https://doi.org/10.5392/JKCA.2020.20.02.689

학술논문 통합 DB 구축을 위한 메타데이터 스키마 비교 분석

Comparison and Analysis of Metadata Schema for Academic Paper Integrated DB

최원준, 황혜경, 김정환, 이강산다정, 임석종
한국과학기술정보연구원, 콘텐츠큐레이션 센터

**Wonjun Choi(cwj@kisti.re.kr), Hyekyong Hwang(hkhwang@kisti.re.kr),
Jeonghwan Kim(kimjh@kisti.re.kr), Kangsandajeong Lee(lksdj@kisti.re.kr),
Seokjong Lim(seoklim@kisti.re.kr)**

요약

국내외 학술논문을 서비스하는 국가과학기술정보센터(NDSL) 데이터베이스는 다양한 정보원으로부터 수집된 데이터가 분산적으로 수집, 구축 및 관리되고 있다. 본 연구에서는 분산된 학술논문 DB를 분석하여 논문데이터를 고부가가치화하고 효율적으로 관리할 수 있는 통합 DB 스키마 도출을 위하여 현재 구축되고 관리되는 학술논문 DB 스키마 및 DB 메타데이터를 분석하였다. 또한, 현재 구매하여 보유하고 있는 Web of Science와 SCOPUS 스키마를 활용하여 비교, 분석을 통한 최종 학술정보 데이터 항목을 정하였다. 본 연구를 통하여 구축되고 서비스되는 학술정보 데이터 항목이 논문, 저자, 초록, 기관, 주제, 저널, 참고문헌 7가지로 요약 도출되었으며 구축중인 핵심콘텐츠로 정의하였다. 본 연구를 통하여 통합 DB 스키마가 만들어졌으며, 향후 이 연구 결과는 고품질의 학술논문 통합 DB 컬렉션 구성과 시스템 최적화 설계를 위한 기반 자료로 활용하고자 한다.

■ 중심어 : | 과학기술정보 | 논문 데이터 | 학술논문 | 문헌 정보 | 정보 서비스 |

Abstract

The National Science and Technology Information Center (NDSL) database, which provides academic papers at home and abroad, collects, builds, and manages data collected from various sources. In this study, we analyzed the DB paper schema and DB metadata that are currently constructed and managed to derive an integrated DB schema that can manage the high-value-added papers and manage them efficiently by analyzing distributed DB papers. Also, the final academic information data items were determined through comparison and analysis using the Web of Science and SCOPUS schemas that are currently purchased and possessed. The academic information data items constructed and serviced through this study were summarized into seven papers, authors, abstracts, institutions, themes, journals, and references, and defined as core contents under construction. The integrated DB schema was created through this study, and the results of this study will be used as a basis for constructing the integrated DB collection of high quality academic papers and designing the integrated system.

■ keyword : | Science Technology Information | Research Data | Academic Paper | Literature Information | Information Service |

* 본 연구는 한국과학기술정보연구원 연구과제로 수행되었습니다.
접수일자 : 2019년 11월 20일
수정일자 : 2019년 12월 10일
심사완료일 : 2019년 12월 10일
교신저자 : 임석종, e-mail : seoklim@kisti.re.kr

