논 단

Rosetta 시스템 도입과 디지털 콘텐츠 마이그레이션 추진

표 혜 리
(중앙도서관 정보관리과)

< 목 차 >

Ⅰ. 서론

디지털도서관팀에서는 차세대학술정보시스템 Alma 가동 지원을 위한 합의서를 체결하면서 SOLARS에 등록되어 있던 디지털 콘텐츠를 어떻게 관리하고 운영할지 고민하였고, 이에 따라 ExLibris사에서 제공하는 디지털 자원관리 보존을 위한 시스템을 사용하기로 결정하였다.

새로운 시스템 도입에 따른 디지털 콘텐츠의 효율적인 마이그레이션을 위해 고문헌, 학위논문, 학술행사 VOD 등 16개 콘텐츠 유형의 원문을 분석하고 데이터클리닝 작업을 진행하였다. 기존 디지털 콘텐츠는 2002년 전자도서관 운영에 이어 2006년 SOLARSⅢ에서 통합 운영되었는데, 디지털 콘텐츠 유형을 계속 추가하여 업로드하면서 경로가 복잡해졌다. 또한 SOLARS 시스템에 구축되어 있었던 MARC 데이터와 서울대학교에서 확장 개발한 XML 데이터를

변환하는 작업이 필요했다.

아울러 디지털 콘텐츠를 서비스하는 기존 뷰어들이 노후화로 인해 최신 정보 환경을 지원하지 않아 이용자 서비스에 불편함이 많았다. 대표적으로 고문헌 자료 같은 디지털 콘텐츠를 플러그인 방식인 DjVu 뷰어로 제공하는 경우 설치 과정에서의 불편함이 있었으며, 모바일 서비스 지원도 불가능하였다. 학술행사 VOD 디지털 콘텐츠의 경우 Microsoft사에서 DRM 시스템 서비스를 중단하여 더 이상 뷰어 서비스를 할 수 없는 상황이었다. 따라서 디지털 콘텐츠 각각의 특성을 반영하면서도 시대에 부합하는 뷰어의 개편이 필요하였다.

Ⅱ. 디지털 자원관리 시스템 Rosetta

1. 도입 배경

Alma는 2015년 도입 당시 디지털 콘텐츠를 관리할 수 있는 기능이 없었고 디지털 콘텐츠를 관리할 수 있는 시스템인 Alma-D는 개발 중이어서, 서울대학교의 디지털 콘텐츠 관리와 서비스를 위한 새로운 시스템 도입이 필요하였다. 2014년 9월 Alma 가동을 위한 합의서에서 ExLibris사가 디지털 자원의 보존 관리를 위한 시스템을 Alma-D가 개발되기 전까지 무상 제공하기로 하였고, ExLibris사는 Digitool과 Rosetta 두 개의 시스템을 제안하였다. 그러나 Digitool의 경우 업그레이드와 지원이 중단될 예정이었으므로 Rosetta 도입 계약을 같은 해 12월 24일 체결하게 되었다.

Rosetta는 디지털 자원을 효과적으로 보존하고 접근할 수 있도록 설계된, 웹을 기반으로 하는 디지털 자원관리 시스템이다. 오디오, 비디오와 문자 콘텐츠를 포함한 많은 양의 디지털 데이터를 저장하여 관리할 수 있고 고품질의 콘텐츠 보존을 위하여 다양한 관리 기능을 제공하고 있다.

2. 도입 경과

Rosetta의 도입 경과는 다음과 같다.

○ Alma 가동 지원을 위한 합의서 작성: 2014. 9. 23.
- ExLibris사 Digitool, Rosetta 제안

○ Rosetta 운영 서버 및 사양 검토: 2014. 10.
- 디지털 콘텐츠 자료 관리 및 서비스용 장비 구입

○ 기존 디지털 콘텐츠의 Rosetta 마이그레이션을 위한 회의: 2014. 12. 17.
- 디지털 콘텐츠 유형별 마이그레이션 범위 확인
- Rosetta 업무 흐름별 환경설정 및 콘텐츠 매칭 작업 등 협의

○ Rosetta 도입 계약 체결: 2014. 12. 24.

