https://doi.org/10.5392/JKCA.2017.17.09.001

커뮤니티 주도적 과학 데이터 큐레이션 협업 환경의 개발

Development of Collaborative Environment for Community-driven Scientific Data Curation

최동훈, 박재원, 김병규, 신진섭
한국과학기술정보연구원

Dong-Hoon Choi(choid@kisti.re.kr), Jae-Won Park(ilonetos@kisti.re.kr),
ByungKyu Kim(bk.kim@kisti.re.kr), Jin-Sup Shin(js.shin@kisti.re.kr)

요약

데이터 재사용 수요가 증가할수록 데이터 큐레이션의 중요성에 대한 인식은 점차 증가하고 있다. 데이터의 폭증으로 인해, 과학자들은 전체 노력의 90%를 자신의 연구에 필요한 데이터의 검색 및 수집에 들이고 있다. 이러한 노력을 절감시키기 위해, 본 논문에서는 과학 데이터의 재사용성을 높이는 데 필수적인 커뮤니티 주도적 데이터 큐레이션 협업 환경의 개발 및 적용에 대해 다룬다. 본 과학 큐레이션 협업 환경은 특정 연구 분야의 연구 결과 간에 상호 연관성을 포획하고 재구성하기 위해 데이터 (또는 데이터 컬렉션) 및 관련 문헌 간의 상호 연결에 초점을 맞추고 있다. 또한 풍부한 문맥 정보를 메타데이터로 제공하여 사용자의 데이터 이해를 돕는다. 데이터 및 데이터-문헌 간의 상호 연결을 영구적으로 접근할 수 있도록 보장하기 위해, DOI 시스템을 이용하여 실현하였다. 이 큐레이션 협업 환경은 국내외 연구자들로 구성된 무정형 단백질 연구 그룹에 의해 커뮤니티 주도적인 큐레이션 데이터베이스 구축에 적용되었다. 이렇게 구축된 데이터베이스는 무정형 단백질 연구자의 과학적 발견을 위한 데이터 검색 및 수집 노력을 절감해 줄 것이다.

■ 중심어 : | 데이터 재사용성 | 커뮤니티 주도적 데이터 큐레이션 | 큐레이션 협업 환경 | 데이터-문헌 상호 연결 | 문맥 정보 |

Abstract

The importance of data curation is increasingly recognized as the need of data reuse drastically grows. Due to recent data explosion, scientists invest almost 90% of their efforts in the retrieval and collection of data needed to their study. In this paper, we deal with the development and application of a collaborative environment for community-driven data curation which is essential to enhance scientific data reusability and citability. The collaborative scientific data curation environment focuses on the cross-linking between data (or data collections) and their associated literatures to capture and organize inter-relations among research results in a specific domain. Also, plenty of contextual information is provided as metadata in order to support users in understanding data. The cross-linking has been realized by using DOI system to guarantee global accessibility to data and their relationships to literatures. The curation environment has been adopted to build a community-driven curated DB by a globally well-known intrinsically-disorderd protein research group. The curated DB will drastically reduce researchers' efforts to retrieve and collect the data required for scientific discovery.

■ keyword : | Data Reusability | Community-Driven Data Curation | Collaborative Curation Environment | Data-Literature Interlinking | Context Information |

* 본 연구는 (2017년도) 한국과학기술정보연구원(KISTI) 주요사업 과제로 수행한 것입니다.
접수일자 : 2017년 06월 21일
수정일자 : 2017년 07월 12일
심사완료일 : 2017년 07월 26일
교신저자 : 신진섭, e-mail : js.shin@kisti.re.kr

Ⅰ. 서론

데이터 공개가 증가하면서 과학자는 데이터 바다에서 자신의 연구에 필요한 데이터를 찾아내서, 다양한 데이터 리파지토리로부터 내려 받은 데이터를 이해하고 서로 다른 포맷의 데이터를 통합하여 자신의 데이터 컬렉션을 생성하기까지 많은 노력을 들이고 있다. 이와 같이 데이터 수집을 포함한 전처리 과정에 드는 연구자의 노력은 전체의 90%(최빈값)에 이른다[1]. 이러한 고통을 경감시키려면, 연구자가 데이터를 손쉽게 찾아서 접근할 수 있고 쉽게 이해할 수 있을 뿐만 아니라 다른 리파지토리의 데이터와 쉽게 통합하여 재사용할 수 있도록, 충분한 메타데이터를 제공하는 것이 중요하다. 데이터의 발견, 접근, 이해, 통합, 재사용을 위한 메타데이터를 정의하고 생성하여 데이터의 부가가치를 제고하려는 노력을 데이터 큐레이션(curation)이라고 한다[2]. 연구 데이터의 큐레이션은 분야 전문가의 지식과 경험을 요구하기 때문에, 분야 전문가, 데이터 사서(data librarian), 데이터 재사용자 등으로 구성된 커뮤니티가 데이터 수명주기에 걸쳐 데이터 이해 및 재사용에 필요한 메타데이터를 정의하고 생성한다. 이렇게 생성된 메타데이터는 연구 데이터의 재사용을 촉진한다. 분야 간의 융합이 대세를 이루고 있는 요즘의 연구 환경에서는 분야 연구자 간의 데이터 재사용 수요가 늘어나면서 이에 따른 데이터 큐레이션 요구가 날로 증가하고 있다.