I. 서 론

국가과학기술정보센터(NDSL)[1]는 기관과 연구자들에게 고품질의 논문 정보를 제공하여 연구자가 해당분야와 관련된 논문을 검색하고 이용할 수 있도록 만들어져 서비스되고 있다. 학술정보통합관리시스템(KSCD[2]는 국내 과학기술 분야의 학술 콘텐츠를 구축 및 서비스하고 있다. 그리고 전자정보국가컨소시엄(KESLI)[1]은 국내 전자저널 구독 기관들이 각자 보유한 정보를 공동 활용할 수 있는 체계를 구축하고 운영하고 있다. 이 외에도 학술연구정보서비스(RISS)[3], DBPia[4], 네이버 학술정보 서비스[5] 등으로 학술정보가 서비스되고 있다. 이러한 다양한 시도를 통해서 국내의 학술 연구 성과 데이터는 지속적으로 축적되고 관리되어 검색 서비스 형태로 제공되고 있다. 이러한 학술논문은 연구자가 생산한 연구성과이자 연구개발을 위해 소비되는 핵심적인 지식자원이다. 국가과학기술정보센터(NDSL)에서 제공되는 학술논문의 지식자원은 해외학술논문 DB, 국내학술논문 DB, 국가R&D논문 DB 등 다양한 지식자원으로 구성되어 있다. 해외 학술논문 DB는 KESLI 컨소시엄 계약을 통해서 수집된 학술논문 메타데이터를 구축하고 관리한다. 국내학술논문 DB는 (한국과학기술정보연구원)KISTI[1]와 국내학술단체가 학술정보공동활용사업 협약을 통하여 수집된 학술지 원문을 관리한다. 국가R&D 논문 DB는 정부지원 연구 개발과제를 수행하여 창출된 논문 성과를 수집하고 관리한다. 이밖에도 저자-기관 식별 DB, 한국인 저자 논문 DB 등 다양한 학술논문 DB를 관리한다. 이와 같이 다양한 채널을 통해서 수집되고 관리되고 있는 분산된 학술논문 DB를 통합적으로 운영하기 위한 DB 분석 작업이 필요하다. 학술정보를 통합 운영, 관리함으로써 소비자가 원하는 데이터를 쉽게 서비스가 가능하고[6][11][12] 공공뿐만 아니라 민간에서도 통합아키텍처를 활용 가능하다[8-10]. 이렇게 통합적으로 데이터베이스 스키마를 재설계하려는 시도는 과거에도 있었다[13-15]. [13] 논문에서는 데이터베이스 통합 방법론으로써 다음의 단계를 명시하였다. 요구사항 분석(통합 데이터베이스 간의 호환성 등), 데이터베이스 모델링, 구현 설계 논리 스키마 구현, 물리 스키마 설계 및 최적화, 작업 프로세스 통합[13]. 이러한 단계를 통하여서 데이터베이스를 통합할 수 있다는 설명이 기술되었으며 본 논문에서는 요구사항 분석과 데이터베이스 모델링 부분은 DB컨설팅을 통하여 나온 산출물이 있으며 저자와 기관 등 식별 데이터를 중심으로 한 데이터베이스 모델이 산출되었다. 실제 데이터베이스 분석 작업은 하지 못한 부분이 있기 때문에 전체 단계에서 논리 스키마를 도출하는 부분을 진행하였다[14]. 논문에서는 데이터베이스 스키마를 설계하기 위해서 정규화 방식과 정보 분석 방법을 사용했는데 해당 논문에서는 정규화 방법이 더 낳은 결과를 가져온 것을 확인되었다[15]. 논문에서는 새로운 데이터 어플리케이션을 위한 스키마 통합 방법론에서 탑 다운 방식과 바텀 업 방식을 비교하여 성능을 평가하였으며 두 가지 방법이 상호 보완적으로 작용할 수 있음을 입증하였다. 기존 방법과 본 논문의 차이점이라고 한다면 정규화 방식을 도입하는 부분은 충분한 검토와 의견수렴이 되어야 하기에 정보 분석 방법을 활용하였으며, 탑 다운 방식보다는 현재 데이터베이스를 우선 확인하고 개선 가능한 부분을 확인하기 위해 바텀 업 방식을 활용하였다. 잘 알려진 Web of Science나 SCOPUS 같은 데이터베이스의 콘텐츠는 지속적으로 분석되고 연구되어지고 있다[16-18]. 본 논문에서는 국내외 학술논문 DB를 분석하고 통합 DB 스키마 생성을 위해 다음과 같이 수행하였다. 첫째, 챕터 II. 챕터 III에서는 개별 학술논문DB 스키마를 분석하여 DB 별 주요 핵심 콘텐츠에 대한 메타데이터를 확인하고 Web of Science와 SCOPUS 데이터베이스의 메타데이터와의 비교를 수행한다. 둘째, 챕터 IV. 에서는 DB 사용현황을 파악함으로써 효율적인 DB 최적화 설계를 위한 미사용 테이블 근거 데이터를 확보하고자 한다. 본 연구를 통하여 산출된 통합 DB 스키마를 기반으로 고품질 학술정보 컬렉션 구성과 시스템 최적화의 기반 자료로 활용할 예정이다.

II. 학술논문 DB 스키마 분석 방법

콘텐츠의 개념 : 논문 학술정보에서의 데이터의 개념은 논문에 기재된 저자명, 타이틀 등 각각의 객체가 데

이터화 되어 데이터베이스에 저장될 수 있는 형태로 가공된 것이다. 콘텐츠의 개념은 여기에서 더 나아가 의미를 부여할 수 있는 데이터의 모임이라고 볼 수 있다. 예를 들면 저자 콘텐츠라고 하였을 때 저자와 관련된 저자명뿐만 아니라 공저자, 교신저자, 저자 ID, 주소, 연락처, 이메일 등을 전부 포괄하는 것이다.

학술논문 DB 분석 스키마는 다음 5가지를 대상으로 수행되었다.

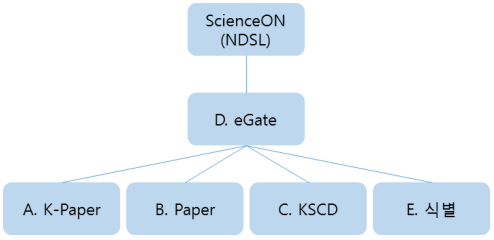

그림 1. 학술논문 국내DB 연계도

[그림 1]은 학술논문 DB 연계도를 나타낸 것이다. ScienceON에서 서비스되고 있는 학술정보는 eGate 데이터베이스를 통해 공급되며 eGate는 다시 K-Paper(국내 연구자 검증 데이터베이스), Paper(국가R&D논문 데이터베이스), KSCD(국내학술논문 데이터베이스), 식별 데이터베이스(저자, 기관 식별 데이터베이스)로 각각 운영되며 관리된다.