○ Rosetta 도입 Kick-off 회의: 2015. 1. 30.
- Rosetta 프로젝트 팀 역할 및 교육 일정 협의
- Rosetta 구축 일정 및 데이터 마이그레이션 방법 논의

○ Rosetta 교육 실시: 2015. 3. 30. ~ 2015. 4. 3.
- 참석자: 디지털도서관팀 및 고문헌자료실
ExLibris사 Nir Sherwinter Rosetta Product Manager, Timothee Lecaudey Project Manager

○ Rosetta의 운영체제 서브 스크립션 연장: 2015. 12. 8.
- 중앙도서관 DAM(Digital Asset Management, 디지털 자원관리)를 위한 Rosetta의 운영체제(OS) 서브 스크립션 연장
- Rosetta의 안정적인 운영을 위한 보안 업데이트 및 소프트웨어 업그레이드 지원(Red Hat Enterprise Linux Server, 3 copies)

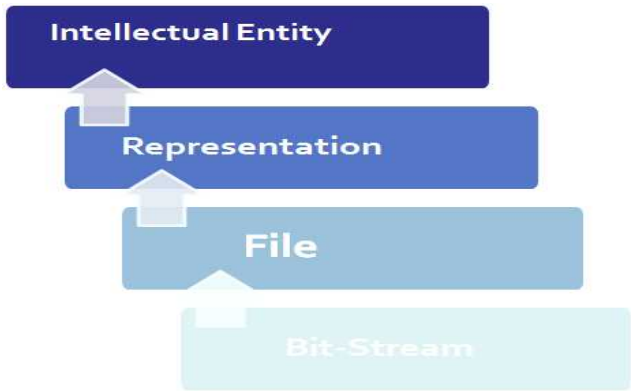
3. 데이터 모델 및 업무 흐름

Rosetta는 디지털 자원을 수집하여 보존하는 시스템으로 전체적인 구조는 PREMIS[1]에서 정의한 데이터 모델과 용어사전을 기반으로 시스템의 업무 흐름이 개발되었다.

데이터 모델은 아래 그림과 같이 비트스트림(Bit-Stream), 파일(File), Representation(이하 REP), Intellectual Entity(이하 IE)로 구성되어 있다. 비트스트림은 파일을 이루고 있는 최소한의 구성요건으로서 독립 실행형 파일로 전환이 어려운 단위이다. 이 비트스트림으로 구성된 파일은 운영 체제가 인식할 수 있는 바이트의 구성으로 이루어진다. 이러한 파일은 한 개 이상이 모여 하나의 REP이 되며, REP은 IE의 변환에 필요한 구조적 메타데이터를 포함하고 있다. IE는 디지털 자원 오브젝트를 뜻하며 최소한 하나

1) PREservation Metadata: Implementation Strategies의 약자. 디지털 아카이브를 위한 메타데이터 요소를 정의한 국제 워킹 그룹

이상의 REP으로 구성된다.