데이터는 관련 조건이나 여건 등에 따라 명확하고 적절한 의미를 갖는다. 이와 같이 어떤 데이터의 의미를 결정하는 관련 조건이나 여건을 문맥이라고 한다. 예를 들어, '채식가에게는 농촌 생활이 좋다'라는 문장을 생각해 보자. 이 문장에서 '채식가에게는'을 빼놓는다면 '농촌 생활이 좋다'는 '모든 사람에게 농촌 생활이 좋다'는 의미를 전달할 수 있어서, 이 문장의 원래 의미를 전달하지 못한다. 즉, '채식가에게는'이 '농촌 생활이 좋다'의 의미를 적절하게 한정하는 문맥이 된다. 최근 들어 학문 분야 간의 융합이 강조되고 있는 추세 속에서, 데이터에 대한 문맥 정보는 타분야 연구자에게 데이터 이해의 용이성을 제공하며 궁극적으로 데이터 재사용에 결정적인 역할을 한다[3][4]. 이러한 이유로 데이터 재사용성을 높이기 위해 문맥 중심으로 데이터 큐레이션을 수행하는 것이 대세를 이루고 있다. 연구 데이터에 대한 문맥 정보는 실험 방법, 장치, 데이터에 간략한 서술, 약어집(codebook), 데이터 포맷, 데이터 분석 도구 등을 말한다. 문맥 중심의 데이터 큐레이션을 성공적으로 달성하려면, 큐레이션 활동에 관련된 이해당사자의 지식과 경험의 다양성에 주목할 필요가 있다. 앞에 언급한 데이터 생산자, 데이터 사서, 데이터 재사용자는 대표적인 이해당사자로, 이들은 각자 서로 다른 전문성을 가지고 데이터에 대한 문맥 정보를 생성하여 데이터의 이해 및 재사용에 기여한다.

일반적으로, 재사용성이 높은 생명과학 데이터는 분야에 독립적인 메타데이터뿐만 아니라 분야에 특정적인 메타데이터를 모두 포함하고 있으며, 후자는 도메인 전문가의 지식과 경험에 의한 문맥 정보로 구성된다[5][6]. 이러한 문맥 정보의 일부는 데이터 컬렉션의 분석 결과로부터 산출된 문헌에 서술되어 있고, 데이터 컬렉션의 생성에 사용된 기존 리파지토리의 데이터는 산출된 논문에 의해 인용된 참고문헌에 서술되어 있다. 따라서 문맥 정보에 의한 데이터 큐레이션은 다수의 관련 논문을 선정하고 이들로부터 메타데이터에 해당하는 데이터 항목을 추출하여 메타데이터로 주석하는 것을 내포한다[7-9]. 데이터-문헌 간의 상호연결은 데이터와 문헌 간의 연관성에 의해 이들을 연결하는 것으로[10][11], 데이터-문헌 간의 상호연결 정보는 어떤 데이터가 어느 문헌에서 논의된 연구 결과를 뒷받침하는 근거인지, 어떤 문헌이 어느 데이터를 인용하는지 등의 정보를 제공한다. 이들 상호연결 정보는 데이터의 이해는 물론이고 재사용에 기여한다.

큐레이션은 데이터의 이해를 돕기 위한 설명뿐만 아니라, 데이터의 신뢰성(trustworthiness)도 보장해야 한다. 데이터의 신뢰성은 이에 근거한 연구결과의 투명성 및 재현성과 밀접한 관계가 있다. 데이터 신뢰성을 보장하기 위해, 연구 결과의 근거가 되는 데이터를 사전에 제출하여 검증 절차에 따라 제3자 검토를 거친다[12][13]. 데이터에 대한 제3자 검토는 자연과학은 물론이고 심리학 및 경제학을 비롯한 사회 과학 분야에서도 실천하고 있다. 검증을 위해 연구자는 생성된 데이터와

함께, 실험 설계, 방법, 장치 등 문맥 정보를 메타데이터로 제출한다. 만일 기존의 데이터 리파지토리로부터 데이터를 수집하여 데이터 컬렉션을 생성했다면, 데이터와 함께 이에 관한 문맥 정보를 수집하여 제출한다. 제출된 데이터의 검증은 전문 지식과 경험을 가진 분야 전문가에 의한 제3자 검토 과정을 수반한다.

위의 서술과 같이 데이터 재사용성을 위한 큐레이션은 이해하기 쉽고 믿고 사용할 수 있는 데이터를 공개하기 위해 데이터 생산자, 데이터 사서, 데이터 재사용자 등 다수의 전문가로 구성된 연구자 커뮤니티가 주도하는 협업적 방법을 요구한다. 본 논문에서는 과학 데이터에 대한 문맥 중심의 커뮤니티 주도적 큐레이션 협업 환경의 개발 및 적용에 대해 다룬다. 데이터의 이해 가능성과 재사용성을 극대화하기 위해, 데이터-문헌 간의 상호연관성 및 데이터에 대한 문맥을 중심으로 큐레이션 데이터 모델을 제시한다. 데이터와 문헌 간의 상호연관성은 이들 간의 DOI(Digital Object Idenfier)[14][16]를 서로 연결하여 영구적 접근성을 보장한다. 또한 연구 커뮤니티에 의한 데이터의 신뢰성을 보장하기 위한 데이터 검토 프로세스(data review process)를 제시한다. 큐레이션 데이터 모델과 데이터 검토 프로세스를 지원하는 문맥 중심의 커뮤니티 주도적 데이터 큐레이션 협업 환경의 구조를 제시한다. 이중에서 필수적인 구성 요소를 중심으로 데이터 큐레이션 환경으로 구현하여, 생명과학 분야의 무정형 단백질 데이터에 적용을 통해 무정형 단백질 데이터베이스를 구축하였다.

과학 데이터 큐레이션 협업 환경을 기반으로 구축된 무정형 단백질 DB를 통해 연구자들은 데이터의 검색 및 수집에 드는 노력을 절감하고 연구에 더욱 많은 노력을 집중할 수 있다. 무정형 단백질 DB는 국제적으로 저명한 선도 연구자가 참여하는 커뮤니티에 의해 그들의 연구 결과와 함께 가능한 한 풍부한 문맥 정보를 제공할 수 있게 되었다.

