준을 분석하고 크로스워크 대상을 선정하여 매핑한 후, 바이오 분야 전문가를 통해 매핑된 요소의 적합성을 검증하고 더 적절한 매핑 요소를 추천받아 데이터셋 및 파일에 대한 메타데이터 요소를 도출하였다. 이를 통해 각 플랫폼의 메타데이터가 의미적으로 연결될 수 있는 가능성과 상호운용성 확보를 위한 기반을 확인할 수 있었다.

ABSTRACT

The 'National Research Data Platform' and the 'Bio Research Data Platform' were recently built and each is actively creating an ecosystem. It is built independently based on other metadata standards, which may cause future interoperability issues. The purpose of this study is to propose a basis for metadata interoperability between the two platforms. To this end, the metadata standards of each platform were analyzed, crosswork targets were selected and mapped, and the suitability of the mapped elements was verified through experts in the bio field. And more appropriate mapping elements were recommended to derive metadata elements for datasets and files. Through this, it was possible to confirm the possibility that the metadata of each platform could be semantically linked and the basis for securing interoperability.

키워드: 연구데이터, 메타데이터, 상호운용성, 바이오 연구데이터플랫폼, 국가 연구데이터플랫폼
research data, metadata, interoperability, bio research data platform, national research data platform

---

* 본 논문은 성균관대학교 및 교육부, 한국연구재단의 4단계 두뇌한국21 사업 대학원혁신으로 지원된 연구임.
** 성균관대학교 일반대학원 문헌정보학과 박사과정(pse3598@skku.edu) (제1저자)
*** 성균관대학교 문과대학 문헌정보학과 교수(ymko@skku.edu) (교신저자)
■ 논문접수일자 : 2022년 5월 14일 ■ 최초심사일자 : 2022년 6월 2일 ■ 게재확정일자 : 2022년 6월 8일
■ 정보관리학회지, 39(2), 159-202, 2022. http://dx.doi.org/10.3743/KOSIM.2022.39.2.159

1. 서 론

1.1 연구의 배경 및 필요성

연구 환경의 디지털화는 연구 생산성을 향상시켜 과학기술 분야에서의 연구 성과물뿐만 아니라 그 과정에서 도출되는 연구데이터의 생산까지 폭발적으로 증가시키고 있다. 데이터 중심 R&D의 활성화로 인해 연구데이터의 활용 수요가 증가함에 따라, 국가 R&D를 통해 축적되는 연구데이터를 다양한 연구자가 공유하고 활용할 수 있도록 하는 체계의 구축이 요구되어 왔으며, 정부에서는 연구데이터의 관리를 위해 국가적 차원에서 법과 제도를 개선하고 인프라를 구축하는 정책을 추진하고 있다. 이러한 인프라 구축 정책의 일환으로 한국과학기술정보연구원(Korea Institute of Science and Technology Information, 이하 KISTI)은 '국가 연구데이터플랫폼'을 구축하여 2018년 12월 시범서비스를 시작으로 2020년 1월 'DataON'이라는 이름으로 정식서비스를 개시하였다.

'국가 연구데이터플랫폼'은 과학기술 전 분야의 각종 연구데이터에 대한 등록부터 검색과 활용까지의 서비스를 제공하는 플랫폼이며, 바이오, 소재 등 분야별 전문센터도 구축되고 있다. 특히 바이오 분야 전문센터인 '바이오 연구데이터플랫폼'은 2020년 정부가 발표한 '생명연구자원 빅데이터 구축 전략'을 근거로 국가생명연구자원정보센터(Korea Bioinformation Center, 이하 KOBIC)에서 '국가 바이오 데이터 스테이션(K-BDS)'이라는 이름으로 구축해왔으며, 2021년 10월 시범 운영을 시작하였다. '국가 연구데이터플랫폼'은 이러한 분야별 전문센터들이 확보하고 연계한 연구데이터를 다시 '국가 연구데이터플랫폼'으로 취합하는 연결망 체계를 갖추고자 노력하고 있다.

각 전문센터에서 구축한 연구데이터가 다시 '국가 연구데이터플랫폼'으로 취합되어 연결망 체계를 갖추기 위해서는 각 플랫폼에서 사용하는 연구데이터의 메타데이터 표준이 의미적으로 연결될 수 있는 상호운용성을 확보하는 것이 필수적이다. 상호운용성은 일반적으로 "둘 또는 그 이상의 시스템이나 구성요소가 특별한 노력 없이도 정보를 교환하거나 교환된 정보를 활용하는 능력"으로 정의되며, 메타데이터에서의 상호운용성은 "구문적, 구조적, 의미적 비일관성을 최소화함으로써 둘 또는 그 이상의 메타데이터 표준 사이에서 데이터를 교환하고, 이를 통해서 정보를 공유하거나 재사용할 수 있는 기회를 극대화하는 능력"으로 정의할 수 있다(남태우, 이승민, 2014, 233-234).

그런데 '국가 연구데이터플랫폼'의 메타데이터 표준과 '바이오 연구데이터플랫폼'의 메타데이터 표준은 각각 다른 표준을 기반으로 독립적으로 구축되었다. '국가 연구데이터플랫폼' 메타데이터의 표준 스키마는 범용적 연구데이터 기술을 위해 수집된 국내 표준 메타데이터인 '연구데이터 관리 및 공유를 위한 메타데이터(TTAK.KO-10.0976, 이하 TTA 표준)'를 중심으로 국내・외 연계 대상기관, OpenAIRE의 메타데이터 스키마 등을 비교 분석하여 설계되었다. 그리고 '바이오 연구데이터플랫폼'의 메타데이터 표준 스키마는 바이오 분야 내에서의 데이터 호환성을 고려하여 바이오 분야에서 국제적으로 통용되는 미국 국립생물공학정보센터(National Center for Biotechnology Information,

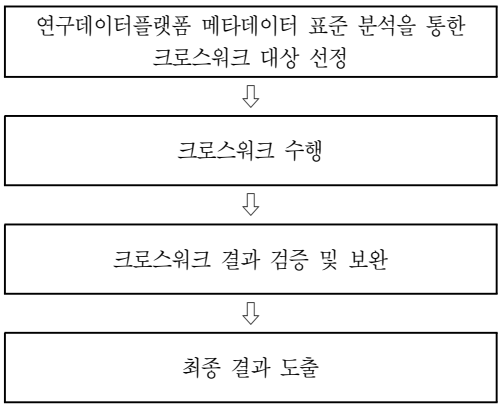
이하 NCBI)와 유럽 생물정보학 연구소(European Bioinformatics Institute, 이하 EBI)의 메타데이터 표준 스키마를 준용하여 설계되었다.

따라서 각각 다른 표준을 기반으로 독립적으로 구축된 두 플랫폼 간에는 향후 상호운용성의 문제가 발생할 수 있으며, 이러한 문제를 극복할 수 있는 방안을 마련할 필요가 있다. 본 연구에서는 '국가 연구데이터플랫폼'의 연구데이터 메타데이터 요소와 '바이오 연구데이터플랫폼'의 연구데이터 메타데이터 요소를 크로스워크(Crosswalk)를 통해 매핑하고, 이를 검증하여 상호운용성을 확보하기 위한 기반을 마련하고자 하였다. 크로스워크란, 여러 메타데이터 간에 메타데이터 요소의 의미와 구조를 매핑하는 것을 말한다(고영만, 서태설, 임태훈, 2007).

1.2 연구 절차

'국가 연구데이터플랫폼'과 '바이오 연구데이터플랫폼'의 연구데이터 메타데이터 요소를 크로스워크를 통해 매핑하고, 이를 검증하여 상호운용성을 확보하기 위한 절차는 다음과 같다.

첫 번째 단계에서는 연구의 대상이 되는 연구데이터플랫폼들의 메타데이터 표준을 분석하고, 이를 통해 크로스워크 대상을 선정하였다. 두 번째 단계에서는 각 표준에서 도출한 메타데이터 요소들을 대상으로 크로스워크를 수행하였다. 세 번째 단계에서는 크로스워크된 메타데이터 요소에 대해 바이오 분야 전공자들을 대상으로 검증을 실시하였다. 마지막 단계에서는 검증 결과를 반영하여 최종 크로스워크 결과를 도출하였다(〈그림 1〉 참조).



〈그림 1〉 연구 절차 도식화

2. 이론적 배경

2.1 연구데이터플랫폼과 연구데이터 메타데이터

연구데이터는 국가별로 다양하게 정의하고 있으나 내용적으로는 유사하다. 우리나라에서는 2020년 3월 17일 시행된 「국가연구개발사업의 관리 등에 관한 규정(공동관리규정, 대통령령 제31297호)」에서 연구데이터를 "R&D 과제 수행 과정에서 실시하는 각종 실험, 관찰, 조사 및 분석 등을 통하여 산출된 사실 자료로서 연구 결과의 검증에 필수적인 데이터"로 정의하고 있다. 연구데이터의 형태는 숫자, 이미지, 텍스트, 동영상, 소리 등 다양하며, 그 역할은 연구 자체에 대한 증빙과 검증 가능성, 데이터의 중복수집 방지와 결과 확산, 연구의 혁신과 잠재적인 새로운 데이터의 이용 촉진으로 요약될 수 있다(UK Data Archive, 2011).

연구데이터는 과학기술 분야뿐만 아니라 인문사회 분야에서도 다양하게 생산되고 있다.

국내의 대표적인 과학기술 분야 연구데이터 플랫폼으로는 KISTI의 국가 연구데이터플랫폼 DataON이 있으며, 인문사회 분야 연구데이터플랫폼은 한국연구재단(National Research Foundation of Korea, 이하 NRF)의 기초학문자료센터(Korean Research Memory, 이하 KRM)이다.

2020년부터 정식서비스를 시작한 DataON은 과학기술 전 분야의 이용자의 연구기획부터 결과물 등록까지 연구데이터와 관련된 연구의 전 과정을 효과적으로 수행할 수 있도록, 각종 연구데이터에 대한 등록부터 검색・활용까지의 서비스를 제공하고 있으며, 연구데이터를 표준화하여 고유 식별번호를 붙이는 등의 기능도 수행한다.

연구데이터의 메타데이터와 관련해서는 주로 메타데이터의 개발에 관한 연구가 도서관, 한의학, 응집물질물리, 임산공학 등 특정 분야에서 이루어져 왔다. 도서관 분야를 대상으로 한 이미화, 이은주, 노지현(2020)의 연구는 국립중앙도서관에서 운영하는 오픈액세스 리포지터리인 OAK(Open Access Korea)에서 연구데이터를 기술하기 위하여 기존 메타데이터의 확장 방안을 제안한 것으로, NRF의 KRM, KISTI의 DataON, 샌디에고 대학 도서관 컬렉션 그리고 미시간 데이터 리포지터리를 분석하였다. 예상준, 장호, 김선태(2019)는 한의학 분야 연구데이터 관리 및 공유를 위한 메타데이터 요소를 설계하고, 한국한의학연구원에서 생산되는 연구데이터를 대상으로 요소 검증을 수행하였다. 이 외에도 Geoscience 분야(김주섭, 김선태, 최상기, 2020), 응집물질물리 분야(김성욱, 김선태, 2020) 연구데이터 관리를 위한 메타데이터 연구가 수행되었으며, 특히 임산공학 분야를 대상으로 한 김주섭 외(2020b)의 연구와 생태 분야를 대상으로 한 김주섭 외(2020a)의 연구에서는 본 연구에서 활용하는 메타데이터 크로스워크 방법을 활용하여 연구데이터를 관리하기 위한 메타데이터 항목을 도출한 바 있다.

'국가 연구데이터플랫폼'의 연구데이터 메타데이터와 관련한 선행연구로는 '국가 연구데이터플랫폼'의 연구데이터 메타데이터의 평가 모델을 개발한 고영만(2019)의 연구가 유일하다. 이 평가 모델은 크게 '연구데이터 메타데이터의 품질평가 모델'과 '연구데이터의 유용성평가 모델'로 나뉘어 구현되었으며, 본 연구에서는 해당 연구에서 도출된 '연구데이터 메타데이터의 품질평가 모델'의 도출 과정 중 일부를 참조하였다.

2.2 바이오 연구데이터와 연구데이터플랫폼

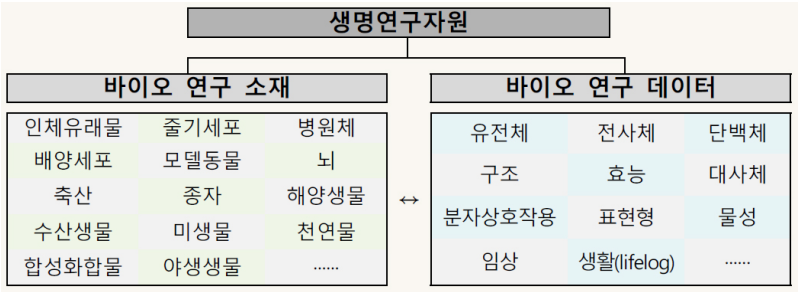
바이오 연구데이터는 실험, 관찰, 조사, 분석 등 바이오 연구를 통해 생산되어 성과 도출에 활용되는 대사물질 종류 정보, 단백질 구조 정보, 유전체 염기 정보, 화합물 반응 정보 등의 객관적 사실 데이터를 의미한다. 바이오 연구 및 산업 활동은 연구에 필요한 소재를 선택해서 실험하고 실험과정에서 생산된 데이터를 분석하여 유의미한 결론을 도출하는 과정으로 이루어진다. 그동안 R&D의 재료로만 인식되던 바이오 소재와 데이터는 데이터양의 급증과 빅데이터를 활용하는 바이오 기술 발전에 따라 기술을 혁신하는 핵심 요소로 부각되고 있다.