통합DB스키마를 위해서 분석된 데이터베이스는 해외학술논문(eGate)[1], 국내학술논문(OCEAN)[2], 국가R&D 논문(Paper)[13], 한국인 저자 논문(K-Paper)[14], 저자, 기관, 용어 식별 데이터베이스[15] 이다. K-Paper 데이터베이스는 국내 연구자가 발표한 성과에 대한 검증, 다양한 분석 서비스를 제공하는 것을 목표로 한다. Paper 데이터베이스는 국가연구개발 연구 성과로 등록된 논문을 적재하고 있다. KSCD 데이터베이스는 국내 학회, 협회에서 수집된 논문 데이터베이스이다. 식별 데이터베이스는 저자 식별 데이터, 기관 식별 데이터베이스를 축적하고 관리하고 있다. eGate는 KESLI[3] 컨소시엄을 통해서 수집된 국내외 학술논문 정보와 해외 출판사의 메타데이터 구매 분 등을 수집 관리하고 있다.

표 1. 학술논문 DB 스키마 구분

구분	분석 DB	내용
A	K-Paper	한국인 저자 DB
B	Paper	국가R&D논문 DB
C	KSCD	국내학술논문 DB
D	eGate	해외학술논문 DB
E	식별	인물, 기관, 용어 DB
F	통합	A+B+C+D+E
W	WoS	Web of Science
S	Scopus	Scopus
WS	W + S	W+S

학술논문 통합DB 구축을 위해서 필요한 주요 논문정보를 스키마 분석을 통해서 확인하는 작업이 필요하다. 학술논문 데이터베이스에는 기본적으로 시스템 관련 테이블과 콘텐츠 관련 테이블로 나눌 수 있다. 본 연구에서는 학술논문 통합DB 구축을 위해 콘텐츠 관련 테이블을 분석하였으며, 학술논문 국내DB 스키마를 Web of Science, Scopus 스키마와 비교 분석하였다. [표 1]은 학술정보 DB를 A, B, C, D, E, F, W, S로 대상 데이터베이스를 구분하였다. F는 A+B+C+D+E DB에서 중복된 스키마를 취합하였고, WS는 W와 S의 중복된 데이터베이스의 스키마를 통합한 것을 나타낸다. 중복된 스키마를 취합한 이유는 각 데이터베이스에서 핵심적인 논문정보로 간주하였기 때문이다. 중복된 스키마에 대한 설명은 다음과 같다. Web of Science, SCOPUS 등 학술정보를 관리하는 데이터베이스에서 필요하다고 생각되는 항목을 구축하고 서비스하고 있다. '중복되었다' 라는 의미는 Web of Science에서도 구축하고 있고 SCOPUS에서도 구축하고 있는 메타데이터 항목을 가리킨다. 각 데이터베이스는 막대한 예산을 투입하여 학술정보를 구축하고 서비스하고 있기에 필요하지 않은 메타데이터 항목은 비용을 들이지 않을 뿐더러 구축되지 않을 것이다.

표 2. WoS, Scopus 분석 스키마

구분	Web of Science	Scopus
기간	1980~2015	2017

[표 2]는 Web of Science DB와 Scopus DB의 분석된 스키마를 나타낸 것이다.

III. 학술논문 DB 스키마 분석

학술논문 통합 DB 구축을 위해 국내 DB 분석 연구 수행 절차는 다음과 같다. 학술논문 DB 스키마를 분석하기 위한 분석 대상 DB를 [표 1]과 같이 먼저 정의한다. [표 3]와 같이 학술논문 DB 별 사용, 미사용 테이블을 정의하고 DB 별 필드 명세서를 취합한다. KSCD와 eGate의 경우에는 사용하고 있지 않은 테이블이 다수 존재하고 있었으며 사용하는 테이블은 일부로 확인이 되었다. 학술논문 통합 DB 스키마를 적용하기 위해 사용 중인 테이블 491개에 대한 내용을 분석하였다.

표 3. 학술논문 DB 사용, 미사용 테이블 구분

구분	분석 DB	사용	미사용	합계
A	K-Paper	10	0	10
B	Paper	67	0	67
C	KSCD	49	123	172
D	eGate	362	671	1033
E	식별	3	0	3
	전체	491	794	1285

학술논문 DB 간 비교를 위해 구분 콘텐츠를 정의한다. 구분의 기준은 모든 테이블에 대해서 사용 용도를 정의하고 W, S, F 사이에 서로 비교가 가능한 콘텐츠가 어느 부분인지를 정의한다.

표 4. 비교가 가능한 비교 콘텐츠

W	S	F
논문	논문	논문
저자	저자	저자
초록	초록	초록
기관	기관	기관
저널	저널	저널
주제	주제(스키마 없음)	주제
참고문헌	참고문헌	참고문헌

[표 4]는 W, S, F에서 비교가 가능한 테이블 콘텐츠가 어느 부분인지를 확인해 주고 있다. 비교가 가능한 콘텐츠는 7가지(논문, 저자, 초록, 기관, 주제, 저널, 참고문헌)으로 요약될 수 있음을 확인할 수 있었다.

위의 결과에서 확인된 7가지 핵심 콘텐츠를 기준으로 개별 학술논문 DB 스키마 분석을 다음과 같이 수행하였다. 학술논문 DB의 스키마를 분석하는 이유는 중복 데이터가 존재하고 중복 데이터에 대한 필드 항목이 일치하는지 여부를 확인하는 것이다.