본 논문의 구성은 다음과 같다. II장 관련 연구에서 과학 데이터 공개를 위해 적용되는 원칙을 소개하고 본 큐레이션 협업 환경이 추구하는 것이 이들 원칙을 만족시키기 위한 것임을 서술한다. 이를 실현하기 위한 연구 방법을 III, IV, V, VI장에 서술하고, VII장에서 끝맺는다.

II. 관련 연구

과학 분야에서 데이터에 대한 이해 및 활용성 제고에 초점을 맞춘 데이터 큐레이션 연구가 다양하게 진행되어 왔으며, 그에 대한 대표적인 결실이 FAIR[17] 가이드라인이다. FAIR는 Findability, Accessibility, Interoperability, Reusability의 머리글자를 딴 약어로, 연구 데이터의 큐레이션 지침으로 제시되어 사용되고 있다. Findability는 데이터의 영구적 발견 가능성을 의미하는 것으로, 영구적 식별자를 부여하여 실현한다. 일반적으로 데이터에 대한 영구적 식별자로 DOI를 사용한다. Accessability는 이 식별자를 가지고 데이터 및 메타데이터를 검색할 수 있어야 한다는 의미이다. DOI의 적용은 글로벌 루트를 통해 접근이 가능하기 때문에 Findability와 Accessability를 보장한다. Interoperability는 사람의 처리 가능성(human actionability)뿐만 아니라 기계 처리 가능성(machine actionability)까지 고려한 것으로, XML, JSON, RDF 등 표현 방법뿐만 아니라 통제 어휘나 온톨로지의 사용을 권장하고 있다. Reusability는 도메인 표준 온톨로지 및 어휘, 데이터의 족보(lineage)를 나타내는 데이터 프로비넌스(provenance), 정확하고 풍부한 메타데이터 등을 포함하고 있다.

본 연구에서 제시하는 과학 데이터 큐레이션 협업 환경에서는 식별자를 DOI로 사용하고 있으며, 커뮤니티가 주도적으로 생성하는 도메인 특정적 메타데이터를 문맥정보로 제공하고 있다. 이로써 기계 처리 가능성과 데이터 프로비넌스를 제외한 모든 FAIR 원칙을 만족시킨다.

문맥 정보에 대한 언급은 매우 다양하며, 연구 분야에 특정적인 메타데이터는 특히 그러하다. 최근 데이터 개방이 화두로 떠오르면서 개방을 위한 기본적인 메타데이터에 문맥 정보를 요구하는 경향이 점차 뚜렷해지고 있다. OpenContext[4]는 원래 고고학 분야의 데이터 큐레이션을 위한 메타데이터로 정의되었으나, 인문학

분야의 데이터 이해에 기본적이고 충실한 메타데이터를 담고 있는 것이 특징이다. 본 연구의 데이터 모델에는 이들 메타데이터를 반영하였다.

기존의 데이터 검토 프로세스는 데이터의 재사용을 위해 주석된 메타데이터의 충실성에 중점을 두는 반면, 본 연구에서 제시하는 데이터 검토 프로세스는 재사용뿐만 아니라 다른 과학자가 믿고 사용할 수 있도록 데이터의 신뢰성 확보에 중점을 두고 있다. 이러한 이유로 본 연구에서는 과학 데이터의 대표적인 사이트인 PDB, MBRB 등의 사례에 기초하여 제3자 검토 모델을 데이터 검토 프로세스로 제시하였다. 이들 사이트의 특징은 데이터 공개가 논문 출판 프로세스와 연계되어 있다는 것이다. 이런 사례에서는 연구자의 업적을 내기 위한 전제 조건으로 데이터의 공개 및 제3자 검토를 내포하기 때문에 데이터 신뢰성 확보가 쉬운 경향이 있다.

Ⅲ. 과학 데이터 큐레이션 협업 환경의 개발 요구 사항

연구자들은 기존 리파지토리의 데이터를 재사용하거나 자신의 데이터(컬렉션)를 검증하기 위한 목적으로 전문 분야의 데이터 기탁(deposit) 사이트를 구축 및 활용하고 있다. 분야마다 차이는 있겠지만, 생명 과학 분야의 데이터 기탁 사이트는 단순한 데이터 저장소의 의미뿐만 아니라 제3자 검토의 의미를 포함하고 있다. GenBank[6], Protein Databank(PDB)[5] 등이 대표적인 사이트로, 데이터의 이해를 돕기 위한 메타데이터를 충분히 제공하고 있으며, 궁극적으로 연구자들의 과학적 발견에 도움이 되고 있다. 이와 같이 데이터의 이해 및 재사용을 위해 필요한 충분한 메타데이터를 제공하기 위해서는 분야 전문가를 비롯한 연구 커뮤니티가 주도하는 강도 높은 데이터 큐레이션 협업 환경을 요구하고 있다. 생명과학 분야의 연구 커뮤니티로부터 식별한 데이터 큐레이션 협업 환경의 기본적인 요구 사항은 다음과 같다.