'바이오 연구데이터'는 '바이오 연구 소재'와

함께 '생명연구자원'에 속한다. 「생명연구자원의 확보·관리 및 활용에 관한 법률(생명연구자원법, 법률16016호)」 제2조에 의하면 '생명연구자원'은 생명공학 연구의 기반이 되는 자원으로 연구 또는 산업적으로 실질적, 잠재적 가치가 있는 동물, 식물, 미생물, 인체유래 연구자원 등 다양한 생물체의 실물(實物, 유전자원을 포함한다)과 관련 정보 등을 의미한다. 바이오 연구 수행 전에 필요한 '연구 소재'와 바이오 연구 수행을 통해 도출되는 '데이터'를 총칭하는 '생명연구자원'은 생명공학 연구의 기반이 되기 때문에, 바이오 연구데이터는 생명공학 분야의 일부분으로 간주될 수 있다.

따라서 본 연구에서는 생명공학 분야의 연구데이터와 관련한 연구를 분석하였다. 생명공학 분야의 연구데이터 관리 현황 및 개선 방안에 관한 연구를 진행한 강주연(2017)은 생명공학 분야의 기초 연구성과가 비교적 쉽게 실용화될 수 있기 때문에 생명공학 분야가 타 학문으로의 확장성이 높으며, 이러한 점이 생명공학 분야 연구데이터의 재사용 빈도를 높이는 것으로 파악하고, 데이터의 신뢰성을 확보할 만큼의 적절한 관리가 이루어지고 있는지에 대한 현황을 분석하였다. 박미영, 안인자, 김준모(2018)는 국내외 생명공학 분야 주요 기관별 연구데이터 공유방법을 조사하고, 연구데이터 공유방법별 연구데이터 관리 구성요소를 확인함으로써 국내의 생명공학 분야 연구데이터 공유 및 활용 방안을 제시 하였다. 또한 김선(2022)은 생명공학 분야 연구자를 대상으로 연구데이터 공유 의도에 영향을 미치는 요인을 분석하기 위해 Ostrom의 집단행동이론을 적용하였으며, 변인들의 영향관계 안에서 학술적 평판의 조절효과를 발견하고자 하였다. 하지만 생명공학 분야의 연구데이터에 대하여 관리 현황이나 공유의도를 파악하고자 하는 연구가 진행된 것에 비해 본 연구에서 다루고자 하는 바이오 연구데이터의 메타데이터와 관련된 내용은 다루어지지 않았다.

미국, EU, 일본, 중국 등 주요국들은 바이오 데이터의 중요성을 인식하여 바이오 데이터의 수집과 공유 체계를 조성하여 운영하고 있다. 미국 국립과학재단(National Science Foundation, 이하 NSF), 국립보건원(National Institutes of Health, 이하 NIH) 등 연방기관들은 데이터 정책과 규정을 마련하여 시행 중이며, 특히 NIH



〈그림 2〉 생명연구자원의 범위

출처: 제3차 국가생명연구자원 관리·활용 기본계획('20~'25)(안)(최기영 외, 2020)

는 2015년부터 NIH 과제에서 생산되는 데이터는 NIH 지정 데이터 저장소인 NCBI에 제출하도록 제도화했다.

국내의 경우, 과학기술정보통신부에서는 연구성과물 관리 제도를 통해 국가 R&D에서 생산된 바이오 연구데이터의 제출을 의무화하도록 하고, 생명정보 연구성과물을 등록받는 공식기관으로 KOBIC을 지정하여 운영하고 있다. 각 부처별로 이러한 체제를 운영함에 따라 국가 차원의 관리체계가 미비하여 데이터의 양적, 질적 측면에서도 관리가 미흡한 상황이었다. 따라서 국내 대부분의 연구자들은 국내에서 생산된 데이터의 대부분을 NCBI, EBI, DDBJ (DNA Data Bank of Japan) 등의 외국 바이오 데이터 등록기관에 등록해왔으며, 이와 같은 외국 공개 DB에 축적된 데이터를 활용해왔다.

그러나 최근 코로나19 상황에서 주요 선진국들의 경우 그간 축적한 바이오 연구데이터를 기반으로 코로나19 치료 약물을 신속하게 개발하여 사람 대상 임상시험에 진입한 바 있어, 바이오 재난 대응을 위한 인프라 역량 강화의 필요성이 재확인되었다. 물론 2011년에도 미국, 유럽에서 데이터 관리 비용 증가 및 데이터 가치 상승으로 자국 데이터의 공개 제한 움직임을 보여 필요성이 확인되었던 사례가 있었다.

이에 따라 정부는 바이오 분야의 성장을 가속화하고, 바이오 경제의 핵심 자원인 데이터와 소재 인프라를 본격 육성하기 위해 2020년 '생명연구자원 빅데이터 구축전략'을 발표하였다. 이를 기반으로 과학기술정보통신부는 관계 부처와 함께 부처, 사업, 연구자별로 흩어져 있는 바이오 연구데이터를 통합 수집하고, 제공하는 데이터 기반의 바이오 연구 환경을 조성하기 위해 '바이오 연구데이터플랫폼'인 '국가 바이오 데이터 스테이션(K-BDS)'을 구축해왔으며, 2021년 10월 시범 운영을 시작하였다.

3. 연구데이터플랫폼의 메타데이터 표준 분석 및 크로스워크 대상 선정

3.1 TTAK.KO-10.0976

'국가 연구데이터플랫폼' 메타데이터의 표준 스키마는 범용적 연구데이터 기술을 위해 수집된 국내 표준 메타데이터인 TTA 표준을 중심으로 국내・외 연계 대상기관의 메타데이터 스키마, OpenAIRE의 메타데이터 스키마 등을 비교 분석하여 설계되었다. TTA 표준은 한국정보통신기술협회에서 정의한 정보통신단체표준으로, 메타데이터 프로젝트그룹(PG606)에 제안되고, 소프트웨어/콘텐츠 기술위원회(TC6)에서 심의를 통과하여 2017년 3월에 승인되었다. 이는 연구과제 수행 도중 생산된 연구데이터를 메타데이터 기반으로 효과적으로 관리, 공유 및 재활용함으로써 연구자와 기관의 자산인 데이터를 보존하고 데이터의 재활용을 통해 연구의 효율성을 제고하는 도구로 활용될 수 있도록 메타데이터 관리 체계, 요소 및 세부 사항을 표준으로 제정한 것이다.

TTA 표준은 여러 종류의 연구데이터를 효과적으로 수집, 저장, 관리, 재활용하기 위해 연구데이터의 관리 단위를 '컬렉션-데이터셋-파일' 구조로 하는 리포지터리 저장소 아키텍처의 구성을 제안하고 있다. '리포지터리'는 기관에서

연구데이터를 수집, 저장, 관리하기 위한 시스템을 의미하며 'Repository', 'Repository URL', 'Identifier' 등 21개 요소로 이루어져 있다. '컬렉션'은 데이터셋을 그룹화하기 위한 논리적 그룹으로 프로젝트, 부서, 연구과제 등으로 컬렉션을 자유롭게 구성할 수 있으며, 'Collection', 'Identifier', 'Title' 등 12개 요소로 이루어져 있다. '데이터셋'은 연구데이터의 관리 및 공유, 재사용성을 높이기 위해 특정한 속성에 따라 묶은 파일 그룹을 의미하며, 'Dataset', 'Identifier', 'Rights' 등 15개 요소로 이루어져 있다. 마지막으로 '파일'은 관리 및 공유, 재사용의 가치가 있는 개별 단위의 연구데이터를 의미하며, 'File', 'Identifier', 'Format' 등 19개 요소로 이루어져 있다. TTA 표준의 전체 요소들은 [부록 1]에서 확인할 수 있다.

3.2 국가 연구데이터플랫폼

'국가 연구데이터플랫폼'은 TTA 표준의 메타데이터 스키마 구조를 참조하여 '리포지터리(Repository)', '프로젝트(Project)', '데이터셋(Dataset)', '파일(File)'의 4가지 계층 구조로 설계되었다. '국가 연구데이터플랫폼'의 스키마 구조는 매년 조금씩 업데이트되고 있으며, 2.1절에서 언급한 것처럼 본 연구는 고영만(2019)의 연구를 참조하여 크로스워크 대상을 도출하였기 때문에, 해당 연구에서 기준으로 한 2019년도의 스키마를 분석 대상으로 한다.

'리포지터리'는 각 연구데이터의 저장 및 관리를 위한 보유기관으로서의 리포지터리에 관한 정보, '프로젝트'는 연구데이터가 생성된 과제 즉 프로젝트와 관련된 정보, '데이터셋'은 연구데이터 관리의 핵심 계층으로 연구데이터 파일에 대한 논리적인 묶음 단위를 의미하며, '파일'은 개별 파일에 대한 정보를 등록한다(한국과학기술정보연구원, 2018). '리포지터리'는 13개 요소로 이루어져 있으며, '프로젝트'는 41개, '데이터셋'은 61개, '파일'은 36개의 요소로 이루어져 있다. '국가 연구데이터플랫폼'의 전체 스키마 요소들은 [부록 2]에서 확인할 수 있다.

고영만(2019)은 연구데이터 메타데이터의 품질평가 모델을 도출하기 위해 '국가 연구데이터플랫폼'의 메타데이터 스키마 구조를 분석하고, 메타데이터의 데이터프로파일링 결과 등을 기반으로 비중요 요소를 제외한 후, 전체 스키마 요소 중 연구자가 직접 입력하는 요소인지 여부를 대표적인 선정 기준으로 삼아 메타데이터 요소를 선정하였다. 시스템에서 입력되는 경우 품질을 평가하는데 큰 의미가 없기 때문이다. 메타데이터 요소의 선정과정을 간략히 살펴보면 다음과 같다. 13개의 리포지터리 요소는 '국가 연구데이터플랫폼' 시스템 운영자 혹은 리포지터리 기관 담당자에 의해 등록될 요소이며, 연구자에 의해 입력되는 대상이 아니므로 매칭 대상에서 제외하였고, 41개의 프로젝트 요소도 대상에서 제외하였다. 프로젝트 관련 사항은 연구자가 연구데이터 등록 시 과제번호만을 입력하며, 이외에 과제 정보는 연구사업 계획 및 개시 단계에서 입력된 정보들을 승계하는 형태이기 때문이다.

그리고 61개의 데이터셋 요소, 36개의 파일 요소 중에도 스키마 구조상의 외래키(Foreign Key), 관리 및 서비스 목적의 메타데이터인 경우에는 연구자가 연구데이터를 등록할 때 기입하는 요소가 아니기 때문에 대상에서 제외하였으며, 파일 요소에서는 데이터셋 요소의 의미

가 중첩되어 데이터셋의 정보로 대체 가능한 요소의 경우에도 대상에서 제외하였다. 최종적으로, 데이터셋 요소 42개, 파일 요소 11개를 추출하였으며, 이렇게 선정된 요소를 TTA 표준을 기준으로 동일 성격 및 기능의 메타데이터 요소를 유형별로 범주화하여 〈표 1〉과 같이 22개의 데이터셋 요소와 9개의 파일 요소를 최종 선정하였다.

해당 선정 결과는 연구데이터 메타데이터의 품질평가 모델 개발을 위한 것이지만, '국가 연구데이터플랫폼'의 메타데이터 스키마 구조를 분석하고, 비중요 요소를 제외하여 메타데이터 요소를 선정하였기 때문에 본 연구에서도 이를 활용하였다. 따라서 본 연구에서는 〈표 1〉의 범주화 결과 요소들을 '국가 연구데이터플랫폼'의 크로스워크 대상으로 정의하였으며, 각 메타데이터 요소의 정의는 [부록 3]에서 확인할 수 있다.

〈표 1〉 '국가 연구데이터플랫폼'의 크로스워크 대상 요소

구분	국가 연구데이터플랫폼			범주화 결과(TTA 표준 매칭)		
	번호	요소명	요소_한글명	번호	요소명	요소_한글명
데이터셋	1	dataset_id	데이터셋ID	1	Identifier	식별자
	2	dataset_id_type	데이터셋ID유형			
	3	title	제목	2	Title	제목
	4	title_eng	영문제목			
	5	crtr	생성자	3	Creator	생성자
	6	crtr_eng	생성자영문명			
	7	crtr_pstinst	생성자소속기관			
	8	pblicte	발행처	4	Publisher	출판사
	9	pblicte_eng	발행처영문명			
	10	pblicte_year	발행연도	5	PublicationYear	출판연도
	11	cntrbtor	기여자	6	Contributor	기여자
	12	cntrbtor_eng_nm	기여자영문명			
	13	cntrbtor_ty	기여자유형			
	14	cntrbtor_org_nm	기여자기관명			
	15	cntrbtor_org_code	기여자기관코드			
	16	sj_cl_nm	주제분류명	7	Subject	주제
	17	ds_lclas	데이터셋-대분류			
	18	ds_mlsfc	데이터셋-중분류			
	19	ds_sclas	데이터셋-소분류			
	20	dc	설명	8	Description	설명
	21	dc_eng	영문설명			
	22	ds_kwrd	데이터셋키워드	9	Keyword	키워드
	23	ds_kwrd_eng	데이터셋영문키워드			
	24	charger_nm	담당자명	10	Contact	연락처
	25	charger_eng_nm	담당자명(영문)			
	26	charger_adres	담당자주소			
	27	charger_tlphon	담당자전화			
	28	charger_email	담당자이메일			
	29	dataset_ty	데이터셋유형	11	ResourceType	데이터의 유형

구분	국가 연구데이터플랫폼			범주화 결과(TTA 표준 매칭)		
	번호	요소명	요소_한글명	번호	요소명	요소_한글명
	30	etc_chartr_atrb	기타특성속성	12	ExtraAtributes	기타특성정보
	31	acces_author	접근권한	13	AccessType	접근유형
	32	embargo_de	엠바고일자	14	EmbargoDate	엠바고기한
	33	ds_lang	데이터셋언어	15	Language	언어
	34	rights	저작권	16	Rights	라이선스
	35	creat_de	생성일	17	CreateDate	생성일
	36	ver	버전	18	Version	버전
	37	locplc	소재지	19	Location of Publisher	소재지
	38	src_url	원천 URL	20	SourceIURL	원천정보 URL
	39	Coverage/Temporal	데이터 보유 기간	21	Coverage	수록범위
	40	Coverage/Spatial	데이터 수집 지역			
	41	Reference	관련 URL	22	Reference	관련정보 URL
파일	1	file_title	파일 제목	1	File_Title	파일 제목
	2	file_dc	파일설명	2	File_Description	파일 설명
	3	file_crtr	파일생성자	3	File_Creator	파일 생성자
	4	file_crtr_eng	파일생성자영문명			
	5	file_crtr_pstinst	파일생성자소속기관			
	6	file_creat_de	파일생성일	4	File_CreateDate	파일 생성일
	7	file_size	파일크기	5	File_Size	파일 크기
	8	file_fom	파일형식	6	File_format	파일 형식
	9	file_ty	파일유형	7	File_Type	파일 유형
	10	etc_chartr_file_atrb	기타특성파일메타	8	File_ExtraAttibutes	파일 기타특성정보
	11	file_src_url	파일원천	9	File_SoureceURL	파일 원천정보URL

3.3 바이오 연구데이터플랫폼

'바이오 연구데이터플랫폼'은 바이오 연구데이터를 보다 체계적으로 수집하고, 수집된 다양한 데이터를 통합 활용할 수 있도록, 데이터 등록 양식 국가 표준을 마련하기 위해 '바이오 연구데이터 표준화 위원회'를 구성하였다. 2020년 12월 '바이오 연구데이터 표준 등록 양식'을 첫 제정하였으며(과학기술정보통신부, 2020), 2021년 12월 표준 등록 양식을 업데이트하여 재개정하였다.