표 5. 저자 콘텐츠 필드 내용

No.	저자	DB
1	논문 제어번호	논문 제어번호
2	저자명	B, C, D, E
3	저자ID	A, B, D, E
4	저자 이메일	B, D, E
5	저자 순서	A, B, E
6	소속기관명	B, C, D
7	등록일자	A, B

[표 5]는 저자 콘텐츠에 대해서 각 DB 별 공통으로 구축되는 필드를 나타낸다. 각 데이터베이스에서 동일하게 가지고 있는 필드이며 DB간에 중복 데이터 값으로 볼 수 있다. 저자정보가 대체적으로 많은 데이터베이스는 B. Paper와 E. 식별 데이터베이스이다. B와 E를 중심으로 저자명과 소속기관 명, 이메일, 저자 순서, 저자ID, 등록일 정보를 제외한 중복되지 않는 항목(기준년도, 직급, ID, 저자역할 등)을 통합적으로 구축한다면 효과적인 저자 콘텐츠 관리가 될 것이다.

표 6. 기관 콘텐츠 필드 내용

No.	기관	DB
1	기관 제어번호	기관 제어번호
2	기관주소	A, B, C, E
3	기관명	B, C, D, E
4	등록자	A, C
5	등록일	A, C

[표 6]은 각 DB 별 중복된 기관 콘텐츠를 나타낸다. 각 데이터베이스에 소속기관명이 기본적으로 존재하며 주로 식별 데이터베이스와 OCEAN 데이터베이스에 많이 존재한다. DB 별로 기관주소, 기관명, 등록자, 등록일이 중복으로 구축되고 있음을 확인하였다. 전화번호, 홈페이지 주소, 기관 분류, 설립 일자, 등록 일자 등의 정보는 중복되지 않고 개별 DB에서 관리되고 있었다.

표 7. 논문 콘텐츠 필드 내용

No.	논문	DB
1	논문 제어번호	논문 제어번호
2	권	B, C, D
3	호	B, C, D
4	시작페이지	B, C, D
5	끝페이지	B, C, D
6	논문명	B, C, D

7	출판언어코드	B, C, D
8	DOI	B, C
9	연구과제번호	B, C
10	등록자ID	B, C
11	등록일자	B, C
12	수정자ID	B, C
13	수정일자	B, C
14	키워드	B, C
15	초록	B, C
16	OpenAccess여부	B, C

표 8. 초록 콘텐츠 필드 내용

No.	초록	DB
1	논문 제어번호	논문 제어번호
2	초록	B, D

[표 7]는 논문 콘텐츠를 나타낸다. OA여부, DOI 등의 논문에 대한 기본 메타 정보가 있으며 등록일자와 수정일자 등은 중복으로 관리되고 있는 항목이다. 특이한 점은 초록 정보가 있는데 보유하고 있는 데이터베이스가 [표 8]의 초록 데이터베이스와 다르기 때문에 데이터의 충실도를 확인하여 비교하는 것이 필요하다. 초록 정보는 D. eGate 데이터베이스와 B. Paper 데이터베이스에 주로 축적되어져 있다.

표 9. 참고문헌 콘텐츠 필드 내용

No.	참고문헌	DB
1	논문 제어번호	논문 제어번호
2	참고문헌 식별자	A, C
3	참고문헌 TYPE	A, B, C
4	저자명	A, B, C
5	권 정보	A, B, C
6	호 정보	A, B, C
7	시작 페이지	A, B
8	년도	A, B, C
9	ISSN	A, B, C
10	DOI	A, B, C
11	기사명	A, B, C
12	참고문헌 원형 정보	B, C
13	매핑 논문제어번호	B, C
14	등록 일자	A, B, C
15	참고문헌URL	A, B, C

[표 9]는 참고문헌 콘텐츠를 나타낸다. 참고문헌에 대한 논문 매핑 제어번호를 가지고 있기 때문에 인용정보 서비스에 활용될 만한 필드 값을 확인할 수 있다. 참고문헌은 주로 eGate 데이터베이스와 Paper 데이터베이스에 많이 존재한다.

표 10. 주제 콘텐츠 필드 내용

No.	주제	DB
1	관리 제어번호	관리 제어번호
2	DDC 명 한글	D
3	DDC 명 영문	D
4	과학기술표준분류	C

[표 10]은 주제 콘텐츠를 나타낸다. 현재는 DDC를 기준으로 주제가 분류되기 때문에 DDC 버전에 대한 업그레이드와 주제 분류 알고리즘 개선이 요구되고 있다. 주제 정보는 주로 eGate 데이터베이스에 축적되고 있다.

표 11. 저널 콘텐츠 필드 내용

No.	저널	DB
1	KOJIC	관리 제어번호
2	자료 유형	D
3	학술지 명 한글	D
4	학술지 명 영문	D
5	발행 국가 코드	D
6	사용 언어	D
7	SCI 여부	D
8	창간 일자	D
9	간기	D
10	전자 저널 대상 여부	D
11	권호 유무	D
12	폐간 일자	D
13	등록 일자	D
14	P-ISSN	D
15	E-ISSN	D
16	DDC 분류	D
17	KSCI 여부	D
18	논문 투고 사이트 URL	D
19	OA 라이센스	D
20	발행 여부	D

[표 11]은 저널 콘텐츠 정보를 나타낸다. 저널에 대한 정보는 eGate 데이터베이스와 OCEAN 데이터베이스에 축적되어져 있다. 저널 정보에 대해서 다른 DB와 매핑을 수행한다면 콘텐츠의 품질을 제고할 수 있다.

