# Apache-Solr 를 이용한 KSCD 학술정보 통합관리시스템 고도화

모은수*, 김병규*, 김정환*
*한국과학기술정보연구원
e-mail : {esmo, bk.kim, kimjh}@kisti.re.kr

# Enhancement of KSCD Information Integrated Management System using Apache-Solr

Eun-Su Mo*, Byung-Kyu Kim*, Jeong Hwan Kim *
*Dept. of Computer Science, Han-Kook University

**요 약**

학술정보 통합관리시스템 OCEAN(Online Collaborator for sociEty & Association Network of KISTI)은 학회정보화지원사업의 일환으로 학술정보를 통합관리하기 위해 2007 년부터 사용된 통합관리 시스템이다. 구 OCEAN 은 Struts2 프레임워크 기반으로 설계되어. 노후화된 프레임워크는 끊임없는 유지보수에도 급변하는 정보기술 환경에 낙후되었고, 보안적 측면 및 신규 기능 구현에 어려움이 있어 2 차년도(2015-2016)에 걸친 시스템 재정비를 통해 최신 프레임워크를 적용하여 신규 시스템 설계하고 구축하였다

## 1. 서론

KISTI(Korea Institute of Science and Technology Information)는 학술정보 DB(DataBase) 구축 및 서비스의 대표기관으로, 과학기술분야 전반의 학술정보를 유통(수집, 가공, 처리, 서비스)을 위해 학술정보화지원사업을 지원하고 있다. 과학기술분야 학회에서 발생되는 고급 학술 정보는 산업, 경제적 측면에서 그 효용 가치가 매우 크고 국가 경쟁력 강화를 위해 전략적 차원으로 첨단과학기술정보를 효율적인 서비스 체제를 구축하는 것이 매우 중요하다[1].

대표적인 정보시스템으로는 Meta 정보에 특화된 과학기술학회마을(이하, 학회마을), 논문 및 보고서에 그리고 최신 과학기술 이슈와 트렌드를 제공하는 NDSL(National Digital Science Library) 그리고 지난 2014 년도에 정부 3.0 의 공공 데이터 개방 정책에 맞춰 개방된 과학기술인용색인 KSCI(Korea Science Citation Index)의 참고문헌 데이터[2] 등이 이에 해당된다.

학술정보통합관리 시스템 OCEAN(Online Collaborator for sociEty & Association Network of KISTI)은 이러한 서비스 시스템의 기반데이터를 구축하기 위해 2007 년부터 사용된 학술정보 통합관리 시스템으로 KSCD(Korea Science Citation Database) 구축가이드[3]

에 맞춰 개발된 시스템이다. 구 OCEAN 은 Struts2 프레임워크 기반으로 설계되어, 노후화된 프레임워크 및 DB 는 끊임없는 유지보수에도 급변하는 정보기술 환경에 맞지않고 보안 이슈 및, 새로운 업무 프로세스를 반영하기 어려운 측면을 내포하고 있었고[4], 또한 학회마을과 KSCI 의 검색엔진이 이원화(Fast, KRISTAL)가 되어 색인 작업이 별도로 수행되어야 하는 문제가 있어 이를 오픈소스 기반의 Apache-Solr 로 통합하였다. Solr 는 Apache Tomcat 과 같은 Servlet Container 와 함께 사용할 수 있다는 장점을 가지고 있고, 단독 어플리케이션 서버 또는 Solr Cloud 라는 분산 서버로 사용이 가능하다는 장점을 가진다[5].

OCEAN 시스템을 개선하기 위해 2015 년도 시스템을 재 설계하였고, 재 구축된 시스템은 다음과 같은 특징을 가진다.

첫째, 기반 프레임워크 변경 및 검색엔진 통합
둘째, DB 스키마재설계 및 표준 도메인 적용
셋째, DOI 관리, DB 구축 프로세스개선 등의 중점 업무프로세스 개선
넷째, 통계 조회, 인용정보관리등의 기능 강화
다섯째, 시스템 운영관리 및 UI(User Interface) 개편

본 논문의 구성은 다음과 같다. 2 장에서는 관련연구를 소개하고 3 장에서는 재 구축된 시스템을 살펴보며 결론과 향후 연구를 제시하며 마무리 한다.