본 연구는 2021년부터 진행된 연구로 처음 제정된 첫 번째 표준 등록 양식을 대상으로 연구를 수행하였기 때문에 재개정된 표준 등록 양식이 반영되지 않았다.

'바이오 연구데이터플랫폼'에서는 정부 바이오 R&D를 통해 생산, 활용되는 모든 데이터를 수집하며, 그 대상은 바이오 주요 연구 분야인 신약, 의료기기 등 15대 바이오 연구 활동에서 필요로 하는 유전체, 이미지, 영상, 생화학분석, 표현형, 임상 및 전임상 데이터 등이다. 2020년 12월 첫 제정된 '바이오 연구데이터 표준 등록 양식'은 15대 분야 중 10개의 분야를 대상으로 작성되었다.

'바이오 연구데이터 표준 등록 양식'은 전 분야 공통(BioProject 정보), 다수 분야 공통(BioSample 정보, Omics 데이터) 양식을 비롯하여, Red 바이오, Green 바이오, White 바이오의 10개 분야에 대한 데이터를 등록할 수 있도록 〈표 2〉와 같이 크게 3개의 레벨로 구성되어 있다. 각 레벨의 하위에는 파트가 존재하고, 각 파트 별로 1개 이상의 메타데이터 요소가 있다.

전체 양식을 살펴볼 때, '국가 연구데이터플랫폼'에서 도출된 크로스워크 대상 요소들과의 매칭은 전 분야 공통 양식인 BioProject 정보와 다수 분야 공통 양식인 BioSample 정보, Omics 데이터를 대상으로 제한하여 진행하는 것이 적절한 것으로 판단되었다.

'국가 연구데이터플랫폼'은 과학기술 분야의 범용 연구데이터를 대상으로 하며, '국가 연구데이터플랫폼'의 크로스워크 대상 요소는 인문 사회 분야를 포함한 모든 범용적 연구데이터를 대상으로 하는 TTA 표준을 기준으로 범주화되었기 때문에, '바이오 연구데이터플랫폼'의 Red 바이오, Green 바이오, White 바이오와 같은 바이오 연구데이터의 세부 분야의 특성을 입력하는 메타데이터 항목과 매칭이 되기 어렵기 때문이다.

이러한 분석을 토대로 BioProject 정보와 BioSample 정보, Omics 데이터의 전체 277개 요소 중 필수 요소인 131개 메타데이터 요소가 크로스워크의 대상으로 선정되었다. 전체 요소들은 [부록 4]를 통해 확인할 수 있다.

〈표 2〉 바이오 연구데이터 표준 등록 양식

Level 1	Level 2	Level 3
전분야 공통	BioProject 정보	
다수 분야 공통	BioSample 정보	
	Omics 데이터	차세대 시퀀싱 데이터
		마이크로어레이 데이터
		염기서열 데이터
		대사체 데이터
		단백체 데이터
Red 바이오 분야	뇌과학 연구 분야	뇌 영상 데이터(매크로 이미징)
		뇌 영상 데이터(마이크로 이미징)
		기타 뇌과학 데이터
	의료기기	의료기기 연구 기본정보
		체외진단기기 연구 데이터
		의료영상 데이터
		의료기기 생체재료 정보
		생체신호 측정장치
	보건(질병 예방)	보건 연구 과제 정보
		보건 데이터 정보
	신약 연구 분야	독성 시험 데이터
		항암 약물의 임상 효능 데이터
		화합물 관련 데이터
		기허가약제 repositioning 연구 정보
		IND(investigational New Drug) 데이터
		바이오마커 데이터
Green 바이오 분야	종자 가축 연구 분야	가축 특성정보
		수산 특성정보
		작물 특성정보
		수목(야생식물)특성정보

Level 1	Level 2	Level 3
	수산 연구 분야	수산양식 데이터
		수산질병 데이터
		수산먹이생물 특성정보
		수산사료 특성정보
	동식물 치료제 연구 분야	동물줄기세포주 특성정보
		농약 정보
	지능형농업 분야	지능형 시설원예 분야
		지능형 축산 분야
	식품 연구 분야	식품 샘플 정보
		식품 성분 정보
		식품 기능성 정보
		식품 가공 정보
White 바이오 분야	환경바이오 연구 분야	곤충 분포 데이터
		DNA PCR 증폭용 프라이머 염기서열 데이터

'바이오 연구데이터 표준 등록 양식'의 BioProject 정보와 BioSample 정보, Omics 데이터에 대해 상세하게 살펴보면 다음과 같다.

3.3.1 BioProject 정보

BioProject 정보는 바이오 연구데이터를 등록할 때, 해당 연구 프로젝트에 대한 개괄적인 정보(제목, 과제, 논문 등)를 등록할 때 쓰이며, 해당 표준에 수록된 모든 각 데이터별 양식을 작성하기 전 특정 데이터 타입에 상관없이 첫 번째로 작성해야 하는 양식이다. 'Name of submitter(등록자의 영문이름)', 'Submission date(제출 날짜)' 등의 요소들로 구성되어 있으며 [부록 4-1]에서 상세한 항목을 확인할 수 있다. 본 연구에서는 필수요소만을 대상으로 했기 때문에, 전체 27개 요소 중 21개 요소가 대상이 되었다.

3.3.2 BioSample 정보

BioSample 정보는 바이오 연구데이터를 등록할 때, 생물 유래 시료를 대상으로 실험을 했을 경우 그 시료에 대한 개괄적인 정보를 등록할 때 쓰이며, Red, Green, White 분야에 수록되어 있는 다수의 각 데이터별 양식을 작성하기 전 필수로 작성해야 하는 양식이다. 즉, 모든 데이터에 적용되는 것은 아니며, 적용되는 데이터는 각 섹션에 별도 표시되어 있다.

각 항목들은 연구에 사용된 생물 샘플 관련 정보(생물종명, 샘플명, 각종 속성, 처리 조건 등)를 다루는데, 생물유래시료를 대상으로 실험을 했을 경우, 샘플이 유래한 생물군에 따라 서로 다른 항목들로 구성된다. 생물군은 인간, 모델생물 및 동물, 무척추동물, 식물, 미생물, 바이러스, 임상병원체, 환경・식품 및 기타 병원체, 메타게놈 및 환경 샘플로 9개의 생물군으로 구분되며, 각 항목들이 해당 생물군에서 의미가 있는지 여부에 따라 입력 필요성이 달라진다. 이처럼 샘플 특성은 샘플의 종류에 따라 입력할 값들이 달라지기 때문에, 본 연구에서는 1번 이상 필수 항목으로 선정된 요소들만 대상으로 하였다. 이에 전체 요소 80개 중 16개 요소를 대상으로 추출하였으며, [부록 4-2]에서 상세한 항목을 확인할 수 있다.

3.3.3 Omics 데이터

생물 유래 데이터에 대한 총체적인 특성에 대한 정보를 제공하는 Omics 데이터는 차세대시퀀싱(Next-Generation Sequencing, 이하 NGS) 데이터, 마이크로어레이(Microarray) 데이터, 염기서열(Nucleotide sequence) 데이터, 대사체(Metabolomics) 데이터, 단백체(Proteomics) 데이터와 같은 5가지 유형으로 구성되어 있다.

Omics는 전체를 뜻하는 말인 옴(-ome)과 학문을 뜻하는 접미사 익스(-ics)가 결합된 말로, 어떤 특정 학문 분야를 말하기보다는 개별 유전자(gene), 전사물(transcript), 단백질(protein), 대사물(metabolite) 연구에 대비되는 총체적인 개념의 데이터 세트를 바탕으로 하는 생물학 분야라고 할 수 있다. '~옴'은 우리말로 전체를 지칭하는 '~체'로 번역된다.

Omics는 인간 유전체 사업(human genome project) 이후 새롭게 등장한 학문 분야로 몇 가지 측면에서 전통적 생물학 연구와 대비된다. Omics에서 다루는 데이터는 대규모-대용량(high-throughput) 기술들로 생산되기 때문에 개별 물질을 연구 대상으로 하는 전통적인 생물학과 달리 데이터의 전산학적 처리가 필수적이다. 또 수학적, 통계적 기법이 연구에 적극 활용된다. [부록 4-3]에서 상세한 항목을 확인할 수 있다.

1) NGS 데이터

NGS 데이터는 DNA 및 RNA 서열의 NGS 기반 high-throughput 시퀀싱 데이터를 등록할 때 쓰이는 양식으로, 해당 양식은 NCBI의 SRA(Sequence Read Archive)을 기준으로 작성되었다. NGS 데이터는 전체 12개 요소 모두 필수값으로 12개가 대상이 된다.

2) 마이크로어레이 데이터

마이크로어레이 데이터는 DNA 및 RNA를 microarray로 profiling한 실험 데이터로, 전사체(gene expression array), 후성유전체(methylation array), 유전체(SNP array)등 모든 종류의 array를 포함한다. 해당 양식은 NCBI GEO(Gene Expression Omnibus)를 기준으로 작성되었으며, 본 연구에서는 전체 19개 요소 중 17개 요소를 대상으로 한다.

3) 염기서열 데이터

염기서열 또는 핵산의 1차 구조(Nucleic Acid Primary Structure)는 DNA의 기본단위 뉴클레오타이드의 구성성분 중 하나인 핵염기들을 순서대로 나열해 놓은 것을 말한다. 유전자는 생물의 유전형질을 결정하는 단백질을 지정하는 기본적인 단위로, 지구상의 모든 생명체들은 염기서열을 통해 단백질을 지정하는 원리를 따른다. DNA상에서 염기가 일렬로 3개씩 모이면 하나의 트리플렛 코드를 형성하여 하나의 아미노산을 지정하게 되는데, 이 트리플렛 코드들이 여러 개 모이면 궁극적으로 하나의 단백질을 지정하게 된다. 즉, 염기가 3개씩 모이면 트리플렛 코드를 형성하여, 단백질 서열로 변환되는 것이다. 해당 양식은 NCBI GenBank를 기준으로 작성되었으며, 본 연구에서는 전체 39개 중 필수 요소인 26개 요소를 대상으로 한다.

4) 대사체 데이터

대사체는 세포, 조직, 체액과 같은 생물학적 시료 내에 존재하는 대사물질들의 총체를 의미한다.

일반적으로 1500달톤(Dalton) 이하의 물질들로 구성되어 있고 생체 내 대사물질(endogenous metabolome: 아미노산, 핵산, 지방산, 당류, 아민류, 당지질, 짧은 펩티드, 비타민, 호르몬 등)과 생체 외 대사물질(exogenous metabolome: 약물, 음식, 첨가제, 독성 물질 등)로 분류된다. 해당 양식은 EBI MetaboLights를 기준으로 작성되었다. 대사체 데이터는 전체 65개 요소 중 19개 요소만이 필수 요소로, 19개 요소를 대상으로 하였다.

5) 단백체 데이터

단백체(단백질체) 또는 프로테옴(proteome)은 세포 내의 단백질의 총합을 뜻한다. 1970년대 영국 케임브리지의 프레드 생어가 세포 내의 유전자들이 모두 단백질로 발현되는지를 알고 싶어서 추진한 단백질 동정에서 비롯된 학문이다. 이와 연관된 말들이 유전체, 전사체, 상호작용체 등이 있다. 단백체의 해석은 1960-1970년대에는 느린 화학적 서열해석법으로 했으나, 질량분석기가 등장함에 따라, 2D젤과 질량분석기를 혼용한 방법이 널리 쓰이게 되었다. 해당 양식은 EBI PRIDE를 기준으로 하여 작성되었으며, 전체 35개 요소 중 필수 요소인 20개 요소를 대상으로 한다.

4. 메타데이터 요소 크로스워크

4.1 크로스워크 수행

'국가 연구데이터플랫폼'의 메타데이터 요소 크로스워크 대상은 22개의 데이터셋 요소와 9개의 파일 요소로 총 31개의 요소이며 이를 '기준 요소'로 삼았다. '바이오 연구데이터플랫폼'의 메타데이터 요소 크로스워크 대상은 선택 입력 요소를 제외한 필수 입력 요소로, 전 분야 공통 양식인 BioProject 정보의 21개 요소, 다수 분야 공통 양식인 BioSample 정보의 16개 요소, Omics 데이터의 하위요소인 NGS 데이터에서 12개 요소, 마이크로어레이 데이터에서 17개 요소, 염기서열 데이터에서 26개 요소, 대사체 데이터에서 19개 요소, 단백체 데이터에서 20개 요소로, 총 131개의 요소이며 본 연구에서는 이를 '매칭 요소'로 명명하였다. 〈표 3〉은 이상의 메타데이터 요소들을 대상으로 항목별 정의를 기준으로 하여 크로스워크를 진행한 결과이다.