7가지 콘텐츠 별 필드 수를 비교하여 대략적으로 콘텐츠 별로 정보의 양을 비교할 수 있다. 이를 위해서 F 데이터베이스에서 각 DB에서 개별적으로 소유하고 있는 필드는 제외하고 전체 DB에서 공통적으로 구축하고 있는 필드를 선별하여 WS와 비교하였다.

표 12. F와 WS DB의 콘텐츠 별 필드 수 비교(F, WS: 중복부분 취합)

DB	논문	저자	초록	참고문헌	기관	주제	저널
F	16	7	2	15	4	3	18
WS	11	7	2	9	1	6	3

[표 12]는 F와 WS DB의 콘텐츠 별 필드 수를 비교한 것이다. 필드 요소가 아닌 수치로만 되어 있지만 수치로써 우선 데이터양을 가늠할 수 있다. [표 12]에서의 F와 WS는 중복부분 만을 취합한 수치이다. 대체적으로 비슷한 개수를 유지하고 있다. 여기에 나와 있는 데이터 필드 수는 콘텐츠 별 기본이 되는 데이터로 볼 수 있다.

표 13. F와 WS의 논문 콘텐츠 필드 비교

논문	F	WS
1	논문관리번호	논문관리번호
2	시작페이지	시작페이지
3	끝페이지	끝페이지
4	논문명	논문명
5	DOI	DOI
6	연구과제번호	-
7	등록자ID	-
8	등록일자	-
9	수정자ID	-
10	수정일자	-
11	키워드	-
12	초록	-
13	OpenAccess여부	-
14	권	-
15	호	-
16	출판언어코드	-
17	-	인용 문헌 개수
18	-	문헌 타입
19	-	출판사 명
20	-	발행 연도
21	-	출판 연도
22	-	문헌 상태 타입

표 14. F와 WS의 저자 콘텐츠 필드 비교

저자	F	WS
1	논문관리번호	논문관리번호
2	저자명	저자명
3	소속기관명	소속기관명
4	저자 순서	저자 순서
5	저자 이메일	-
6	저자ID	-
7	등록일자	-
8	-	도시
9	-	국가
11	-	공저자

표 15. F와 WS의 초록 콘텐츠 필드 비교

초록	F	WS
1	문서 제어번호	문서 제어번호
2	초록	초록

표 16. F와 WS의 참고문헌 콘텐츠 필드 비교

참고문헌	F	WS
1	문서 제어번호	문서 제어번호
2	참고문헌 ID	참고문헌 ID
3	참고문헌 발행 연도	참고문헌 발행 연도
4	참고문헌 제목	참고문헌 제목
5	참고문헌 출처 권	참고문헌 출처 권
6	참고문헌 출처 호	참고문헌 출처 호
7	참고문헌 시작 페이지	참고문헌 시작 페이지
8	참고문헌 저자	참고문헌 저자
9	참고문헌 타입	-
10	DOI	-
11	ISSN	-
12	참고문헌 원형 정보	-
13	매핑 논문제어번호	-
14	등록 일자	-
15	참고문헌URL	-
16	-	참고문헌 끝 페이지

표 17. F와 WS의 기관 콘텐츠 필드 비교

기관	F	WS
1	기관명	기관명
2	기관주소	-
3	등록자	-
4	등록일	-

표 18. F와 WS의 주제 콘텐츠 필드 비교

주제	F	WS
1	문서 제어번호	문서 제어번호
2	DDC 명 영문	DDC 명 영문
3	DDC 명 한글	-
4	-	SCI 여부
5	-	코드
6	-	분류
7	-	백분율

표 19. F와 WS의 저널 콘텐츠 필드 비교

저널	F	WS
1	저널명 영문	저널명 영문
2	저널명 한글	P-ISSN
3	KOJIC	E-ISSN
4	발행 국가 코드	-
5	사용 언어	-
6	SCI 여부	-
7	창간 일자	-
8	간기	-
9	전자 저널 대상 여부	-
10	권호 유무	-
11	폐간 일자	-
12	등록 일자	-
13	자료 유형	-
14	DDC 분류	-
15	KSCI 여부	-
16	논문 투고 사이트 URL	-
17	OA 라이센스	-
18	발행 여부	-

[표 13-표 19]은 논문, 저자, 초록, 참고문헌, 기관, 주제, 저널 콘텐츠 별 F와 WS의 비교를 나타낸 것이다. 회색으로 표기된 부분은 F와 WS의 공통 정보를 나타낸 것이다. 논문의 경우 연관 과제와 OA, 키워드와 초록 정보는 DB 간 연계를 통해서 데이터 충실도를 확보하여 학술정보 서비스에 중요한 데이터로 활용될 수 있다. 저자 콘텐츠의 경우에는 저자 이메일을 통해서 저자 식별을 높일 수 있으며, 국가정보와 공저자 정보는 WS에 비해 아쉬운 부분으로 남는다. 이러한 정보를 추가로 확보한다면 진보된 큐레이션 서비스가 가능할 것이다. 초록의 경우에는 항목이 같았으며, 참고문헌의 경우에는 F와 WS를 비교했을 때에 F는 DOI 정보와 ISSN, 참고문헌 URL 등의 정보가 추가적으로 더 구축하고 있었다. 이러한 정보는 연계정보로 활용이 가능하다. 기관정보의 경우에는 주소까지 구축하고 있으며, 주제 분류는 DDC 분류로 동일한 분류를 사용하고 있음을 확인할 수 있다. 저널정보는 언어, 국가코드 등의 추가적인 정보를 더 많이 확보하고 있다. 저자 콘텐츠를 제외한 나머지 콘텐츠에서 추가적인 정보를 확보하고 있어 이러한 정보를 통합하여 콘텐츠를 개발한다면 유용한 학술정보 서비스가 될 것이다.