1. 문맥 정보에 의한 데이터 재사용성 제고

데이터의 재사용성을 제고하려면, 사용자가 쉽게 데이터를 이해할 수 있고 원하는 포맷으로 변환하여 다른 데이터와 통합하여 사용할 수 있도록 데이터에 대한 충분한 문맥 정보를 메타데이터로 제공해야 한다. 이러한 메타데이터는 데이터 생산자, 데이터 관리자, 데이터 재사용자 등 데이터에 관한 이해당사자에 따라 문맥 정보 제공을 위한 메타데이터가 다르다. 데이터 생산자는 자신이 설계한 실험 방법을 통해 생성한 데이터, 데이터 생성에 사용된 장비나 장치, 자신이 생성한 데이터와 다른 데이터 리파지토리로부터 수집한 데이터를 통합한 데이터 컬렉션, 데이터 컬렉션을 위해 사용된 데이터 리파지토리 및 통합 도구, 데이터 컬렉션의 해석을 위해 사용된 분석 도구 등을 중요한 문맥 정보로 간주한다. 이 밖에도 데이터 생산자가 제공하는 메타데이터는 데이터 컬렉션(생성 또는 수집된 데이터), 생성된 데이터에 관한 서술, 서술에 사용된 데이터 코드집, 데이터를 생성한 실험 설계 및 절차, 샘플, 관찰 기록 등이고, 데이터 관리자가 제공하는 메타데이터는 데이터 컬렉션, 제목, 데이터에 관한 서술, 생산자, 출판사, 날짜, 저작권, 분류, 주제, 위치, 접근 방법 등이며, 재사용자가 제공하는 메타데이터는 데이터 컬렉션, 데이터 생산자, 데이터 프로비넌스(데이터 재사용에 의한 데이터 컬렉션의 생성 과정 또는 데이터 족보), 리파지토리 정보(리파지토리의 이름, 위치, 접근 방법 등에 대한 정보), 이전의 활용 사례(prior use case) 및 이에 관련된 연구 목적, 데이터 오류, 데이터 재사용 시의 주의사항 등이다. 문헌에 대해서도 문맥 정보를 고려해 볼 수 있다. 분야의 전문성을 갖춘 연구자는 기존의 문헌을 조사할 때 본문 전체를 읽지 않고 이해하는 데 필요한 부분만을 선택적으로 읽는다. 예를 들어, 초록, 서론의 연구 동기, 데이터 분석에 의해 산출된 그림 및 표, 이에 관련된 캡션 등에 중점을 둔다. 이들이 문헌에 대한 대표적인 문맥 정보라고 할 수 있다. 데이터 큐레이션 협업 환경은 이와 같이 다양한 관점의 문맥 정보를 풍부하게 제공하여, 사용자의 데이터의 이해를 돕고 데이터의 재사용성을 제고하고자 한다.

2. 데이터–문헌간 상호연결 정보관리 및 영구적

접근성 보장

연구자는 일상적으로 연구 주제와 관련 있는 기존의 문헌을 분석하여 자신의 연구 문제를 정의하고, 데이터를 생성 또는 수집하고 분석하여 자신의 연구를 증명한다. 즉, 자신이 설계한 실험 방법이나 장치에 의해 데이터를 생성하고, 기존 리파지토리로부터 관련 데이터를 수집하여 자신이 생성한 데이터와 통합하여 데이터 컬렉션을 생성한다. 통계 분석 프로그램을 비롯한 분석 도구를 사용하여 이들 데이터를 분석한 후, 해당 분야의 전문 지식과 경험을 바탕으로 분석 결과를 해석한다. 또한 기존의 관련 연구와 비교 분석을 통해 자신의 연구를 차별화한다. 이와 같이 데이터와 문헌은 연구 결과의 가치를 결정하는 중요한 역할을 한다. 협업 환경은 데이터와 문헌 간의 연관성을 중심으로 데이터-문헌 간의 상호연결 정보를 관리하고, 이들 정보에 대한 영구적 접근성을 보장하고자 한다. 이렇게 하면, 연구자의 데이터와 문헌 정보에 대한 지속적인 재사용 및 인용이 가능하다.

3. 데이터 검토 프로세스에 의한 데이터 신뢰성 확보

사용자가 데이터를 사용하려면 믿을 수 있어야 한다. 신뢰성 확보 방법은, 분야에 따라 다르지만, 생명과학 분야에서는 GenBank, PDB 등 기탁 사이트를 통해 데이터를 공개하기 전에 제3자 검토(peer review) 프로세스를 거친다. 데이터의 신뢰성(trustworthiness)은 이와 같은 기탁에 의해 보장될 수 있으며, 제3자 검토 프로세스는 해당 분야의 전문가에 의해 수행된다. 제3자 검토 프로세스는 데이터 생산자의 데이터(컬렉션)의 제출, 분야 전문가의 심사, 승인 또는 거절 단계로 구성된다. 데이터 생산자는 제출시 데이터와 함께 데이터 설명서(data descriptor: data paper라고도 함)를 기탁 사이트에 제출한다. 데이터 설명서는 데이터 생성 및 수집에 관련된 실험 방법, 장치, 데이터 리파지토리, 프로그램 도구 등 데이터의 검증뿐만 아니라 재사용에 필요한 다양한 메타데이터를 담고 있다. 분야 전문가는 데이터(컬렉션) 및 데이터 설명서에 대한 심사를 거쳐 심사 보고서를 작성하여 데이터 생산자에게 피드백한다. 심사 결과에 따라 승인되거나 수정 및 재심사를 거치기도 한다. 이와 같이 데이터 검토 프로세스는 논문 심사 프로세스와 유사하게 보일 수 있지만, 위에 언급한 데이터에 대한 서술에 근거하여 데이터의 검증에 중점을 두고 검토하기 때문에 논문의 제3자 검토에 비해 훨씬 단순하다. 협업 환경은 제3자 검토 프로세스를 지원하여 데이터의 신뢰성 보장하고자 한다.