〈표 3〉 크로스워크 결과

번호	구분	기준 요소	매칭 요소
1	데이터셋	Identifier	Grant ID, Grant NTIS ID, Library ID, Confirm update of existing submission, Existing genome accession
2		Title	Grant title, Grant title in Korean, organism, sample name, Title(1), Title(2), detailed sample title
3		Creator	biomaterial provider, collected by
4		Publisher	Project title, Project title in Korean
5		PublicationYear	Submission date
6		Contributor	Name of submitter, Name of submitter in Korean, Name of submitter's organization, Name of submitter's organization in Korean, Department of submitter, Department of submitter in Korean, Funding agency, host

번호	구분	기준 요소	매칭 요소
7		Subject	Project description, Project description in Korean
8		Description	Sample type(1), host disease, isolate, sex, tissue, Library strategy, Library source, Library selection, Library layout, Platform, Instrument model, Design description, Summary, Overall design, source name, characteristics: tag, molecule, label, description, platform, extract protocol, label protocol, hybridization protocol, scan protocol, data processing, Assembly method, Program version, Genome coverage, Total raw read length, Read throughput, The number of contigs, Contig length, N50, Sequencing technology, Confirm full genome, Description for subset of the genome, Confirm final version, Confirm de novo assembly, Reference assembly name or accession, Molecule Type, Topology, Source Organelle/Location information, Chimera check, Chimera check program name, Cultured or Uncultured, Primer Type, Submission Set/Batch, Additional characteristic, Sample collection protocol, Extraction protocol, Technique type, Assay type, Assay definition, Chromatography protocol, Mass spectrometry protocol, NMR sample protocol, NMR spectroscopy protocol, NMR assay protocol, Data transformation protocol, Metabolite identification protocol, Calibration standard, Subcellular, Experiment type, Sample processing protocol, Data processing protocol, Search parameters, Instrument, Modification, Enzyme, LC system, PTM, Sample fraction, Fractionation, plex
9		Keyword	Keywords(1), Factors, Keywords(2)
10		Contact	Primary e-mail of submitter, Address of submitter, Address of submitter in Korean, Country of submitter
11		ResourceType	sample type(2)
12		ExtraAttributes	age, Add features, Period of Creation, Submission Type
13		AccessType	-
14		EmbargoDate	Release date
15		Language	-
16		Rights	-
17		CreateDate	collection date
18		Version	-
19		Location of Publisher	geographic location, latitude and longitude
20		SourceIURL	-
21		Coverage	isolation source
22		Reference	-
23	파일	File_Title	Filename
24		File_Description	Reference
25		File_Creator	-
26		File_CreateDate	-
27		File_Size	-
28		File_format	FASTA file, Raw, Peak list, Search list, FASTA
29		File_Type	Filetype
30		File_ExtraAttibutes	raw data file, processed data file, Acquired raw data files, Data processed peak table, Concentration data
31		File_SourceURL	-

크로스워크 결과 '국가 연구데이터플랫폼'의 31개 '기준 요소' 중 21개 요소와 '매칭 요소'의 요소들이 매칭되었다. '기준 요소'는 TTA 표준을 기준으로 동일 성격 및 기능의 메타데이터 요소를 유형별로 범주화한 요소들로 매칭을 진행하였기 때문에, 모든 요소들에 고르게 매

칭이 되지 않았고, 매칭 결과가 약간 부자연스러운 경우가 있다. 그러나 전체 요소를 1:1로 매칭하기에는 한계가 있기 때문에, 세부 요소들까지의 전체 요소에 대한 매칭이 이루어지지 못한 제한사항은 후속 연구를 통해 정밀한 조정을 거쳐 보완될 필요가 있다.

4.2 크로스워크 결과 검증을 통한 최종 크로스워크 결과 도출

메타데이터 크로스워크 결과에 대해 S 대학과 C 대학 소속 바이오 분야 연구자 7인을 대상으로 검증을 실시하였다. 이는 바이오 분야 전문성을 가진 검증자를 통해 매칭된 요소의 결과를 검증하고 더 적절한 매칭 요소의 추천을 통해 크로스워크 결과를 보완할 수 있도록 하기 위한 것이다.

검증자들은 PEET(약학대학입문자격시험)을 통과하여 약학대학 학부를 졸업한 후, 약학대학원에 재학 중인 박사 또는 석사과정생으로 바이오 분야에서 최소 5년 이상의 연구 경력을 가진 연구자들을 대상으로 섭외하였다. 바이오 분야의 범위가 다양하나 약학을 전공한 연구자를 대상으로 한 이유는 본 연구가 초기 설계 단계에서 Red 바이오 분야의 신약 연구 분야를 대상으로 크로스워크를 수행하려고 설계되었기 때문이다. 연구를 진행하면서 크로스워크 대상 요소들과의 매칭 대상 요소를 전 분야 공통 양식인 BioProject 정보와 다수 분야 공통 양식인 BioSample 정보, Omics 데이터로 제한하게 되었으며, 검증 참여자들은 이 검증에 필요한 바이오 분야 배경지식을 충분히 갖춘 연구자들로 구성되었다.

검증자들에게 이메일을 통해 검증 자료와 설명 자료를 함께 배포하여, 해당 검증의 목적과 방법을 상세히 설명하고 검증할 수 있도록 하였다. 검증은 2022년 2월 4일에서 10일까지 총 7일간 이루어졌으며, 7명의 검증자에게 모두 회신을 받았다(질문지 샘플 및 검증자 소속기관 [부록 5] 참조). 검증자들의 검증 내용을 반영한 검증 결과는 〈표 4〉와 같으며, 매칭 요소가 수정된 경우 이탤릭체로 표시하였고, 상세한 매칭 결과는 [부록 6]에서 확인할 수 있다.

〈표 4〉 검증 결과

번호	구분	기준 요소	매칭 요소
1	데이터셋	Identifier	Grant ID, Grant NTIS ID, Library ID, Existing genome accession
2		Title	*Project title*, *Project title in Korean*, Grant title, Grant title in Korean, Title(1), Title(2)
3		Creator	collected by
4		Publisher	*Name of submitter*, *Name of submitter in Korean*
5		PublicationYear	Submission date
6		Contributor	Funding agency, *biomaterial provider*
7		Subject	*detailed sample title*
8		Description	*Project description*, *Project description in Korean*, *geographic location*, *host*, host disease, isolate, *isolation source*, *latitude and longitude*, sex, tissue, Library strategy,

번호	구분	기준 요소	매칭 요소
			Library source, Library selection, Library layout, Platform(1), Instrument model, Design description, Summary, Overall design, source name, molecule, description, platform(2), extract protocol(1), label protocol, hybridization protocol, scan protocol, data processing, Assembly method, Genome coverage, Total raw read length, Read throughput, The number of contigs, Contig length, N50, Sequencing technology, Confirm full genome, Description for subset of the genome, Confirm de novo assembly, Reference assembly name or accession, Molecule Type, Topology, Source Organelle/Location information, Chimera check, Chimera check program name, Cultured or Uncultured, Primer Type, Submission Set/Batch, *Factors*, Sample collection protocol, Extraction protocol(2), Technique type, Assay type, Assay definition, Chromatography protocol, Mass spectrometry protocol, NMR sample protocol, NMR spectroscopy protocol, NMR assay protocol, Data transformation protocol, Metabolite identification protocol, Calibration standard, Subcellular, Sample processing protocol, Data processing protocol, Search parameters, Modification, Enzyme, LC system, PTM, Sample fraction, Fractionation, plex
9		Keyword	*Sample type(1), organism, sample name, characteristics: tag, label*, Keywords(1), Keywords(2)
10		Contact	Primary e-mail of submitter
11		ResourceType	sample type(2), *Experiment type*
12		ExtraAttributes	*Address of submitter, Address of submitter in Korean, Country of submitter,* age, Add features, *Additional characteristic*, Submission Type, *Instrument*
13		AccessType	-
14		EmbargoDate	Release date
15		Language	-
16		Rights	-
17		CreateDate	collection date, *Period of Creation*
18		Version	*Program version, Confirm final version, Confirm update of existing submission*
19		Location of Publisher	*Name of submitter's organization, Name of submitter's organization in Korean, Department of submitter, Department of submitter in Korean*
20		SourceIURL	-
21		Coverage	-
22		Reference	-
23	파일	File_Title	Filename
24		File_Description	Reference
25		File_Creator	-
26		File_CreateDate	-
27		File_Size	-
28		File_format	FASTA file, Raw, Search list, FASTA
29		File_Type	Filetype
30		File_ExtraAttibutes	raw data file, processed data file, Acquired raw data files, Data processed peak table, Concentration data, *Peak list*
31		File_SourceURL	-

본 연구에서는 검증을 양적 연구의 방법인 설문 형태로 진행하지 않고, 질적 연구의 방법으로 검증을 진행하였기 때문에, 검증 결과를 취합할 때 기준을 세워 진행하였다.

7명의 검증자가 모두 동일한 의견을 제시하기는 쉽지 않기 때문에, 2명 이상의 검증자가 같은 요소로 수정을 제안한 경우에는 해당 요소로 '매칭 요소'를 수정하였고, 각각 다른 요소로 수정을 제안한 경우 검증자들을 대상으로 재검토를 요청하여 1개 요소를 최종 선정하였다.

검증 결과 33개의 '매칭 요소'가 다른 '기준 요소'와 매칭되도록 수정되었다. 일부 결과를 보면, 기존에 '기준 요소' Contributor에 매칭되었던 Name of submitter, Name of submitter in Korean는 '기준 요소' Publisher에 재배치되었음을 확인할 수 있는데, 이는 Publisher가 '데이터셋을 웹에 공개 또는 출판하는 주체'로 정의되기 때문에 Contributor보다는 Publisher가 적절하다는 검증 결과가 반영된 것이다.

또한 기존에 '기준 요소' Title에 매칭되었던 detailed sample title은 '기준 요소' Subject에 재배치되었다. '매칭 요소'가 상위 레벨의 요소들과 함께 명기되지 않고 해당 요소명으로만 표현되어 있어서, 요소명 자체로 볼 때는 detailed sample title이 기존에 매칭된 결과가 맞게 보일 수 있으나, detailed sample title은 '샘플을 설명할 수 있는 상세한 제목'이라고 정의되며, 'Omics 데이터' - '마이크로어레이데이터' - 'Sample detail'이라는 상위 요소들을 가지며([부록 4-3] 참조), 바이오 분야에서 연구데이터를 검색하는 연구자의 입장에서는 해당 요소가 주제로서 검색할 수 있는 대상이 되기 때문에 '기준 요소' Subject와의 매칭이 더 적절하다.

'매칭 요소'가 재배치되지는 않았지만, 위와 같은 맥락으로 '기준 요소' File_Description에 매칭된 Reference의 경우에도 용어 자체로 보면 '기준 요소' Reference에 매칭 가능해 보이지만, 해당 '매칭 요소'의 상위 요소들을 보면 'Omics 데이터' - '차세대시퀀싱(NGS) 데이터' - '파일'로 이루어져 있고([부록 4-3] 참조), Reference 즉, 참조 서열은 NGS 데이터의 파일을 도출할 때 참조가 되는 잘 알려진 서열을 의미하는 정보로 정의되며, 파일을 설명하는 요소로 볼 수 있기 때문에, '기준 요소'의 File_Description과 매칭되는 것이 적합하다.

또한 전체 '매칭 요소'들이 31개 '기준 요소' 중 21개 요소와 매칭되었다. 이는 크로스워크 매칭 결과와 매칭된 요소의 수는 동일하지만 매칭된 요소의 항목이 다르다. 일부 결과를 살펴보면 크로스워크 결과에서는 '기준 요소' Version과 매칭된 '매칭 요소'가 없었으나, 검증 결과에서는 3개의 '매칭 요소'와 매칭되었으며, '기준 요소' Coverage와 매칭되었던 isolation source는 '기준 요소' Description과 매칭되는 것으로 수정된 것을 확인할 수 있다.

'기준 요소' Description의 경우 일부 요소들의 변경이 있었으나, 여전히 73개의 '매칭 요소'가 매칭되었다. '기준 요소'의 항목들의 수가 '매칭 요소'의 항목들의 수보다 현저하게 적은 상황에서 포괄적인 의미를 가진 Description과 같은 요소에 많은 요소들이 매칭될 수 밖에 없기 때문이다. 이에 대해 검증자들은 해당 요소를 추가적으로 구분하여 '기준 요소' Description을 세분함으로써 Library strategy, extract protocol, label protocol, data processing 등의 요소들은 '연구에 사용된 프로토콜(실험 또는 연구 방법)'

의 항목으로 구분하고, Confirm full genome, Description for subset of the genome, Confirm de novo assembly와 같은 요소들은 '분석 대상에 대한 정보'의 항목으로 구분할 것을 제안하였다.

이러한 한계를 극복하기 위해서는 '국가 연구데이터플랫폼'의 메타데이터 스키마와 '바이오 연구데이터플랫폼'의 메타데이터 스키마 담당자들 간의 충분한 협의를 바탕으로 세밀한 매칭이 필요하며, 필요시 담당자 및 메타데이터 전문가의 의견을 반영하여 '국가 연구데이터플랫폼'의 메타데이터 스키마의 조정도 필요한 것으로 파악되었다.