통합 DB 스키마 도출 : 위와 같은 비교를 토대로 다음과 같은 핵심콘텐츠 별 통합 DB 스키마를 생성하였다. 현재까지의 분석 부분으로 향후 일부 수정이 될 수 있다.

표 20. 논문

No	컬럼명
1	논문관리번호
2	저널구분코드
3	해외학술지구분코드
4	저널명
5	ISSN번호
6	출판년월
7	권
8	호
9	시작페이지
10	끝페이지
11	논문명
12	논문명(타언어)
13	출판언어코드
14	ImpactFactor
15	DOI
16	연구분야코드
17	NDSL_CNTNTS_CTRL_NO
18	E-Gate ID
19	WOS UT
20	SCOPUS EID
21	KCI ID
22	URL
23	파일공개여부
24	키워드
25	초록
26	사사표기여부
27	사사표기
28	OpenAccess여부
29	6T기술분류코드
30	SCI_여부
31	AHCI_여부
32	SSCI_여부
33	SCIE_여부
34	ESCI여부

표 21. 저자

No	컬럼명
1	논문관리번호
2	저자순번
3	저자명(ShortName)
4	저자명(FullName)
5	과학기술인번호
6	저자역할구분코드
7	저자식별코드
8	소속기관코드
9	소속기관명
10	제1저자여부(Y,N)
11	교신저자여부(Y,N)

표 22. 초록

No	컬럼명
1	구분
2	논문관리번호
3	언어
4	초록언어
5	초록데이터

표 23. 참고문헌

No	컬럼명
1	참고문헌관리번호
2	논문관리번호
3	DOC_TYPE
4	제목
5	DOI
6	ISSN
7	통권
8	권
9	호
10	발행년도
11	시작페이지
12	종료페이지
13	웹문서명
14	인용일
15	저자
16	참고문헌 소스 관리번호
17	참고문헌 소스 식별
18	참고문헌 유형
19	URL
20	출판사(발행기관)
21	피인용횟수
22	피인용횟수 수정일자
23	대학명
24	학과명
25	학술대회명
26	학술지명
27	학술지약어명
28	학위구분
29	화제의논문

표 24. 기관

No	컬럼명
1	기관 ID
2	논문관리번호
3	기관 구분
4	기관 성격
5	기관 분류
6	기관 명 영문
7	기관 명 한글
8	기관 명
9	기관 약칭 영문
10	기관 홈페이지 URL
11	연구재단 대분류 명 한글
12	우편 번호
13	주소
14	국가명
15	국가주소
16	기관주소
17	기타주소
18	년도 정보

표 25. 주제

No	컬럼명
1	고유번호
2	상위코드
3	수준
4	DDC
5	영문 명칭
6	KISTEP 코드 1
7	KISTEP HanName
8	KISTEP EngName
9	KISTI 코드 1
10	KISTI HanName
11	KISTI EngName
12	SCI 주제분류코드
13	SCOPUS 주제분류코드

표 26. 저널

No	컬럼명
1	DDC 코드
2	ISSN
3	E-ISSN
4	학술지 명 영문
5	학술지 명 한글
6	서지제어번호
7	창간 일자
8	창간년월
9	NDSL 폐간년
10	출판국
11	출판사
12	영문출판사
13	OA 라이센스
14	OECD WOS 코드
15	간기
16	구분(j-저널,p-프로시딩)
17	자료형태(p-인쇄, e-전자, a-인쇄+전자)
18	저널구분(1-인쇄 소장, 2- 전자 라이선스)
19	학술지 URL
20	발행 국가 코드
21	발행국명
22	사용 언어

IV. 학술논문 DB 사용현황 분석

학술논문 DB 사용현황 분석을 위해 1달간 각 DB에서 들어오는 세션을 모니터링 할 수 있는 코드를 수행하였다. [표 27]은 각 DB 별 세션 모니터링을 수행한 결과를 빨간색으로 기재한 표이다. 미사용 테이블이라고 생각되었던 부분에서 사용하고 있는 테이블로 확인된 내용이 발견되었으며, 전체 DB 크기로 보았을 때에 Paper 는 57%, OCEAN은 36%, eGate는 20% 정도를 사용하고 있었다. 미사용 테이블이라고 생각되는 테이블 판단 근거는 DB구축을 수행하는 수행자 입장에서 정리한 내용과 DB컨설팅 과정을 통해서 나온 미사용 테이블 리스트를 근거로 판단한 내용이며, 최종 실질

검증은 DB 세션 모니터링을 분석 과정을 통해서 확인되었다. 세션 모니터링을 통해서 확인된 미사용 테이블을 정리하게 되면 시스템 운영 관리 측면에서 스토리지 용량을 줄이고 추가 서버 구매 비용을 줄일 수 있으며, 필요한 콘텐츠만을 정제하여 구축할 수 있어 일석이조 이상의 효과를 얻을 수 있다.