Ⅳ. 데이터-문헌 상호연결 및 문맥 정보 기반의 큐레이션 데이터 모델

데이터 큐레이션은 데이터에 대한 문맥 정보, 데이터-문헌 간의 상호연결 정보를 메타데이터로 생성하고 관리하는 활동이다. 먼저 데이터에 대한 문맥정보를 고려해 보자. 새로 생성된 데이터는 측정값의 집합인 실험 절차, 실험에 사용된 장비나 장치, 실험 조건, 장소, 일자, 데이터 포맷, 주요 측정값에 대한 설명 등을 메타데이터로 포함하고 있다. 반면, 기존의 리파지토리로부터 수집된 데이터는 앞에 언급된 메타데이터 이외에 리파지토리 정보, 접근 방법, 분석 도구 등을 메타데이터로 포함하고 있다.

데이터-문헌 간의 상호연결 정보는 어떤 데이터가 어느 문헌에 근거가 되고 있는지를 나타내는 근거 관계를 표현하거나, 어떤 데이터가 어느 데이터를 인용하고 있는지를 나타내는 인용 관계를 표현한다. 데이터와 데이터 간의 인용 관계는 데이터 컬렉션을 생성할 때 사용된 데이터 간의 재사용 관계를 나타낸다. 최근에 데이터 설명서의 출판이 증가하는 상황에서 이것을 인용 관계라고 간주할 수도 있다.

큐레이션 데이터 모델은 데이터의 문맥 정보와 데이터-문헌 간의 상호연결 정보를 모두 반영하여 설계하였다. 데이터 및 문맥 정보에 대한 영구적 접근성을 보장하기 위해, 데이터에 DOI를 부여하고 이에 대한 문맥 정보를 속성으로 모델링하였다. 문헌은 데이터와 마찬가지로 문맥 정보를 포함하고 있으나, 본 논문에서는 제목, 연구자, 초록, 키워드, 연구 주제, 표, 그림, 연구 재단으로 제한한다. 문헌은 출판 당시에 DOI를 부여한 상태에서 출판하기 때문에 출판된 모든 문헌은 DOI가

부여되어 있다고 가정한다. 따라서, 데이터-문헌 간의 상호연결 정보는 데이터 DOI와 문헌 DOI 간의 매핑 테이블로 모델링한다. 이와 같이 모델링한 것을 개체-관계 모형으로 간략히 표현하면 [그림 1]과 같다.

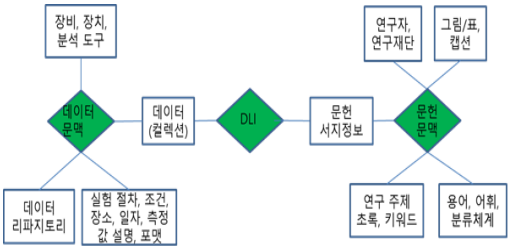

그림 1. 큐레이션 데이터 모델

V. 연구 커뮤니티 중심의 데이터 검토 프로세스 모델링

데이터 재사용자가 데이터를 믿고 활용할 수 있으려면 데이터의 신뢰성이 확보되어야 한다. 이를 위해 데이터 검토 프로세스는 해당 분야 전문가의 제3자 검토를 필요로 한다. 전통적인 데이터 검토 프로세스로 PDB를 예로 들 수 있다. 연구자가 검토에 필요한 각종 메타데이터와 함께, 단백질 3차 구조 데이터, 실험 데이터, 검증 프로그램에 의한 계산 결과값을 PDB 사이트에 제출하면, PDB 사이트는 이들 제출된 데이터에 대해, PDB내에서 사용하는 식별자(PDB ID)를 할당하고 연구자에게 반환한다. 분야 전문가로 구성된 심사위원들이 이들 데이터를 심사 및 검증한다. 검토 결과를 연구자에게 피드백하고, 필요 시 수정을 요청하거나 승인한다. 승인된 데이터를 PDB 엔트리로 저장하고, 보존 사이트나 관련 사이트에게 배포한다.

본 연구에서 모델링한 데이터 검토 프로세스도 이와 유사하지만, 앞에 언급한 연구 커뮤니티의 요구 사항을 충족시키기 위해, 몇 가지를 추가적으로 고려하였다. 우선, 앞의 큐레이션 데이터 모델에서 언급한 문맥 정보와 데이터 문헌 간의 연결 정보를 충분히 제공하기 연구자에게 관련 메타데이터의 주석을 큐레이션 DB 사이트에 직접 반영하도록 요구하고 있다. 그리고, 승인된 데이터에 국제 표준 식별 체계를 할당하여, 영구적 접근성을 보장한다. PDB에서 PDB ID는 제출하면 즉시 할당되지만, 본 모델에서는 승인된 데이터에 한해 PID(persistent ID)를 할당한다. 끝으로, PID가 할당되면 연구자는 이 데이터를 PID 시스템에 등록하여 국제적으로 재사용할 수 있도록 개방한다.

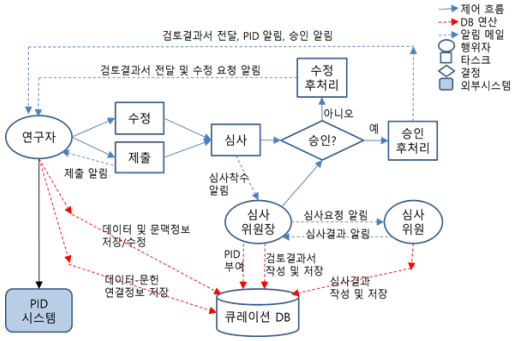

그림 2. 데이터 검토 프로세스 모델

VI. 커뮤니티 주도적 큐레이션 협업 환경의 설계 및 구현

1. 주요 사용자 기능

1) 사용자 등록(ORCID 포함): 회원 가입, 사용자 역할 및 권한 요청, 이메일 로그인
2) 검색: 데이터 이름, 데이터 생산자, 메타데이터, 키워드, 문헌 DOI, 문헌 제목, 문헌 저자를 기반으로 하는 검색 기능
3) 재사용
   - 화면표시: 검색결과 리스트(데이터 이름, 데이터 DOI, 데이터 집합(데이터컬렉션), 메타데이터, 데이터를 근거로 하는 문헌, 데이터 생산자 등)
   - 내려받기: 데이터를 특정 포맷(주로 CSV 포맷)으로 내려받기
4) 제출: 제출자는 데이터, 메타데이터, 검증 결과 등을 제출
   - 검증 프로그램 실행: 웹 서비스로 제공된 계산 프