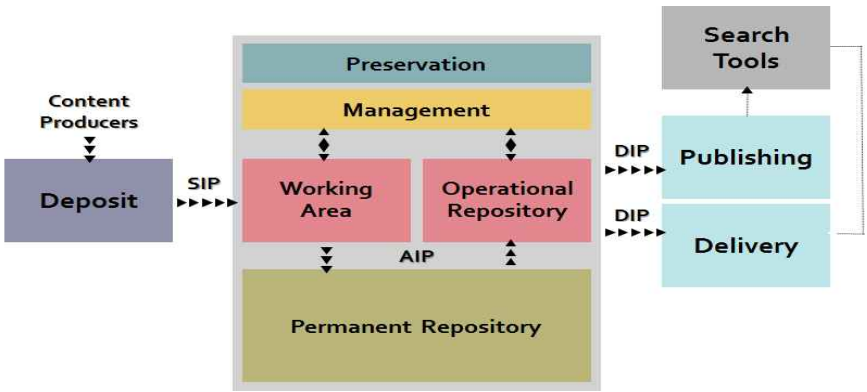


[그림 1] Rosetta 데이터 모델

책으로 비유하자면, 책의 한 페이지에 있는 글자들이 비트스트림이고, 한 페이지는 파일에 해당한다. 각각의 페이지들이 모여 하나의 장인 REP을 이루게 되고 이 장들이 모여 결국 한 권의 책인 IE를 완성하는 것이다. 이 때 Persistent Identifier(이하 PID)라는 고유 식별번호가 IE마다 부여되는데, 이는 다른 위치나 컬렉션으로 이동하더라도 IE가 존재하는 한 변동되지 않는 영구번호이기도 하다.

전체적인 구조와 업무 흐름을 살펴보면 아래 그림과 같이 Deposit(콘텐츠 업로드), Preservation(보존), Delivery(이용)으로 구분된다. Deposit은 디지털 콘텐츠를 업로드하는 일련의 과정을 모두 포함한다. 콘텐츠의 메타데이터가 작성되고 그에 해당하는 파일들이 첨부되어 IE를 생성한다. 구축된 콘텐츠는 Rosetta의 보존 모듈로 이동하여 저장되는데 이때 시스템에서는 SIP(Submission Information Package)이라는 정보패키지를 생성하여 전송한다. 보존 모듈에서는 제출된 정보들에 대한 시스템 내부의 검증과 프로세스를 진행하여 AIP(Archival Information Package)라는 보존용 정보패키지를 생성한다. 이때 AIP는 METS[2] XML의 형식으로 저장된다. 보존 모듈에 저장된 정보는 이용자들이 검색할 수 있도록 통합검색시스템 Primo로 보내지게 되는데 이 작업을 Publishing이라고 하며 이때 보내지는 정보의 패키지를 DIP(Dissemination Information Package)라고 한다.