5. 결 론

'국가 연구데이터플랫폼'과 '바이오 연구데이터플랫폼'의 메타데이터 스키마가 각기 다른 기준을 참고하여 지속적으로 업데이트되고 있는 상황에서, '국가 연구데이터플랫폼'을 통해 각 전문센터의 연구데이터가 취합되어 연결망 체계를 갖추기 위해서는 상호운용성을 확보할 수 있는 체계의 구축이 필수적이다. 본 연구에서는 '국가 연구데이터플랫폼'과 '바이오 연구데이터플랫폼'의 메타데이터 표준을 분석하고, 크로스워크 대상을 선정하여 매칭을 한 후 검증과 보완을 거쳐 최종 크로스워크 결과를 도출하였다. 최종적으로 크로스워크 대상으로 선정된 '국가 연구데이터플랫폼'의 31개 요소와 '바이오 연구데이터플랫폼'의 131개 요소가 모두 매칭되었으며, 최종 매칭 결과는 [부록 6]에서 확인할 수 있다.

본 연구를 통해 바이오 분야에서 쓰이는 요소와 '국가 연구데이터플랫폼'의 요소가 매칭되어 의미적으로 연결될 수 있는 가능성을 밝혔으며, 두 플랫폼의 상호운용성을 확보하기 위한 기반을 확인할 수 있었다.

융합연구의 중요성이 강조되는 오늘날, 분야별로 각각 구축된 연구데이터플랫폼은 이른바 부서 간 장벽을 일컫는 사일로에 갇히게 되는 문제를 발생시킬 수 있다. 따라서 바이오 분야의 연구데이터 메타데이터를 '국가 연구데이터플랫폼'과 적절하게 연결하여 바이오 분야의 연구데이터를 '국가 연구데이터플랫폼'에서 취합하여 제공할 수 있는 후속 연구가 진행된다면, 연구자들은 '연구데이터플랫폼'에서 본인의 연구 분야 이외의 다양한 분야의 연구데이터를 발견하게 됨으로써, 발전된 후속연구와 융합연구가 활성화될 수 있을 것이다.

참 고 문 헌

강주연 (2017). 생명공학분야의 연구데이터 관리 현황 및 개선 방안에 관한 연구. 석사학위논문, 전북대학교 일반대학원 기록관리학과.

고영만 (2019). 국가연구데이터플랫폼의 메타데이터 품질 및 유용성 평가 모델 연구. 한국과학기술정보연구원.

고영만, 서태설, 임태훈 (2007). 의미 호환을 위한 메타데이터 매핑 연구. 정보관리학회지, 24(4), 223-238. https://doi.org/10.3743/KOSIM.2007.24.4.223

국가연구개발사업의 관리 등에 관한 규정. 대통령령 제31297호

김선 (2022). 생명공학 분야 연구자의 연구데이터 공유 의도에 영향을 미치는 요인에 관한 연구: 학술적 평판의 조절효과를 중심으로. 정보관리학회지, 39(1), 45-68. https://doi.org/10.3743/KOSIM.2022.39.1.045

김성욱, 김선태 (2020). 응집물질물리분야 연구데이터 관리 방안 연구. 정보관리학회지, 37(3), 77-106. https://doi.org/10.3743/KOSIM.2020.37.3.077

김주섭, 김선태, 최상기 (2020). Geoscience 연구데이터 관리를 위한 기능별 세부 요소 및 중요도에 관한 연구. 한국문헌정보학회지, 54(1), 411-440. http://dx.doi.org/10.4275/KSLIS.2020.54.1.411

김주섭, 윤희남, 권용수, 김선태 (2020a). 생태 분야 연구데이터를 위한 메타데이터 설계: DCAT을 중심으로. 한국도서관·정보학회지, 51(4), 249-278. https://doi.org/10.16981/kliss.51.4.202012.249

김주섭, 한연중, 유원재, 김선태 (2020b). 임산공학 분야 연구데이터 관리를 위한 메타데이터 설계에 관한 연구. 한국문헌정보학회지, 54(4), 169-194. http://dx.doi.org/10.4275/KSLIS.2020.54.4.169

남태우, 이승민 (2014). 정보자원의 기술과 메타데이터. 서울: 한국도서관협회, 233-234.

대한민국. 과학기술정보통신부 (2020), 바이오 연구 데이터 표준 등록 양식.

박미영, 안인자, 김준모 (2018). 생명공학분야의 연구데이터 공유 사례에 관한 연구. 한국비블리아학회지, 29(1), 393-416. https://doi.org/10.14699/kbiblia.2018.29.1.393

생명연구자원의 확보·관리 및 활용에 관한 법률. 법률 제16016호

연구데이터 관리 및 공유를 위한 메타데이터. TTAK.KO -10.0976

예상준, 장호, 김선태 (2019). 한의학 연구 데이터 관리 및 공유를 위한 메타데이터 요소 설계. 한국문헌정보학회지, 53(2), 223-246. http://doi.org/10.4275/KSLIS.2019.53.2.223

이미화, 이은주, 노지현 (2020). 연구데이터 관리를 위한 OAK 메타데이터 확장 방안 연구. 한국도서관·정보학회지, 51(3), 27-51. http://doi.org/10.16981/kliss.51.3.202009.27

최기영, 성윤모, 조명래, 박영선, 김경규, 김현수, 박능후, 문성혁, 이의경, 박종호 (2020). 제3차 국가생명연구자원 관리·활용 기본계획('20~'25)(안). 국가과학기술자문회의 심의회의.

한국과학기술정보연구원 (2018). 연구데이터 공유·확산체제 구축. 과학기술정보통신부.

UK Data Archive (2011). Managing and Sharing Data: Best Practice For Researchers. UK: UK Data Archive

• **국문 참고문헌에 대한 영문 표기**
(English translation of references written in Korean)

Act on Securing, Management and Utilization of Bioresearch Resources. Law No. 16016

Choi, Gi-young, Sung, Yun-Mo, Cho, Myoung-rae, Park, Youngsun, Kim, Kyung-Kyu, Kim Hyeonsu, Park, NeungHoo, Moon, Seong-Hyeok, Lee, Eui-Kyung, & Park, Jong-Ho (2020). The 3rd National Life Research Resource Management and Utilization Basic Plan (2020~'25) (draft). Council of National Science and Technology Advisory Council.

Gang, Ju-Yeon (2017). A Study on Analysis of the Current Status and the Method to Improve Biotechnology Research Data Management. Master's thesis, Graduate School of Archives and Records Management, Chonbuk National University.

Kim, Juseop, Han, Yeonjung, Youe, Wonjae, & Kim, Suntae (2020b). A study on the design of metadata for research data management in forestry engineering. Journal of the Korean Society for Library and Information Science, 54(4), 169-194. http://dx.doi.org/10.4275/KSLIS.2020.54.4.169

Kim, Juseop, Kim, Suntae, & Choi, Sangki (2020). A study on functional details and importance of geoscience research data management. Journal of the Korean Society for Library and Information Science, 54(1), 411-440. http://dx.doi.org/10.4275/KSLIS.2020.54.1.411

Kim, Juseop, Yoon, Heenam, Kwon, Yong-su, & Kim, Suntae (2020a). Metadata design for ecological research data: focused on DCAT. Journal of Korean Library and Information Science Society, 51(4), 249-278. https://doi.org/10.16981/kliss.51.4.202012.249

Kim, Sun (2022). An exploratory study of biotechnology scientists' research data sharing intention: the moderating effects of academic reputation. Journal of the Korean Society for Information Management, 39(1), 45-68. https://doi.org/10.3743/KOSIM.2022.39.1.045

Kim, Sungwook & Kim, Suntae (2020). A study on the research data management methods for the condensed matter physics. Journal of the Korean Society for Information Management, 37(3), 77-106. https://doi.org/10.3743/KOSIM.2020.37.3.077

Ko, Young-Man (2019). Study on a Model for Metadata Quality and Usability Assessment of the National Research Data Platform. Korea Institute of Science and Technology Information.

Ko, Young-Man, Seo, Tae-sul, & Lim, Tae-Hoon (2007). A study on metadata mapping for semantic interoperability. Journal of the Korean Society for Information Management, 24(4), 223-238. https://doi.org/10.3743/KOSIM.2007.24.4.223

Korea Institute of Science and Technology Information (2018). Establishing a system for sharing

and disseminating research data.

Lee, Mihwa, Lee, Eun-Ju, & Rho, Jee-Hyun (2020). A preliminary study on extending OAK metadata for research data. Journal of Korean Library and Information Science Society, 51(3), 27-51. http://doi.org/10.16981/kliss.51.3.202009.27

Nam, Taewoo & Lee, Seungmin (2014). Resource Description and Metadata. Seoul: Korean Library Association.

Park, Miyoung, Ahn, Inja, & Kim, Junmo (2018). A study on use case of research data sharing in biotechnology. Journal of the Korean BIBLIA Society for library and Information Science, 29(1), 393-416. https://doi.org/10.14699/kbiblia.2018.29.1.393

Republic of Korea. Ministry of Science and ICT (2020). Bio Research Data Standard Registration Form.

The Integrated metadata for the scientific data. TTAK.KO - 10.0976

Yea, Sang-Jun, Jang, Ho, & Kim, Suntae (2019). Metadata element design for Korean medicine research data management and re-use. Journal of the Korean Society for Library and Information Science, 53(2), 223-246. http://doi.org/10.4275/KSLIS.2019.53.2.223

[부록 1] TTA 표준

리포지터리 메타데이터		컬렉션 메타데이터		데이터셋 메타데이터		파일 메타데이터	
ID	요소명	ID	요소명	ID	요소명	ID	요소명
R1	Repository	C1	Collection	D1	Dataset	F1	File
R2	Repository URL	C2	Identifier	D2	Identifier	F2	Identifier
R3	Identifier	C2.1	Identifier Type	D2.1	Identifier Type	F2.1	Identifier Type
R3.1	Identifier Type	C3	Title	D3	Title	F3	Title
R4	Repository Name	C3.1	Title Type	D3.1	Title Type	F3.1	Title Type
R4.1	Repository Name Type	C4	Date	D4	Creator	F4	Creator
R5	Type	C4.1	Date Type	D5	Publisher	F5	Publisher
R6	Repository Language	C5	Description	D6	Publication Year	F6	Publication Year
R7	Subject	C6	Subject	D7	Date	F7	Contributor
R7.1	Subject Scheme	C6.1	Subject Scheme	D7.1	Date Type	F7.1	Contributor Type
R7.2	Subject ID	C6.2	Subject ID	D8	Description	F8	Date
R7.3	Subject Name	C6.3	Subject Name	D9	Subject	F8.1	Date Type
R8	Institution Name	C7	Creator	D9.1	Subject Scheme	F9	Description
R9	Institution Country	C8	Contact	D9.2	Subject ID	F10	Subject
R10	Database Access Type	C9	Rights	D9.3	Subject Name	F10.1	Subject Scheme
R11	Data Access Type	C10	Keyword	D10	Contributor	F10.2	Subject ID
R12	Data License Name	C11	Access Type	D10.1	Contributor Type	F10.3	Subject Name
R13	Data License Url	C12	Access Restriction	D11	Contact	F11	Contact
R14	Data Upload			D12	Rights	F12	Rights
R15	Versioning			D13	Keyword	F13	Keyword
R16	Enhanced Publication			D14	Access Type	F14	Access Type
R17	Quality Management			D15	Access Restriction	F15	Access Restriction
R18	Description					F16	Coverage
R19	Responsibility Type					F17	Type
R20	Institution Contact					F18	Format
R21	Repository Contact					F19	Size
						F19.1	Unit

[부록 2] 국가연구데이터플랫폼 메타데이터 표준

리포지터리 메타데이터			프로젝트 메타데이터		
번호	요소명	요소_한글명	번호	요소명	요소_한글명
1	rpstr_id	리포지터리ID	1	task_id	과제ID
2	rpstr_nm	리포지터리명	2	task_id_type	과제ID유형
3	rpstr_url	리포지터리URL	3	rpstr_id	리포지터리ID
4	rpstr_repic_nm	리포지터리대체명	4	detail_task_no	세부과제번호
5	rpstr_ty	리포지터리유형	5	task_nm	과제명
6	rpstr_dc	리포지터리설명	6	task_eng_nm	과제영문명
7	rpstr_instt	리포지터리기관명	7	task_rspnber_nm	과제책임자명
8	rpstr_lang	리포지터리언어	8	task_rspnber_eng_nm	과제책임자영문명
9	rpstr_nation	리포지터리국가	9	task_rspnber_rsrch_no	과제책임자과기번호
10	rgst_id	등록자ID	10	task_rspnber_orgnzt_nm	과제책임자기관명
11	rgsde	등록일자	11	mngt_rrcs_nm	주관연구기관명
12	updusr_id	수정자ID	12	mngt_rrcs_eng_nm	주관연구기관영문명
13	updt_de	수정일자	13	mngt_rrcs_code	주관연구기관코드
			14	prtcpnt	참여자
			15	prtcpnt_eng_nm	참여자영문명
			16	prtcpnt_rsrch_no	참여자과학기술인등재번호
			17	prtcpnt_orgnzt_nm	참여자의기관명
			18	base_year	기준년도
			19	rsrch_bgnde	연구시작일
			20	rsrcct	연구비
			21	rsrch_endde	연구종료일
			22	task_manage_orgnzt	과제관리기관
			23	task_fund_orgnzt_nm	과제펀딩기관명
			24	stsc_lclas	과학기술표준분류-대분류
			25	stsc_mlsfc	과학기술표준분류-중분류
			26	stsc_sclas	과학기술표준분류-소분류
			27	kwrd	키워드
			28	kwrd_eng	키워드영문
			29	lang	언어
			30	nation	국가
			31	cttpc_nm	담당자명
			32	cttpc_eng_nm	담당자명(영문)
			33	cttpc_adres	담당자주소
			34	cttpc_tlphon	담당자전화
			35	cttpc__email	담당자이메일
			36	dc	설명
			37	dmp_uri	DMP_URL
			38	rgst_id	등록자ID
			39	rgsde	등록일자
			40	updusr_id	수정자ID
			41	updt_de	수정일자