로그램을 수행하여 데이터를 검증하고, 그 결과를 이메일로 받음
- 메타데이터 및 데이터 제출: 검증 결과를 비롯하여 지정된 메타데이터와 데이터를 DB에 저장
- 심사상태 변경 및 제출 알림: 시스템은 상태를 'submitted'로 변경하고, 제출 사실을 제출자와 심사위원장에게 이메일로 알림

5) 심사: 심사위원장이 심사기준에 근거하여 수행
- 심사관리번호 부여: 제출된 데이터에 심사관리 번호 부여
- 검증 프로그램 실행 및 실행결과 이메일 알림
- 심사위원 선정 및 전달: 심사위원장이 수작업으로 실행
- 심사결과 작성: 심사위원이 작성하여 DB에 저장
- 검토보고서 작성: 심사위원장은 심사 결과를 바탕으로 검토보고서를 작성하여 DB에 저장
- 심사상태 변경 및 검토결과 알림: 심사위원장은 상태를 'accepted/rejected/change/revised' 중의 하나로 변경하고, 제출자에게 검토결과를 이메일로 알림

6) 수정: 제출자는 답변서를 작성하여 심사위원장에게 알림
- 답변서 작성: 제출자는 검토 결과에 대한 답변서를 작성하여 DB에 저장하며, 재수정 시 새로운 답변 내용을 추가하여 저장할 수 있도록 Append 기능 제공
- 수정된 데이터 제출: 필요시, 기존 데이터 Update
- 상태 변경 및 답변서 제출 알림: 시스템은 상태를 'revised'로 변경하고, 심사위원장에게 답변서 제출 사실을 이메일로 알림

7) 승인된 데이터에 DOI 할당 및 등록
- 승인된 데이터에 DOI 할당: 기탁 관리자가 승인된 데이터에 대해 DOI를 부여
- 상태 변경 및 승인 알림: 심사위원장이 제출자에게 승인 및 부여된 DOI를 이메일로 알림
- DOI 등록: 제출자가 DOI 등록관리 시스템에 DOI, 메타데이터, 인용정보를 등록

8) 저장 (주석 및 보존): 사용자는 각자의 역할에 따라, 수집 및 생성, 재사용 및 기탁관리에 요구되는 데이터-문헌 상호연결 및 문맥 정보 등 메타데이터를 생성하여 DB에 저장
- 데이터 생산자(제출자): 연구 분야에 특정된 메타데이터 및 문맥 정보를 생성하여 DB에 저장
- 심사위원장: 데이터 보존 관리에 따른 메타데이터 추가
- 데이터 재사용자: 재사용에 따른 메타데이터 추가

2. 큐레이션 협업 환경의 구조

데이터와 문헌 간의 연결 정보는 이들의 식별자 간의 매핑으로 표현할 수 있다. 연구자들은 데이터, 문헌 및 이들 강의 정보를 영구적으로 접근하기를 원한다. 이를 위해 문헌뿐만 아니라 데이터에도 DOI(digital object identifier)를[14] 할당하고, 데이터 및 문헌 간의 매핑을 DOI를 이용하여 표현한다. 이들을 DOI 시스템에 등록하여 영구적 접근을 보장한다. 데이터-문헌 간의 연결 정보 관리를 위한 DOI 시스템을 중심으로, 큐레이션 데이터 모델 및 데이터 검토 프로세스 모델을 지원하는 데이터 큐레이션 협업 환경의 구조는 [그림 3]과 같다.

DOI 시스템은 Global Root Server와 DOI RA (registration agency)의 Local Handle Server로 구성되어 있다. Global Handle Registry와 DOI RA의 Local Store 간의 연합을 통해 상호 연결(Cross-Link)에 의한 영구적인 데이터 연결성 및 접근성을 보장한다. 데이터 큐레이션은 메타데이터의 일관성 및 정확성 등 분야 전문가 커뮤니티의 협력에 의한 데이터 검토 프로세스를 필수적으로 수반한다. Deposit Process Management는 데이터 검토 프로세스에 따라 타스크의 처리 및 상태 관리 기능을 수행한다. Curated Data Management는 각종 문맥 정보를 위한 메타데이터 관리를 담당한다. 이들을 웹 환경으로 지원하기 위해 웹 기반의 큐레이션 인터페이스를 제공한다. 이를 통해 사용자는 데이터의 기탁 시 메타데이터 정의 및 주석, 데이터의 신뢰성을 담보하기 위한 제3자 검토, 데이터 및 문헌에 의한 검색, 데이터-문헌 간의 상호 연결 정보에 의한 탐색, 데이터 재사용 등을 수행한다. Information Extraction & Organization은 문헌으로부터 문맥 정보의 추출과 재구

성을 위한 텍스트 분석 및 마이닝 도구, 주석을 위한 통제 어휘 등을 제공하기 위한 것으로, 별도의 외부 타스크에 의존한다. 그 이유는 자동화 도구의 활용이 큐레이션의 정확성을 떨어뜨리기 때문이다. 일반적으로 고강도의 큐레이션에는 분야 전문가의 수작업에 의존하고 있다. 외부 타스크의 적용으로 사용자는 탄력적으로 자동화 도구를 사용할 수 있다.