## 2. 관련연구

### 2.1. KSCD 통합관리시스템 설계 및 구현에 관한 연구[4]

- KSCD 통합관리시스템 설계 및 구현에 관한 연구에서는 기존 OCEAN 시스템의 문제점을 도출하여 5 가지의 목표를 수립하였다. 5 가지의 목표는 다음과 같다. 첫째, 국가표준 프레임워크인 전자정부 프레임워크 도입으로 시스템의 노후화 및 보안관련 이슈를 해

결. 둘째, DB 표준화를 통한 데이터 품질 향상. 셋째, 노후화된 DB Schema 재설계. 넷째, 국제적 표준을 준수하는 데이터 연계 기반 마련. 다섯째, 중점 업무 기능 고도화를 통한 업무능률 향상 등 시스템을 재 정비 하도록 하는 연구이다.

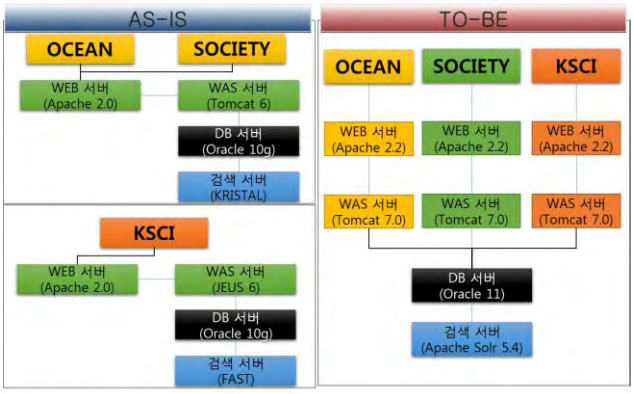
### 2.2. 아파치 스파크 기반 검색 엔진의 설계 및 구현[5]

아파치 스파크 기반 검색엔진 설계 및 구현 연구에서는 스파크를 기반으로 한 In Memory 처리를 apache Hadoop 기반의 분산 처리와 대비하여 성능차이를 비교하였고, 스파크 기반 검색엔진이 기존 방식보다 나은 성능을 보임을 확인하였다. Solr 검색엔진의 성능을 간접적으로 확인할 수 있는 연구이다.

## 3. 학술정보 통합관리 시스템 중점 개선사항

### 3.1. 프레임워크 및 시스템 구성 변경

Struts2 기반 프레임워크기반을 국가 표준 프레임워크인 전자정부 프레임워크로 적용하였다. 보안을 위해 Spring Security 와 시큐어 코딩이 적용되었다. [그림 1]과 같이 OCEAN 과 학회마을이 같은 웹 서버를 이용하도록 구현돼있어 향후 유지보수 및 관리를 위해 시스템이 분리되었다. 검색엔진이 이원화(Fast, Kristal)의 문제가 있어 이를 Apache-Solr 로 통합하여 관리 포인트를 줄일 수 있게 하였다.