데이터셋 메타데이터			파일 메타데이터		
번호	요소명	요소_한글명	번호	요소명	요소_한글명
1	dataset_id	데이터셋ID	1	file_id	파일ID
2	dataset_id_type	데이터셋ID유형	2	File/Identifier_type	파일ID 유형
3	task_id	과제ID	3	dataset_id	데이터셋ID
4	title	제목	4	file_title	파일 제목
5	title_eng	영문제목	5	file_eng_title	파일영문제목
6	register	등록자명	6	file_dc	파일설명
7	register_pstinst	등록자소속기관	7	file_register	파일등록자명
8	crtr	생성자	8	file_register_pstinst	파일등록자소속기관
9	crtr_eng	생성자영문명	9	file_crtr	파일생성자
10	crtr_pstinst	생성자소속기관	10	file_crtr_eng	파일생성자영문명
11	pblicte	발행처	11	file_crtr_pstinst	파일생성자소속기관
12	pblicte_eng	발행처영문명	12	file_creat_de	파일생성일
13	pblicte_year	발행연도	13	file_updt_de	파일수정일
14	cntrbtor	기여자	14	file_kwrd	파일키워드
15	cntrbtor_eng_nm	기여자영문명	15	file_kwrd_eng	파일영문키워드
16	cntrbtor_ty	기여자유형	16	file_size	파일크기
17	cntrbtor_org_nm	기여자기관명	17	file_fom	파일형식
18	cntrbtor_org_code	기여자기관코드	18	file_ty	파일유형
19	sj_cl_nm	주제분류명	19	etc_chartr_file_atrb	기타특성파일메타
20	ds_lclas	데이터셋-대분류	20	atmc_extrc_file_atrb	자동추출파일속성
21	ds_mlsfc	데이터셋-중분류	21	file_src_url	파일원천
22	ds_sclas	데이터셋-소분류	22	file_acces_author	파일접근권한
23	dc	설명	23	file_embargo_de	파일엠바고일자
24	dc_eng	영문설명	24	file_rights	파일저작권
25	ds_kwrd	데이터셋키워드	25	file_pblicte	파일발행처
26	ds_kwrd_eng	데이터셋영문키워드	26	file_colct_offic	파일수집처
27	charger_nm	담당자명	27	cmmnty_code	커뮤니티코드
28	charger_eng_nm	담당자명(영문)	28	preview	미리보기
29	charger_adres	담당자주소	29	file_viewr_at	파일뷰어여부
30	charger_tlphon	담당자전화	30	file_viewr_nm	파일뷰어명
31	charger_email	담당자이메일	31	file_path	파일경로
32	colct_offic	수집처	32	thembnail	썸네일정보
33	provd_offic	제공처	33	rgst_id	등록자ID
34	dataset_ty	데이터셋유형	34	rgsde	등록일자
35	dataset_fom	데이터셋형식	35	updusr_id	수정자ID
36	base_year	기준년도	36	updt_de	수정일자
37	confmer	승인자			
38	confm_rqester	승인요청자			
39	confm_de	승인일자			
40	etc_chartr_atrb	기타특성속성			
41	acces_author	접근권한			
42	embargo_de	엠바고일자			

데이터셋 메타데이터			파일 메타데이터		
번호	요소명	요소_한글명	번호	요소명	요소_한글명
43	doi	DOI			
44	sds_lang	데이터셋언어			
45	rights	저작권			
46	search_perm_at	검색허용여부			
47	data_civ_code	데이터구분코드			
48	creat_de	생성일			
49	regist_de	등록일			
50	ver	버전			
51	locplc	소재지			
52	src_url	원천URL			
53	cmmnty_code	커뮤니티코드			
54	doi_rhsde	DOI등록일			
55	rgst_id	등록자ID			
56	rgsde	등록일자			
57	updusr_id	수정자ID			
58	updt_de	수정일자			
59	Coverage/Temporal	데이터 보유 기간			
60	Coverage/Spatial	데이터 수집 지역			
61	Reference	관련 url			

[부록 3] '국가연구데이터플랫폼' 크로스워크 대상 메타데이터 요소의 정의

구분	범주화 결과(TTA 표준 매칭)			
	번호	요소명	요소_한글명	요소 정의
데이터셋	1	Identifier	식별자	각각의 연구데이터를 구분하기 위하여 정해진 규칙에 따라 부여한 이름
	2	Title	제목	데이터셋의 이름 또는 명칭
	3	Creator	생성자	데이터셋을 생성 또는 묶은 사람, 조직
	4	Publisher	출판사	데이터셋을 웹에 공개 또는 출판하는 주체
	5	PublicationYear	출판연도	데이터셋이 웹상에서 접근 또는 이용이 가능해진 연도
	6	Contributor	기여자	데이터셋의 수집, 생산, 예산지원 등에 관계된 사람이나 기관
	7	Subject	주제	데이터셋의 내용이 지닌 주제 정보
	8	Description	설명	데이터셋의 내용에 대한 설명
	9	Keyword	키워드	데이터셋의 내용이 지닌 주제에 대한 자연어
	10	Contact	연락처	데이터셋에 대한 문의사항에 대응할 수 있는 담당자의 연락처 정보
	11	ResourceType	데이터의 유형	데이터의 유형
	12	ExtraAtributes	기타특성정보	데이터셋의 기타 특성정보
	13	AccessType	접근유형	데이터셋에 접근하거나 이용할 수 있는 유형
	14	EmbargoDate	엠바고기한	데이터셋의 엠바고 일자
	15	Language	언어	데이터셋의 언어
	16	Rights	라이선스	데이터셋의 라이선스 정보
	17	CreateDate	생성일	데이터셋의 생성일
	18	Version	버전	데이터셋의 버전
	19	Location of Publisher	소재지	데이터셋의 소재 위치
	20	SourceIURL	원천정보 URL	데이터셋의 원천정보의 URL
	21	Coverage	수록범위	데이터셋의 공간적 시간적 수록범위
	22	Reference	관련정보 URL	데이터셋의 관련정보 URL
파일	1	File_Title	파일 제목	파일의 이름 또는 명칭
	2	File_Description	파일 설명	파일의 내용에 대한 설명
	3	File_Creator	파일 생성자	파일을 생성한 사람, 조직
	4	File_CreateDate	파일 생성일	파일의 생성일
	5	File_Size	파일 크기	파일의 크기
	6	File_format	파일 형식	파일의 형식
	7	File_Type	파일 유형	파일의 유형
	8	File_ExtraAttibutes	파일 기타특성정보	파일의 기타 특성정보
	9	File_SourceURL	파일 원천정보URL	파일의 원천정보 URL

[부록 4] 바이오연구데이터플랫폼 메타데이터 표준

[부록 4-1] 전분야 공통 – BioProject 정보 / 대상 21개

파트	요소명	요소_한글명	필수(M)/ 선택(O)
Part1. Submitter information (등록자 인적사항)	Name of submitter	등록자의 영문 이름	M
	Name of submitter in Korean	등록자의 국문 이름	M
	Primary e-mail of submitter	등록자의 주 이메일 주소	M
	Secondary e-mail of submitter	등록자의 부 이메일 주소	O
	Name of submitter's organization	등록자 소속 기관의 영문명	M
	Name of submitter's organization in Korean	등록자 소속 기관의 국문명	M
	Department of submitter	등록자 소속 학과/부서의 영문명	M
	Department of submitter in Korean	등록자 소속 학과/부서의 국문명	M
	Phone number of submitter	등록자의 전화번호	O
	Address of submitter	등록자의 영문 주소	M
	Address of submitter in Korean	등록자의 국문 주소	M
	Country of submitter	등록자의 국가	M
	Researcher ID of submitter	등록자의 연구자 고유번호	O
Part2. Date (날짜)	Submission date	제출 날짜	M
	Release date	공개 날짜	M
Part3. Title and description of the project (프로젝트의 제목 및 설명)	Project title	프로젝트의 영문 제목	M
	Project title in Korean	프로젝트의 국문 제목	M
	Project description	프로젝트의 영문 설명	M
	Project description in Korean	프로젝트의 국문 설명	M
Part4. Grant information (연구 과제 정보)	Grant ID	연구과제 번호	M
	Grant NTIS ID	연구과제의 NTIS 번호	M
	Grant title	연구과제의 영문 제목	M
	Grant title in Korean	연구과제의 국문 제목	M
	Funding agency	연구비 지원 기관	M
Part5. Publication and patent information (논문 및 특허 성과 정보)	PubMed ID of publication	논문의 PubMed ID	O
	DOI of publication	논문의 DOI	O
	Patent information	특허정보	O

[부록 4-2] 다수 분야 공통 – BioSample 정보 / 대상 16개

<table>
<tr><th>파트</th><th colspan="2">요소명</th><th>요소_한글명</th></tr>
<tr><td>Part1. Sample type
(샘플 종류)</td><td colspan="2">sample type (1)</td><td>샘플 종류 (1)</td></tr>
<tr><td></td><td colspan="3"><table>
<tr><th>No</th><th>가능한 값</th><th>설명</th></tr>
<tr><td>1</td><td>Human</td><td>인간</td></tr>
<tr><td>2</td><td>Model organism or animal</td><td>모델생물 및 동물</td></tr>
<tr><td>3</td><td>Invertebrate</td><td>무척추동물</td></tr>
<tr><td>4</td><td>Plant</td><td>식물</td></tr>
<tr><td>5</td><td>Microbe</td><td>미생물</td></tr>
<tr><td>6</td><td>Virus</td><td>바이러스</td></tr>
<tr><td>7</td><td>Clinical of host-associated pathogen</td><td>임상병원체</td></tr>
<tr><td>8</td><td>Environmental/food/other pathogen</td><td>환경, 식품 및 기타 병원체</td></tr>
<tr><td>9</td><td>Metagenome or environmental</td><td>메타게놈 및 환경샘플</td></tr>
</table></td></tr>
<tr><td rowspan="16">Part2. Sample attribute
(샘플 특성)

※ 전체 79개 항목이며, 대상으로 한 값(1회 이상 필수, 15개 항목)만 표기</td><td>1</td><td>age</td><td>나이</td></tr>
<tr><td>3</td><td>biomaterial provider</td><td>생물질 제공자</td></tr>
<tr><td>13</td><td>collected by</td><td>수집자</td></tr>
<tr><td>14</td><td>collection date</td><td>수집 일자</td></tr>
<tr><td>26</td><td>geographic location</td><td>지리적 장소</td></tr>
<tr><td>30</td><td>host</td><td>숙주</td></tr>
<tr><td>33</td><td>host disease</td><td>숙주 질병</td></tr>
<tr><td>41</td><td>isolate</td><td>분리</td></tr>
<tr><td>42</td><td>isolation source</td><td>분리 소스</td></tr>
<tr><td>45</td><td>latitude and longitude</td><td>위도 및 경도</td></tr>
<tr><td>47</td><td>organism</td><td>생명체 명</td></tr>
<tr><td>57</td><td>sample name</td><td>샘플명</td></tr>
<tr><td>60</td><td>sample type (2)</td><td>샘플 타입 (2)</td></tr>
<tr><td>63</td><td>sex</td><td>성별</td></tr>
<tr><td>72</td><td>tissue</td><td>조직</td></tr>
</table>

[부록 4-3] 다수분야 공통 – Omics 데이터

Omics 데이터는 크게 5개로 나누어져 있는데, 이 중 일부 중복되는 요소명에는 언급되는 순서에 따라 (1) (2)로 표시하여 구분하였다. 전체 170개 요소 중 중복되는 요소명은 모두 6개 이며, 해당 요소는 Title, Platform, Growth protocol, Treatment protocol, Keywords, Extraction Protocol이다. 이 중 본 연구에서는 필수 요소만 크로스워크 대상으로 선정하였기 때문에, 크로스워크 대상인 요소는 Title, Platform, Keywords, Extraction Protocol의 4개 요소이다.

1. NGS 데이터 / 대상 12개

파트	요소명	요소_한글명	필수(M)/ 선택(O)
Part1. NGS metadata (차세대 시퀀싱 속성정보)	Library ID	라이브러리 ID	M
	Title (1)	제목 (1)	M
	Library strategy	시퀀싱 기법	M
	Library source	라이브러리 출처	M
	Library selection	라이브러리 선택 방법	M
	Library layout	라이브러리 레이아웃	M
	Platform (1)	플랫폼 (1)	M
	Instrument model	기기모델	M
	Design description	디자인 설명	M
Part2. Files	Filetype	파일 타입	M
	Filename	파일명	M
	Reference	참조 서열	M

2. 마이크로어레이 데이터 / 대상 17개

파트	요소명	요소_한글명	필수(M)/ 선택(O)
Part1. Series (시리즈)	Title (2)	제목 (2)	M
	Summary	요약	M
	Overall design	전반적 디자인	M
Part2. Sample detail (샘플 상세정보)	detailed sample title	상세 샘플 제목	M
	source name	소스 명	M
	characteristics: tag	특성 태그	M
	molecule	분자	M
	label	라벨	M
	description	설명	M
	platform (2)	플랫폼 (2)	M
Part3. Protocol (실험 프로토콜)	growth protocol (1)	성장 프로토콜 (1)	O
	treatment protocol (1)	처리 프로토콜 (1)	O
	extract protocol (1)	추출 프로토콜 (1)	M
	label protocol	라벨링 프로토콜	M
	hybridization protocol	혼성화 프로토콜	M
	scan protocol	스캔 프로토콜	M
	data processing	데이터 프로세싱	M
Part4. Result files (결과 파일)	raw data file	원시 데이터 파일	M
	processed data file	가공 데이터 파일	M