표 27. DB 별 세션 모니터링 결과 비교

학술논문 DB	사용 테이블	미사용 테이블	합계
A. K-Paper	10	0	10
B. Paper	67(38)	0	67(38, 57%)
C, OCEAN	49(31)	123(31)	172(62, 36%)
D. eGate	362(146)	671(63)	1033(209, 20%)
E. 식별	3	0	3
F. 전체	491	794	1285

데이터베이스에 들어오는 세션 정보는 해당 DB에 접근 권한이 존재한다는 의미이며, SELECT 권한이 있는 접근 계정에 대한 세션 모니터링을 수행하여 실제 사용자 아이디 별 조회 테이블을 파악하였다. eGate에서는 3개의 아이디가 주요하게 사용되고 있었고, RPMS에서는 2개, 그리고 OCEAN에서는 3개의 아이디가 주로 사용됨을 확인하였다. eGate, RPMS, OCEAN 데이터베이스를 사용하고 있는 사용자 또는 응용프로그램 등이 많을 것이라 짐작되었지만 실제 DB세션 모니터링을 3달간 수행한 결과 해당 DB에 접근한 사용자는 eGate에서는 3개의 아이디, RPMS에서는 2개의 아이디, OCEAN에서는 3개의 아이디로 적은 사용자가 사용하고 있었다. 추후에 데이터베이스 시스템 통합을 고려한다면 사용자가 많을 경우 수정해야 하는 사용자 또는 응용프로그램이 많은데 사용자가 많지 않기 때문에 해당 아이디를 사용하는 사용자나 응용프로그램만 변경하면 손쉽게 통합 작업을 수월하게 할 수 있음을 확인하였다.

V. 결론

본 논문에서는 국내외 학술논문 통합 DB 구축을 위하여 개별 DB 스키마를 분석하고 Web of Science, SCOPUS 데이터베이스와 비교하여 핵심콘텐츠 별 주요 필드를 정리하여 통합 DB 스키마를 도출하였다. 또한, 향후 학술정보 시스템 최적화 설계를 위한 DB 사용현황을 분석하였다. 개별 DB 스키마 분석을 통하여 어느 DB의 콘텐츠가 충실하게 구축이 되었는지 와 어떤 항목을 구축하고 있는지를 확인할 수 있었으며, DB 사용현황을 통하여 사용하지 않는 자원을 효과적으로 운용할 수 있는 내용도 확인하였다. F와 WS 비교를 통하여 저자 콘텐츠를 제외한 부분에서 WS보다 많은 강점을 확인할 수 있었다. 본 논문에서 도출된 통합 DB 스키마를 기반으로 데이터베이스 간 API 연계로 핵심콘텐츠를 구축할 수 있는 부분이 단기간 효율을 높일 수 있는 방법이다. 그리고 시스템 최적화를 위해 미사용 테이블을 정리하고 효율적으로 관리하기 된다면 운영, 관리에 소요되는 비용을 줄일 수 있는 방법이 될 것이다. 학술논문 통합 DB 스키마를 기반으로 각 데이터베이스를 연계시켜 일정 주기별 데이터 구축 작업은 앞으로 수행할 과제이며, 본 논문의 한계점은 통합 DB 스키마가 도출되었지만 실제 데이터베이스 연계는 아직 되어있지 않은 부분이다. 학술논문 DB스키마 분석과정을 통해 산출된 핵심 콘텐츠 별 요소를 잘 정제하여 통합하는 큐레이션 작업을 통하여 학술논문 DB 구축에 적용한다면 손쉬운 분석 통계, 우수연구자 분석, 연구자 논문성과 검증 지원, 약탈 학술지 리스트 관리 및 우수 학술지 선정과 추천, 기계학습 데이터 변환 및 지원, 토픽을 중심으로 한 연구성과 연계 등의 서비스가 가능하다.

참 고 문 헌

[1] 강남규, 신용주, 박근철, 주원균 외, “NTIS-NDSL 활용한 국가R&D논문 수집 방안에 관한 연구” 한국인터넷정보학회 학술발표대회 논문집, pp.231-232, 2010.

[2] 김병규, 강무영, 최선희, “협회 기술정보관리 및 유통 시스템 구축에 관한 연구,” 정보관리연구, Vol.37, No.3, pp.117-137, 2006.

[3] 이성숙, "학술연구정보서비스(RISS)의 발전방안 연구," 한국도서관정보학회지, Vol.37, No.3, pp.103-129, 2006.

[4] 이태영, 차웅석, 김남일, "한국 저널에 수록된 보완대체의학 관련 연구 동향 분석-DBPia에 수록된 논문을 중심으로," 한국의사학회지, Vol.22, No.1, pp.69-80, 2009.

[5] 박상근, "인문학 분야의 인용 데이터정보원 비교 분석: 네이버 전문정보, KCI," 정보관리학회지, Vol.30, N.1-87, pp.33-50, 2013.