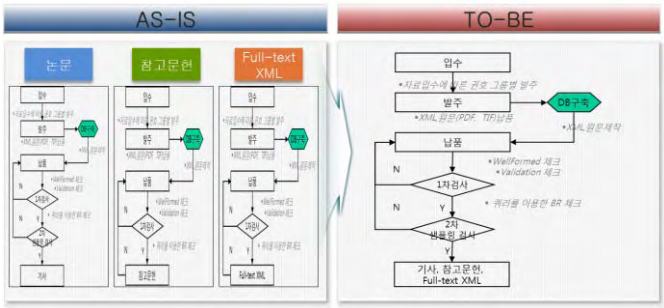

[그림 1] OCEAN System H/W Architecture

### 3.2. 데이터 베이스 재설계 및 데이터 마이그레이션

OCEAN DB 는 지속적인 업무프로세스 개선과 중점업무등의 필요성에 의해 테이블 및 컬럼이 추가되어왔다. 과거의 경우 Meta 데이터를 구축하는 것이 중점이였다면 현시점에서는 Meta 데이터뿐만이 아닌 그림, 테이블, 데이터, DOI 등 시대적인 요구사항이 많아졌기 때문이다. 이는 향후 업무프로세스 개선 및 유지보수 시 문제점을 내포하고 있었고, 필연적으로 개선할 필요가 있었다. 재 설계된 DB 는 메뉴와 기능에 맞춰 테이블을 통합하고, 연계 시스템인 과학정보학회마을, KSCI 의 데이터도 포괄적으로 관리 할 수 있게 재 설계되었다. 또한 데이터 유형별 표준화 도메인과 DiCMS(Digital Contents Management System) 표준용어를 적용되었다. 테이블 170 개, 약 1.3 억건이 해당되며, 데이터 마이그레이션은 1,2 차에 걸쳐 수행되었다. 마이그레이션 절차는 추출/이행/전환/보정/기존 자료정비 절차로 진행되었고, 데이터 이관을 위해 ETL(Extract Transform Loader)도구인 Talend 가 활용하였다. 1 차는 OCEAN 을 대상으로 진행되었으며, 2 차는 관련서비스시스템(학회마을, KSCI)을 테이블 10 개, 약 8,000 만 건을 대상으로 진행되었다.


[그림 2] OCEAN System Workflow

정보화 프로세스는 [그림 3]과 같이 데이터수집부터 서비스까지 KSCD 가공 지침에 따라 one-stop 으로 DB 구축을 수행할 수 있는 기능이다. 이 중 논문, 참고문헌, Full-text XML 의 구축 프로세스가 유사하나 개발 시점의 차이로 인한 콘텐츠 별 관리 메뉴 사용으로 사용 및 이용 통계 부분에 대한 활용이 어려움이 있었고, 각각 분리 관리되 업무담당자의 업무 효율에 저해되는 요소로 통합 관리할 수 있도록 프로세스가 통일 되었다.




[그림 3] 정보화 프로세스 개선사항

## 4. 결론

2007 년부터 사용된 OCEAN 은 KSCD 학술정보통합관리 시스템으로 정보화지원사업을 위한 통합관리 시스템이다. 학술정보 지원사업의 중심이 되는 OCEAN 은 끊임없는 시스템 개선과 유지보수에도 급변하는 정보기술 환경과 보안적인 측면에 있어 2 차년도에 거쳐 시스템을 재 구축하였다. 최신 프레임워크 적용과 중점 업무 프로세스 개선 그리고 DB 재설계 및 검색엔진 전환을 수행하였다. 이를 통해 2017 년에도 지속적이고 고품질의 데이터를 서비스 할 수 있게 되었다.

향후 연구로는 KISTI 가 2016 년 DOI-RA 의 10 번째 기관이 되어 DOI 등록 업무를 수행할 수 있게 됨에 따라 CrossRef/DOI 중심으로 설계된 DOI 관리 프로세스가 Korea DOI Center 도 수용할 수 있도록 개선이 필요하다.

## 참고문헌

[1] 김병규, 강무영, 박재원. (2004). KRISTAL-2002 기반의 학술정보관리시스템의 설계 및 구현. 한국정보과학회 학술발표논문집, 31(2Ⅱ), 211-213.

[2] KISTI, 정부 3.0 에 발맞춰 공공데이터 개방 - http://www.kisti.re.kr/promote/post/news/2269

[3] 강무영 외 7 인, KSCD 구축가이드라인, 한국과학기술정보연구원, 2015

[4] 서선경, 김병규, 최선희, 강무영. (2015). KSCD 통합관리시스템 설계 및 구현에 관한 연구. 한국정보관리학회 학술대회 논문집,61-65

[5] 박기성, 최재현, 김종배, 박제원 (2017). 아파치 스파크 기반 검색엔진의 설계 및 구현. 한국정보통신학회논문지, 21(1), 17-28.