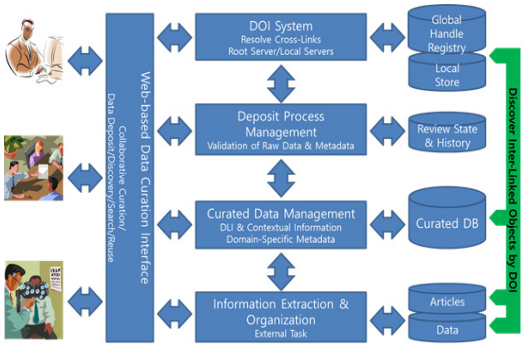

그림 3. DOI 중심의 데이터 큐레이션 협업 환경 구조

3. 구현 및 적용 사례: IDP DB 구축

무정형 단백질(IDP, intrinsically disordered protein)은 단백질 구조에 관한 정설을 깨는 새로운 특성을 갖는 단백질이다. 단백질은 3차 구조를 가질 때 기능을 한다는 것이 기존의 정설이었으나 2차 구조를 갖는 단백질도 특정 기능을 한다는 것이 새롭게 밝혀졌다[15]. 이러한 단백질을 무정형 단백질이라고 한다. 이에 관한 연구는 30년 이상 지속되고 있으며 이미 60여개 이상의 무정형 단백질이 알려졌다. 이들 단백질은 PDB에 기탁되어 있지 않기 때문에, 연구자는 연구 수행 중에 무정형 단백질인 듯한 것을 밝혀냈다고 하더라도 이것이 새로운 것인지 아닌지 알아내려면 많은 노력을 기울여야 한다. 이러한 어려움을 완화하기 위해, 무정형 단백질 연구 커뮤니티가 중심이 되어, 무정형 단백질에 관한 문맥 정보를 충분히 제공하여 이해의 용이성을 도모하는 동시에 데이터에 DOI를 적용하여 영구적으로 접근할 수 있도록, 데이터의 재사용을 우선적으로 고려한 IDP DB를 구축한다. IDP DB 스키마는 [그림 4]와 같다.

그림 4. IDP DB 스키마

IDP DB는 현재 프로토타입 수준인 데이터 큐레이션 협업 환경을 활용하여 구축되었으며, 큐레이션 데이터 모델과 데이터 검토 프로세스 모델의 타당성을 검증하였다. 현재 60여편의 논문으로부터 추출된 50여개의 IDP 관련 데이터를 포함하고 있으며, 레코드 일부를 보여주고 있다[16].

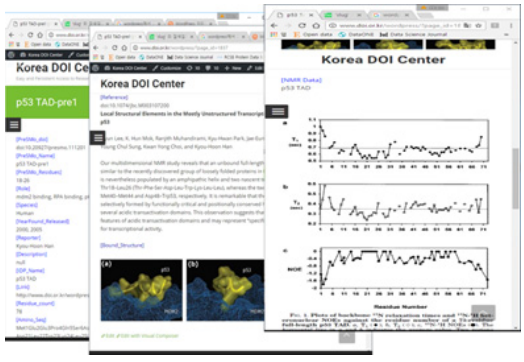

그림 5. IDP DB 웹페이지의 예

Ⅶ. 결론 및 향후 연구

본 논문에서는 과학 분야의 연구자 커뮤니티가 주도하는 데이터 큐레이션을 지원하기 위한 협업 환경의 개발 및 적용 사례를 제시하였다. 과학 데이터의 이해 가

능성과 재사용성을 극대화하기 위해, 데이터에 대한 문맥 정보를 메타데이터로 정의, 생성 및 관리할 수 있고 데이터-문헌 간의 상호연결 정보를 DOI 간의 매핑으로 생성 및 관리할 수 있도록, 큐레이션 데이터 모델을 제시하였다. 분야 전문가에 의한 제 3자 검토를 통한 데이터의 신뢰성을 보장하기 위해 연구 커뮤니티 중심의 데이터 검토 프로세스를 제시하였다. 연구 커뮤니티의 데이터 큐레이션 요구사항에 따라 큐레이션 협업 환경의 사용자 기능을 도출하였고, 큐레이션 데이터 모델과 데이터 검토 프로세스 모델을 중심으로 이들 기능을 지원할 수 있도록 커뮤니티 주도적 데이터 큐레이션 협업 환경을 설계하여 개괄적인 구조를 제시하였다.

연구 커뮤니티 주도적 과학 데이터 큐레이션 환경의 프로토타입을 구현하여, 생명과학 분야의 IDP DB 웹사이트 구축에 적용하였다. 구축 중에 IDP DB 웹사이트의 사용성 및 데이터 재사용성을 제고하기 위해, IDP 커뮤니티와 함께, 분석 및 설계 검토/구현 테스트를 수행하였고 데이터 및 메타데이터에 대한 FAIR 원칙을 검토하였다. FAIR 원칙의 검토를 통해 본 협업 환경이 IDP는 물론이고 타분야 과학 데이터의 개방에 적용 가능함을 공동으로 인식하게 되었다. 이를 위해서는, 도메인에 특화된 메타데이터를 각 분야의 커뮤니티가 정의해야 한다.