3. 염기서열 데이터 / 대상 26개

파트	요소명	요소_한글명	필수(M)/ 선택(O)
Part1. Sequencing technology 정보	Assembly date	어셈블리 일자	O
	Assembly method	어셈블리 방식	M
	Program version	프로그램 버전	M
	Assembly name	어셈블리명	O
	Genome coverage	게놈 커버리지	M
	Total raw read length	전체 원시데이터 길이	M
	Read throughput	총 염기서열 read 수	M
	The number of contigs	전체 contig 수	M
	Contig length	전체 contig 길이	M
	N50	-	M
	Sequencing technology	시퀀싱 기술	M
	Confirm full genome	전체 게놈 포함 여부 확인	M
	Description for subset of the genome	게놈 영역 설명	M
	Confirm final version	최종버전 여부	M
	Confirm de novo assembly	드 노보 어셈블리 여부	M
	Reference assembly name or accession	참조 어셈블리 이름	M
	Confirm update of existing submission	업데이트 여부	M
	Existing genome accession	기존 게놈 등록번호	M
Part2. Sequences/Nucleotide 정보	Molecule Type	분자 유형	M
	Topology	위상	M
	Source Organelle/Location information	원천 소기관/위치 정보	M
	Chimera check	키메라 서열 확인	M
	Chimera check program name	키메라 확인 프로그램명	M
	Chimera check program version	키메라 확인 프로그램 버전	O
	Cultured or Uncultured	배양 여부	M
	Primer Type	프라이머 유형	M
	Forward primer name	포워드 프라이머명	O
	Forward primer sequence	포워드 프라이머 시퀀스	O
	Reverse primer name	리버스 프라이머명	O
	Reverse primer sequence	리버스 프라이머 시퀀스	O
Part3. Set/Batch 정보	Submission Set/Batch	제출 세트/뱃치	M
Part4. Feature 정보	Add features	특성 추가	M
Part5. File 정보	Assembly state	어셈블리 상태	O
	Confirm AGP file for unplaced scaffolds	비배치 스캐폴드 AGP 파일 유무	O
	Confirm AGP file for unlocalized scaffolds	비위치화 스캐폴드 AGP 파일 유무	O
	How to assemble using AGP	AGP 이용 조립방법	O
	Confirm annotation in AGP	AGP 주석 유무	O
	AGP file	AGP 파일	O
	FASTA file	FASTA 파일	M

4. 대사체 데이터 / 대상 19개

파트	요소명	요소_한글명	필수(M)/ 선택(O)
Part1. Descriptors (설명자)	Keywords (1)	키워드 (1)	M
	Factors	요인	M
Part2. Sample details (샘플 상세정보)	Variant	변이	O
	Organism part	부위	O
	Additional characteristic	추가 특성	M
Part3. Sample collection	Sample collection protocol	샘플 수집 프로토콜	M
Part4. Extraction	Extraction protocol (2)	추출 프로토콜 (2)	M
	Derivatization	유도체화	O
	Post extraction	후 추출	O
	Internal standard	내부 표준	O
Part5. Platform	Technique type	기술 유형	M
	Assay type	어세이 유형	M
	Assay definition	어세이 정의	M
Part6. Chromatography	Chromatography protocol	크로마토그래피 프로토콜	M
	Chromatography instrument model	크로마토그래피 장비 모델	O
	Autosampler model	오토샘플러 모델	O
	Column model	컬럼 모델	O
	Column type	컬럼 유형	O
	Guard column	가드 컬럼	O
	Column temperature	컬럼 온도	O
	Column temperature unit	컬럼 온도 단위	O
	Flow rate	유량	O
	Solvent for chromatography	크로마토그래피 용매	O
	Gradient	경사	O
	Injection	주입	O
Part7. Mass spectrometry	Mass spectrometry protocol	질량 분석 프로토콜	M
	Scan polarity	스캔 극성	O
	Scan M/Z range	스캔 M/Z 범위	O
	Mass spectrometry instrument model	질량 분석 장비 모델	O
	Ion Source	이온원	O
	Mass analyzer	질량 분석기	O
	Capillary voltage	모세관 전압	O
	Capillary voltage unit	모세관 전압 단위	O
	Capillary source temperature	모세관 소스 온도	O
	Capillary source temperature unit	모세관 소스 온도 단위	O
	MS/MS data acquisition	MS/MS 데이터 확보 방법	O
Part8. NMR sample	NMR sample protocol	NMR 샘플 프로토콜	M
	NMR tube type	NMR 튜브 유형	O
	Solvent for NMR	NMR 용매	O
	Sample pH	샘플 산성도	O
	Temperature	온도	O
	Temperature unit	온도 단위	O

파트	요소명	요소_한글명	필수(M)/ 선택(O)
Part9. NMR spectroscopy	NMR spectroscopy protocol	NMR 분광 프로토콜	M
	NMR instrument model	NMR 장비 모델	O
	NMR probe	NMR 프로브	O
	Number of transients	Transient의 개수	O
	Pulse sequence name	펄스 시퀀스명	O
	Magnetic field strength	자기장 강도	O
	Number of data points	데이터 포인트 수	O
	Spectral width	스펙트럼 폭	O
	Relaxation delay	이완지연	O
	Acquisition time	획득 시간	O
Part10. NMR assay	NMR assay protocol	NMR 어세이 프로토콜	M
	Data processing and software	데이터 처리 및 소프트웨어	O
	Peak alignment	피크 정렬	O
	Peak deconvolution	피크 deconvolution 방법	O
Part11. Data transformation	Data transformation protocol	데이터 변환 프로토콜	M
	Normalization	정규화	O
Part12. Metabolite identification	Metabolite identification protocol	대사물질 식별 프로토콜	M
	Metabolite quantification	대사물질 정량화	O
	Calibration standard	보정 표준	M
	Calibration	보정 방법	O
Part13. Files (결과 파일)	Acquired raw data files	원시 데이터 파일	M
	Data processed peak table	각 peak에 대한 intensity data	M
	Concentration data	대사체 ID 및 대사체 농도에 대한 정량 값	M

5. 단백체 데이터 / 대상 20개

파트	요소명	요소_한글명	필수(M)/ 선택(O)
Part1. Descriptors (설명자)	Keywords (2)	키워드 (2)	M
	Period of Creation	데이터 생산 기간	M
	Submission Type	데이터 완성도	M
Part2. Sample details (샘플 상세정보)	Subcellular	세포소기관	M
Part3. Experiment and dataset details (실험 및 데이터셋 상세정보)	Experiment type	실험 유형	M
	Sample processing protocol	샘플 처리 프로토콜	M
	Data processing protocol	데이터 처리 프로토콜	M
	Acquisition Protocol	획득 프로토콜	O
	Digestion Protocol	소화 프로토콜	O
	Extraction Protocol (3)	추출 프로토콜 (3)	O
	Growth Protocol (2)	성장 프로토콜 (2)	O
	Separation Protocol	분리 프로토콜	O
	Treatment Protocol (2)	처리 프로토콜 (2)	O

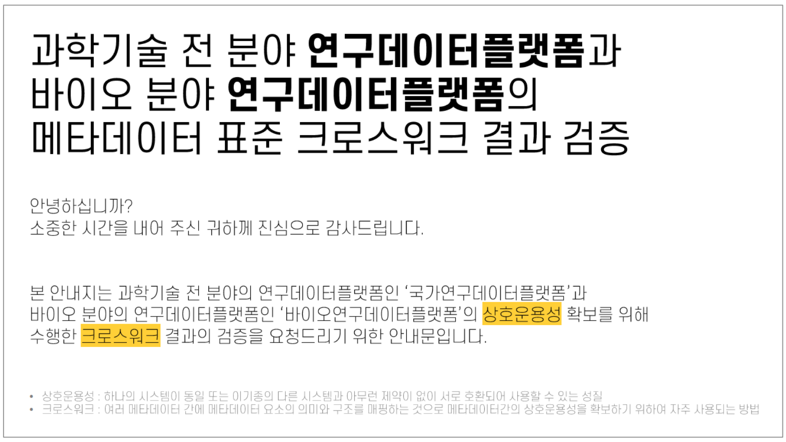
파트	요소명	요소_한글명	필수(M)/ 선택(O)
	Search parameters	검색 설정값	M
	LC operation method	LC 운용 방법	O
	MS operation method	MS 운용 방법	O
	UniProt DB	UniProt 데이터베이스	O
	Experimental factor	실험 요인	O
	Instrument	장비	M
	Quantification method	정량화 방법	O
	Modification	변형	M
	Enzyme	효소	M
	LC system	LC 시스템	M
	PTM	단백질 번역 후 변형	M
	Sample fraction	샘플 분획	M
	Fractionation	분획	M
	plex	플렉스	M
Part4. File (결과 파일)	Raw	원시데이터	M
	Peak list	피크 목록	M
	Search list	검색 목록	M
	Quantification	정량화	O
	FASTA	-	M
	Spectrum library	스펙트럼 라이브러리	O
	Gel image	겔 이미지	O
	Other	이 외의 것	O

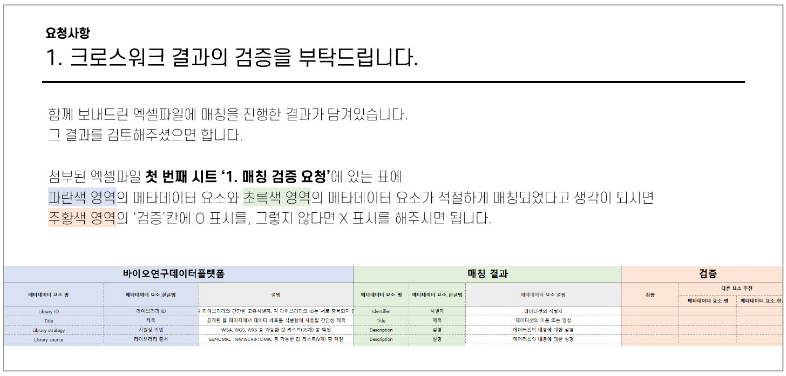
[부록 5] 크로스워크 결과 검증자 목록 및 질문지

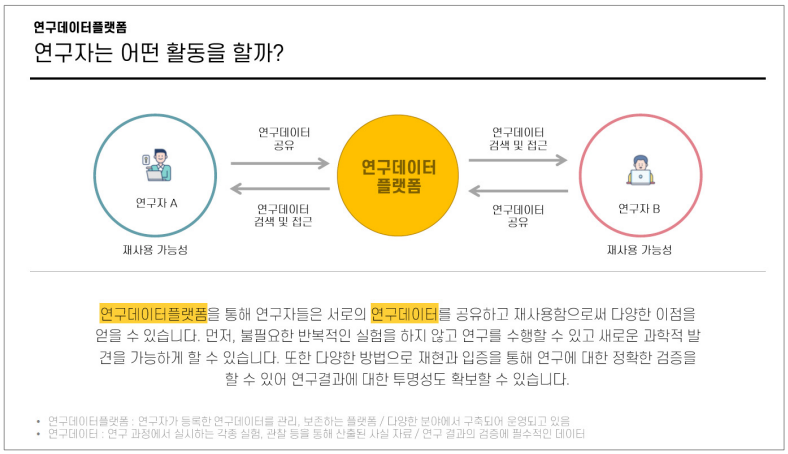
[부록 5-1] 검증자 목록

번호	성명	소속기관
1	구**	충남대학교 약학대학 대학원
2	김**	충남대학교 약학대학 대학원
3	김**	충남대학교 약학대학 대학원
4	백**	서울대학교 약학대학 대학원
5	이**	서울대학교 약학대학 대학원
6	정**	충남대학교 약학대학 대학원
7	정**	서울대학교 약학대학 대학원

[부록 5-2] 질문지







※ 이하는 연구데이터플랫폼과 메타데이터, 상호운용성에 대한 설명이기 때문에 생략함

[부록 6] 최종 매칭 결과: 범주화 요소를 기준으로

1. 데이터셋

기준 요소				매칭 요소					
국가연구데이터플랫폼		범주화 결과(TTA 표준 매칭)		바이오연구데이터플랫폼					
요소명	요소_한글명	요소명	요소_한글명	Level 1	Level 2	Level 3	Part	요소명	요소_한글명
dataset_id	데이터셋 ID	Identifier	식별자	Part Ⅱ. 전 분야 공통	2. BioProject 정보	-	Part4. Grant information (연구과제정보)	Grant ID	연구과제 번호
								Grant NTIS ID	연구과제의 NTIS 번호
dataset_id_type	데이터셋 ID유형			Part Ⅲ. 다수 분야 공통	4. Omics(오믹스) 데이터	4.1 NGS (차세대 시퀀싱) 데이터	Part1. NGS metadata (차세대 시퀀싱 속성정보)	Library ID	라이브러리 ID
						4.3 Nucleotide sequence (염기서열) 데이터	Part1. Sequencing technology 정보	Existing genome accession	기존 게놈 등록번호
title	제목	title	제목	Part Ⅱ. 전 분야 공통	2. BioProject 정보	-	Part3. Title and description of the project (프로젝트의 제목 및 설명)	Project title	프로젝트의 영문 제목
								Project title in Korean	프로젝트의 국문 제목
						-	Part4. Grant information (연구과제정보)	Grant title	연구과제의 영문 제목
								Grant title in Korean	연구과제의 국문 제목
title_eng	영문제목			Part Ⅲ. 다수 분야 공통	4. Omics(오믹스) 데이터	4.1 NGS (차세대 시퀀싱) 데이터	Part1. NGS metadata (차세대 시퀀싱 속성정보)	Title (1)	제목 (1)
						4.2 Microarray (마이크로어레이) 데이터	Part1. Series (시리즈)	Title (2)	제목 (2)
crtr	생성자	Creator	생성자	Part Ⅲ. 다수 분야 공통	3. BioSample 정보	-	Part2. Sample attribute (샘플 특성)	collected by	수집자
crtr_eng	생성자 영문명								
crtr_pstinst	생성자 소속기관								
pblicte	발행처	Publisher	출판사	Part Ⅱ. 전 분야 공통	2. BioProject 정보	-	Part1. Submitter information (등록자인적사항)	Name of submitter	등록자의 영문 이름
pblicte_eng	발행처 영문명							Name of submitter in Korean	등록자의 국문 이름