[6] H. S. Lee and J. H. Kwon, "The Survey about the Personalized Data Curation in the Age of Big Data," Korean Institute of Information Technology, Vol.14, No.1, pp.124-127, 2013.

[7] Y. J. Im, S. K. Baek, and S. J. Yeon, "Select and focus on securing competitiveness in the Big Data era," The Journal of The Korean Institute of Communication Sciences, Vol.29, No.11, pp.3-10, 2012.

[8] H. J. Seo and S. H. Myeong, "Policy Alternatives for User-oriented Public Data Utilization - Focusing on ICT Managers' Perception in Private Sector," Journal of Korean Associastion for Regional Information Society, Vol.17, No.3, pp.61-86, 2014.

[9] J. H. Kim, "A Study on the Perceptions of University Researchers on Data Management and Sharing," JOURNAL OF THE KOREAN SOCIETY FOR LIBRARY AND INFORMATION SCIENCE, Vol.49, No.3, pp.413-436, 2015.

[10] S. Y. Bang, H. D. Ha, and C. J. Kim, "A Study on BigData-based Software Architecture Design for Utilizing Public Open Data," Journal of Korean Institute of Information Technology, Vol.13, No.10, pp.99-107, 2015.

[11] J. H. Kim, S. E. Hong, Y. J. Kim, and H. J. Kim, "Proposal of Concept and Implementation of Datacon for Data Sharing and Utilizing," Journal of Korean Institute of Information Technology, Vol.14, No.1, pp.123-132, 2016.

[12] S. H. Shin, Y. J. Yoon, M. S. Yang, J. M. Kim, and K. R. Shon, "A Data Cleansing Strategy for Improving Data Quality of National R&D Information - Case Study of NTIS," Journal of the Korea Society of Computer and Information, Vol.16, No.6, pp.119-130, 2011.

[13] C. Batini, M. Lenzerini, and S. B. Navathe, "A comparative analysis of methodologies for database schema integration," ACM Comput. Surv., Vol.18, No.4, pp.323-364, 1986.

[14] Shoval, Peretz and Even-Chaime, Moshe, "Database schema design: An experimental comparison between normalization and information analysis," ACM SIGMIS Database, Vol.18 pp.30-39, 1987.

[15] Fong, Joseph, Karlapalem, Kamalakar, Li, Qing, and Kwan, Irene, "Methodology of Schema Integration for New Database Applications: A Practitioner's Approach," J. Database Manag, Vol.10, pp.3-18, 1999.

[16] Mongeon, Philippe and Paul-Hus, Adèle, "The Journal Coverage of Web of Science and Scopus: a Comparative Analysis," Scientometrics, Vol.106, 2015.

[17] E Falagas, Matthew, I Pitsouni, Eleni, alietzis, George, and Pappas, Georgios, "Comparison of PubMed, Scopus, Web of Science, and Google Scholar: Strengths and weaknesses," FASEB journal : official publication of the Federation of American Societies for Experimental Biology, Vol.22, pp.338-342, 2008.

[18] Salisbury, Lutishoor, "Web of Science and Scopus: A Comparative Review of Content and Searching Capabilities," The Charleston Advisor, 11, 2009.

저 자 소 개

최 원 준(Wonjun Choi) 정회원

- 1999년 3월 ~ 2006년 2월 : 원광대학교 수리통계학 학사
- 2013년 3월 ~ 2017년 2월 : 한국과학기술연합대학교 과학기술정보학 박사
- 2017년 2월 ~ 현재 : 한국과학기술정보연구원 연구원

〈관심분야〉 : 과학기술정보, 네트워크 분석, 데이터 분석

황 혜 경(Hyekyong Hwang) 정회원

- 1992년 2월 : 서울여자대학교 도서관학과(학사)
- 1999년 2월 : 연세대학교 문헌정보학과(석사)
- 2014년 2월 : 연세대학교 문헌정보학과(박사)
- 1999년 ~ 현재 : 한국과학기술정보연구원 책임연구원

〈관심분야〉 : 콘텐츠 큐레이션, 오픈액세스

김 정 환(Jeonghwan Kim) 정회원

- 2013년 : 충남대학교 문헌정보학(박사)
- 2014년 : 유럽핵입자물리연구소(CERN) 방문연구원
- 2016년 ~ 현재 : 영국물리학회 출판부(IOPP) 자문위원
- 2006년 ~ 현재 : 한국과학기술정보연구원 책임연구원

〈관심분야〉 : 데이터베이스

이강산다정(Kangsandajeong Lee) 정회원

- 2015년 : 중앙대학교 문헌정보학(박사)
- 2016년 ~ 2017년 : University of Washington visiting librarian
- 2018년 ~ 현재 : 한국과학기술정보연구원 박사후연구원

〈관심분야〉 : 데이터베이스, 메타데이터

임 석 종(Seokjong Lim) 정회원

- 1998년 2월 : 중앙대학교 문헌정보학 석사
- 2009년 8월 : 중앙대학교 문헌정보학 박사
- 2005년 ~ 현재 : 한국과학기술정보연구원 선임연구원

〈관심분야〉 : 오픈액세스, 학술커뮤니케이션, 학술정보, 메타데이터, 계량분석