향후, 큐레이션 협업 환경의 정교한 구현을 위해, 연구 커뮤니티 내의 메일 시스템, 기존 학술정보 DB, 텍스트 분석 및 마이닝 도구를 비롯한 외부 시스템과 연동을 우선적으로 고려해야 한다. 큐레이션 협업 환경은 메일 시스템과의 연동을 통해 큐레이션 DB에 데이터를 저장함과 동시에 이해 당사자에게 알림 메일을 자동으로 보낼 수 있으며, 기존 학술 정보와의 연동을 통해 필요한 서지 정보의 중복 입력 및 저장 노력을 제거할 수 있다. 텍스트 분석 및 마이닝 도구와의 연동으로, 큐레이션 협업 환경은 데이터-문헌 간의 연결 정보에 근거하여 관련 문헌으로부터 연구 주제에 밀접한 중요한 어휘를 자동으로 추출할 수 있다[18]. 여기에 통제 어휘를 사용하면 메타데이터 주석의 모호성을 제거하여, 데이터 이해의 용이성을 제고할 수 있다. 끝으로, IDP DB의 공개에 앞서 연구 커뮤니티 구성원들에게 무정형 단백질 DB 웹사이트에 대한 가치 및 유용성 평가를 체계적으로 수행해야 할 것이다.

참 고 문 헌

[1] B. Howe and T. Lewis, "Enabling Collaborative Research Data Management with SQLShare, 2012, https://www.slideshare.net/billhoweuw/research-data-managementi22012

[2] I. Faniel, D. Minor, and C. L. Palm, "Putting Research Data into Context: Scholarly, Professional, and Educational Approaches to Curating Data for Reuse," ASIST 2014.

[3] I. Faniel, E. Yakel, K. Fear, and E. Kansa, "A Context-driven Approach to Data Curation for Reuse," International Digital Curation Conference, Amsterdam, February 22, 2016.

[4] I. Faniel, E. Kansa, S. W. Kansa, J. Barrera-Gomez, and E. Yakel, "The Challenges of Digging Data: A Study of Context in Archaeological Data Reuse," JCDL 2013, pp.295-304.

[5] http://www.rcsb.org/pdb/home/home.do

[6] https://www.ncbi.nlm.nih.gov/genbank

[7] M. E. Cusick, "Literature-curated protein interaction datasets," Nat Methods, Vol.6, No.1, pp.39-465, 2009.

[8] D. S. Kwon, S. Kim, S. Y. Shin, Andrew Chatr-aryamontri, and W. John Wilbur, "Assisting manual literature curation for protein-protein interactions using BioQRator," Database, 2014.

[9] D. G. Jamieson, M. Germer, F. Sarafraz, G. Nenadic, and D. L. Robertson, "Towards semi-automated curation: using text mining to recreate the HIV-1, human protein interaction database," Database, 2012.

[10] M. S. Mayernik, J. Phillips, and E. Nienhouse,

"Linking Publications and Data: Challenges, Trends, and Opportunities," D-Lib Magazine, Vol.22, No.5/6, 2016(11).

[11] M. Hoogerwerf, M. Lösch, J. Schirrwagen, S. Callaghan, P. Manghi, K. Iatropoulou, D. Keramida, and N. Rettberg, "Linking Data and Publications: Towards a Cross-Disciplinary Approach," The International Journal of Digital Curation, Vol.8, No.1, 2013.

[12] B. Lawrence, C. Jones, B. Mathews, S. Palmer, and S. Callaghan, "Citation and Peer Review of Data: Moving Towards Formal Data Publication," The International Journal of Digital Curation, Vol.6, No.2, 2011.

[13] H. M. Berman, J, Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The Protein Data Bank," Nucleic Acids Research, Vol.28, No.1, pp.235-242, 2000.

[14] https://www.doi.org

[15] H. Lee, K. H. Mok, R. Muhandiram, K. H. Park, J. E. Suk, D. H. Kim, J. Chang, Y. C. Sung, K. Y. Choi, and K. H. Han, "Local Structural Elements in the Mostly Unstructured Transcriptional Activation Domain of Human p53," The Journal of Biological Chemistry, Vol.275, No.38, pp.29426-294323, 2000.

[16] https://www.doi.or.kr

[17] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, K. Blomberg, J. W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, C. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. G. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. G. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S. A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, and B. Mons, "The FAIR Guiding Principles for scientific data management and stewardship," Scientific Data 2016.

[18] Life Science Solutions, "Automated vs manual literature curation: extracting more information from scientific literature," Elsevier, 2014.

저 자 소 개

최 동 훈(Dong-Hoon Choi) 정회원

- 1981년 2월 : 서울대학교 계산통계학과(학사)
- 1983년 2월 : 한국과학기술원 전산학과(석사)
- 1989년 6월 : 노스웨스턴대학교 전산학과(박사)
- 2004년 2월 ~ 현재 : 한국과학기술정보연구원 책임연구원

<관심분야> : 데이터베이스

박 재 원(Jae-Won Park) 정회원

- 1998년 2월 : 성균관대학교 정보공학과(석사)
- 1998년 : ㈜삼창기업 부설연구소 (연구원)
- 1999년 2월 ~ 2000년 12월 : 연구개발정보센터(연구원)
- 2001년 1월 ~ 현재 : 한국과학기술정보연구원 선임연구원

<관심분야> : 정보공학, 국제표준식별체계

김 병 규(ByungKyu Kim) 정회원

- 2002년 2월 : 충남대학교 컴퓨터과학과(석사)
- 2002년 12월 ~ 2007년 5월 : 한국과학기술정보연구원(연구원)
- 2007년 5월 ~ 현재 : 한국과학기술정보연구원 선임연구원

<관심분야> : 학술콘텐트 인용분석 및 응용연구

신 진 섭(Jin-Sup Shin) 정회원

- 2000년 8월 : 충남대학교 화학과(석사)
- 2005년 2월 : 충남대학교 컴퓨터과학과(석사)
- 2010년 9월 ~ 현재 : 한국과학기술원(박사과정)
- 2005년 7월 ~ 현재 : 한국과학기술정보연구원 선임연구원

<관심분야> : 과학 데이터 서비스, 국제표준식별체계