기준 요소				매칭 요소					
국가연구데이터플랫폼		범주화 결과(TTA 표준 매칭)		바이오연구데이터플랫폼					
요소명	요소_한글명	요소명	요소_한글명	Level 1	Level 2	Level 3	Part	요소명	요소_한글명
pblicte_year	발행연도	Publication Year	출판연도	Part Ⅱ. 전 분야 공통	2. BioProject 정보	-	Part2. Date (날짜)	Submission date	제출 날짜
cntrbtor	기여자	Contributor	기여자	Part Ⅱ. 전 분야 공통	2. BioProject 정보	-	Part4. Grant information (연구 과제 정보)	Funding agency	연구비 지원 기관
cntrbtor_eng_nm	기여자 영문명								
cntrbtor_ty	기여자 유형			Part Ⅲ. 다수 분야 공통	3. BioSample 정보	-	Part2. Sample attribute (샘플 특성)	biomaterial provider	생물질 제공자
cntrbtor_org_nm	기여자 기관명								
cntrbtor_org_code	기여자 기관코드								
sj_cl_nm	주제 분류명	Subject	주제	Part Ⅲ. 다수 분야 공통	4. Omics(오믹스) 데이터	4.2 Microarray (마이크로어레이) 데이터	Part2. Sample detail (샘플 상세정보)	detailed sample title	상세 샘플 제목
ds_lclas	데이터셋 -대분류								
ds_mlsfc	데이터셋 -중분류								
ds_sclas	데이터셋 -소분류								
dc	설명	Description	설명	Part Ⅱ. 전 분야 공통	2. BioProject 정보	-	Part3. Title and description of the project (프로젝트의 제목 및 설명)	Project description	프로젝트의 영문 설명
								Project description in Korean	프로젝트의 국문 설명
				Part Ⅲ. 다수 분야 공통	3. BioSample 정보	-	Part2. Sample attribute (샘플 특성)	geographic location	지리적 장소
								host	숙주
								host disease	숙주 질병
								isolate	분리
								isolation source	분리 소스
								latitude and longitude	위도 및 경도
								sex	성별
								tissue	조직
					4. Omics(오믹스) 데이터	4.1 NGS (차세대 시퀀싱) 데이터	Part1. NGS metadata (차세대 시퀀싱 속성정보)	Library strategy	시퀀싱 기법
								Library source	라이브러리 출처
								Library selection	라이브러리 선택 방법

기준 요소				매칭 요소					
국가연구데이터플랫폼		범주화 결과(TTA 표준 매칭)		바이오연구데이터플랫폼					
요소명	요소_한글명	요소명	요소_한글명	Level 1	Level 2	Level 3	Part	요소명	요소_한글명
								Library layout	라이브러리 레이아웃
								Platform (1)	플랫폼 (1)
								Instrument model	기기모델
								Design description	디자인 설명
						4.2 Microarray (마이크로어레이) 데이터	Part1. Series (시리즈)	Summary	요약
								Overall design	전반적 디자인
							Part2. Sample detail (샘플 상세정보)	source name	소스 명
								molecule	분자
								description	설명
								platform (2)	플랫폼 (2)
							Part3. Protocol (실험 프로토콜)	extract protocol (1)	추출 프로토콜 (1)
								label protocol	라벨링 프로토콜
								hybridization protocol	혼성화 프로토콜
								scan protocol	스캔 프로토콜
								data processing	데이터 프로세싱
						4.3 Nucleotide sequence (염기서열) 데이터	Part1. Sequencing technology 정보	Assembly method	어셈블리 방식
								Genome coverage	게놈 커버리지
								Total raw read length	전체 원시데이터 길이
								Read throughput	총 염기서열 read 수
								The number of contigs	전체 contig 수
								Contig length	전체 contig 길이
								N50	N50
								Sequencing technology	시퀀싱 기술
								Confirm full genome	전체 게놈 포함 여부 확인
								Description for subset of the genome	게놈 영역 설명

기준 요소				매칭 요소					
국가연구데이터플랫폼		범주화 결과(TTA 표준 매칭)		바이오연구데이터플랫폼					
요소명	요소_한글명	요소명	요소_한글명	Level 1	Level 2	Level 3	Part	요소명	요소_한글명
								Confirm de novo assembly	드 노보 어셈블리 여부
								Reference assembly name or accession	참조 어셈블리 이름
dc_eng	영문설명						Part2. Sequences/ Nucleotide 정보	Molecule Type	분자 유형
								Topology	위상
								Source Organelle/Location information	원천 소기관/위치 정보
								Chimera check	키메라 서열 확인
								Chimera check program name	키메라 확인 프로그램명
								Cultured or Uncultured	배양 여부
								Primer Type	프라이머 유형
							Part3. Set/Batch 정보	Submission Set/Batch	제출 세트/뱃치
						4.4 Metabolomics (대사체) 데이터	Part1. Descriptors (설명자)	Factors	요인
							Part3. Sample collection	Sample collection protocol	샘플 수집 프로토콜
							Part4. Extraction	Extraction protocol (2)	추출 프로토콜 (2)
							Part5. Platform	Technique type	기술 유형
								Assay type	어세이 유형
								Assay definition	어세이 정의
							Part6. Chromatography	Chromatography protocol	크로마토그래피 프로토콜
							Part7. Mass spectrometry	Mass spectrometry protocol	질량 분석 프로토콜
							Part8. NMR sample	NMR sample protocol	NMR 샘플 프로토콜
							Part9. NMR spectroscopy	NMR spectroscopy protocol	NMR 분광 프로토콜

기준 요소				매칭 요소					
국가연구데이터플랫폼		범주화 결과(TTA 표준 매칭)		바이오연구데이터플랫폼					
요소명	요소_한글명	요소명	요소_한글명	Level 1	Level 2	Level 3	Part	요소명	요소_한글명
							Part10. NMR assay	NMR assay protocol	NMR 어세이 프로토콜
							Part11. Data transformation	Data transformation protocol	데이터 변환 프로토콜
							Part12. Metabolite identification	Metabolite identification protocol	대사물질 식별 프로토콜
								Calibration standard	보정 표준
						4.5 Proteomics (단백체) 데이터	Part2. Sample details (샘플 상세정보)	Subcellular	세포소기관
							Part3. Experiment and dataset details (실험 및 데이터셋 상세정보)	Sample processing protocol	샘플 처리 프로토콜
								Data processing protocol	데이터 처리 프로토콜
								Search parameters	검색 설정값
								Modification	변형
								Enzyme	효소
								LC system	LC 시스템
								PTM	단백질 번역 후 변형
								Sample fraction	샘플 분획
								Fractionation	분획
								plex	플렉스
ds_kwrd	데이터셋 키워드	Keyword	키워드	Part Ⅲ. 다수 분야 공통	3. BioSample 정보	-	Part1. Sample type (샘플 종류)	Sample type (1)	샘플 종류 (1)
							Part2. Sample attribute (샘플 특성)	organism	생명체 명
								sample name	샘플명
					4. Omics(오믹스) 데이터	4.2 Microarray (마이크로어레이) 데이터	Part2. Sample detail (샘플 상세정보)	characteristics: tag	특성 태그
								label	라벨
						4.4 Metabolomics (대사체) 데이터	Part1. Descriptors (설명자)	Keywords (1)	키워드 (1)
ds_kwrd_eng	데이터셋 영문키워드					4.5 Proteomics (단백체) 데이터	Part1. Descriptors (설명자)	Keywords (2)	키워드 (2)

기준 요소				매칭 요소					
국가연구데이터플랫폼		범주화 결과(TTA 표준 매칭)		바이오연구데이터플랫폼					
요소명	요소_한글명	요소명	요소_한글명	Level 1	Level 2	Level 3	Part	요소명	요소_한글명
charger_nm	담당자명	Contact	연락처	Part Ⅱ. 전 분야 공통	2. BioProject 정보	-	Part1. Submitter information (등록자인적사항)	Primary e-mail of submitter	등록자의 주 이메일 주소
charger_eng_nm	담당자명(영문)								
charger_adres	담당자주소								
charger_tlphon	담당자전화								
charger_email	담당자이메일								
dataset_fom	데이터셋형식	ResourceType	데이터의 유형	Part Ⅲ. 다수 분야 공통	3. BioSample 정보	-	Part2. Sample attribute (샘플 특성)	Sample type (2)	샘플 종류 (2)
					4. Omics(오믹스) 데이터	4.5 Proteomics (단백체) 데이터	Part3. Experiment and dataset details (실험 및 데이터셋 상세정보)	Experiment type	실험 유형
etc_chartr_atrb	기타특성속성	ExtraAttribute	기타특성정보	Part Ⅱ. 전 분야 공통	2. BioProject 정보	-	Part1. Submitter information (등록자인적사항)	Address of submitter	등록자의 영문 주소
								Address of submitter in Korean	등록자의 국문 주소
								Country of submitter	등록자의 국가
				Part Ⅲ. 다수 분야 공통	3. BioSample 정보	-	Part2. Sample attribute (샘플 특성)	age	나이
					4. Omics(오믹스) 데이터	4.3 Nucleotide sequence (염기서열) 데이터	Part4. Feature 정보	Add features	특성 추가
						4.4 Metabolomics (대사체) 데이터	Part2. Sample details (샘플 상세정보)	Additional characteristic	추가 특성
						4.5 Proteomics (단백체) 데이터	Part1. Descriptors (설명자)	Submission Type	데이터 완성도
							Part2. Sample details (샘플 상세정보)	Instrument	장비
acces_author	접근권한	AccessType	접근유형						

기준 요소				매칭 요소					
국가연구데이터플랫폼		범주화 결과(TTA 표준 매칭)		바이오연구데이터플랫폼					
요소명	요소_한글명	요소명	요소_한글명	Level 1	Level 2	Level 3	Part	요소명	요소_한글명
embargo_de	엠바고일자	EmbargoDate	엠바고 기한	Part Ⅱ. 전 분야 공통	2. BioProject 정보	-	Part2. Date (날짜)	Release date	공개 날짜
sds_lang	데이터셋 언어	Language	언어						
rights	저작권	Rights	라이선스						
creat_de	생성일	CreateDate	생성일	Part Ⅲ. 다수 분야 공통	3. BioSample 정보	-	Part2. Sample attribute (샘플 특성)	collection date	수집 일자
					4. Omics(오믹스) 데이터	4.5 Proteomics (단백체) 데이터	Part1. Descriptors (설명자)	Period of Creation	데이터 생산 기간
ver	버전	Version	버전	Part Ⅲ. 다수 분야 공통	4. Omics(오믹스) 데이터	4.3 Nucleotide sequence (염기서열) 데이터	Part1. Sequencing technology 정보	Program version	프로그램 버전
								Confirm final version	최종버전 여부
								Confirm update of existing submission	업데이트 여부
locplc	소재지	Location of Publisher	소재지	Part Ⅱ. 전 분야 공통	2. BioProject 정보	-	Part1. Submitter information (등록자인적사항)	Name of submitter's organization	등록자 소속 기관의 영문명
								Name of submitter's organization in Korean	등록자 소속 기관의 국문명
								Department of submitter	등록자 소속 학과/부서의 영문명
								Department of submitter in Korean	등록자 소속 학과/부서의 국문명
src_url	원천URL	SourceIURL	원천정보 URL						
Coverage/ Temporal	데이터 보유 기간	Coverage	수록범위						
Coverage/Spatial	데이터 수집 지역								
Reference	관련 url	Reference	관련정보 URL						

2. 파일

기준 요소				매칭 요소					
국가연구데이터플랫폼		범주화 결과(TTA 표준 매칭)		바이오연구데이터플랫폼					
요소명	요소_한글명	요소명	요소_한글명	Level 1	Level 2	Level 3	Part	요소명	요소_한글명
file_title	파일 제목	File_Title	파일 제목	Part Ⅲ. 다수 분야 공통	4. Omics(오믹스) 데이터	4.1 NGS (차세대 시퀀싱) 데이터	Part2. Files	Filename	파일명
file_dc	파일설명	File_ Description	파일 설명	Part Ⅲ. 다수 분야 공통	4. Omics(오믹스) 데이터	4.1 NGS (차세대 시퀀싱) 데이터	Part2. Files	Reference	참조 서열
file_crtr	파일생성자	File_ Creator	파일 생성자						
file_crtr_eng	파일생성자 영문명								
file_crtr_pstinst	파일생성자 소속기관								
file_creat_de	파일생성일	File_Create Date	파일 생성일						
file_size	파일크기	File_Size	파일 크기						
file_fom	파일형식	File_format	파일 형식	Part Ⅲ. 다수 분야 공통	4. Omics(오믹스) 데이터	4.3 Nucleotide sequence (염기서열) 데이터	Part5. File 정보	FASTA file	FASTA 파일
						4.5 Proteomics (단백체) 데이터	Part4. File (결과 파일)	Raw	원시데이터
								Search list	검색목록
								FASTA	FASTA
file_ty	파일유형	File_Type	파일 유형	Part Ⅲ. 다수 분야 공통	4. Omics(오믹스) 데이터	4.1 NGS (차세대 시퀀싱) 데이터	Part2. Files	Filetype	파일 타입
etc_chartr_file_atrb	기타특성파일 메타	File_ExtraAttributes	파일 기타특성 정보	Part Ⅲ. 다수 분야 공통	4. Omics(오믹스) 데이터	4.2 Microarray (마이크로어레이) 데이터	Part4. Result files (결과 파일)	raw data file	원시 데이터 파일
								processed data file	가공 데이터 파일

기준 요소				매칭 요소					
국가연구데이터플랫폼		범주화 결과(TTA 표준 매칭)		바이오연구데이터플랫폼					
요소명	요소_한글명	요소명	요소_한글명	Level 1	Level 2	Level 3	Part	요소명	요소_한글명
						4.4 Metabolomics (대사체) 데이터	Part13. Files (결과 파일)	Acquired raw data files	원시 데이터 파일
								Data processed peak table	각 peak에 대한 intensity data
								Concentration data	대사체 ID 및 대사체 농도에 대한 정량 값
						4.5 Proteomics (단백체) 데이터	Part4. File (결과 파일)	Peak list	피크 목록
file_src_url	파일원천	File_Sourece URL	파일 원천정보 URL						