2) Metadata Encoding and Transmission Standard의 약자로 디지털 자원의 메타데이터를 인코딩하고 전송하는 표준



[그림 2] Rosetta 시스템 모델링

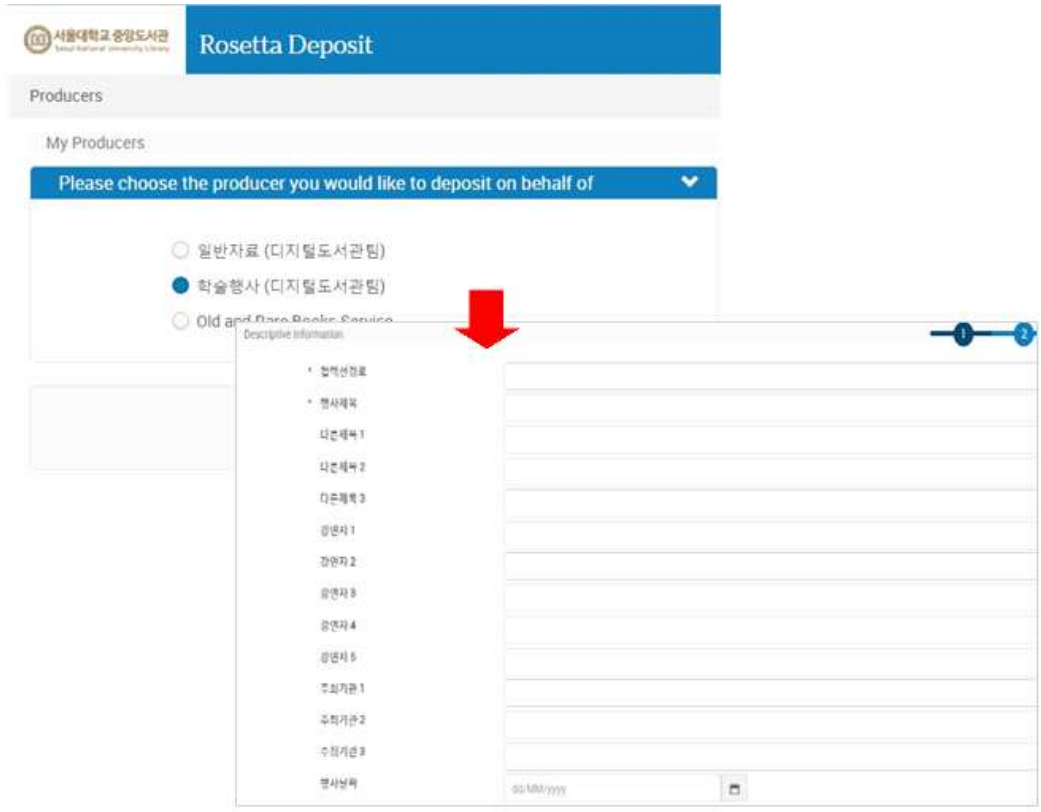
4. 구성 및 기능

Rosetta는 사용자의 권한과 활동의 종류에 따라 Deposit, Management, Administration 세 개의 홈페이지로 구분되어 있다.

[그림 3] Rosetta 로그인 화면

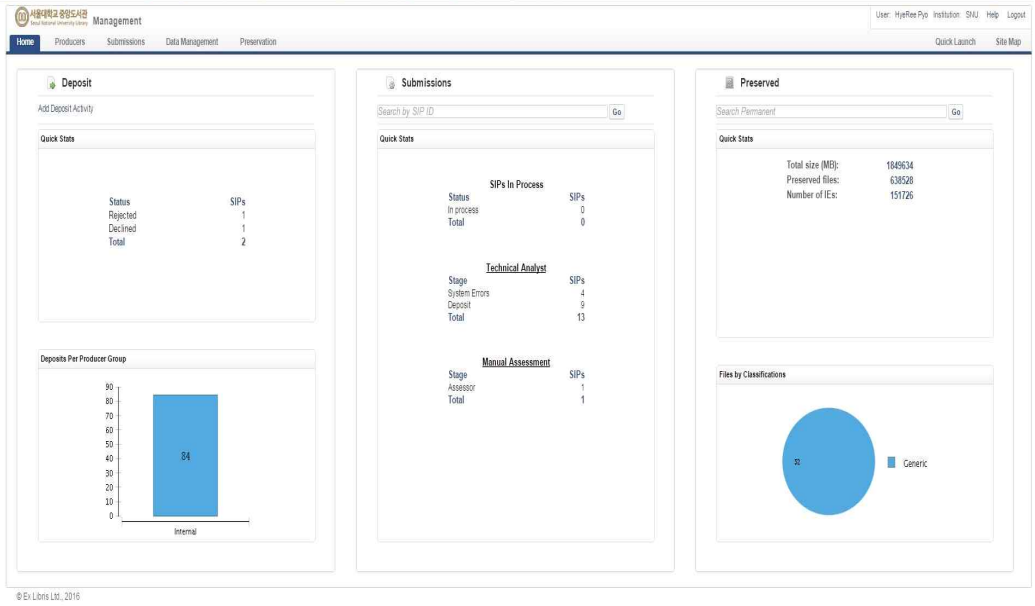
Deposit에서는 메타데이터 입력, 파일 업로드 등 콘텐츠를 수집하는 제반의 업무를 진행한다. Rosetta에서는 더블린 코어를 기반으로 메타데이터가 작성되는데, 콘텐츠의

특성에 맞게 입력할 메타데이터 항목을 만들어둔 템플릿으로 자료를 업로드할 수 있다. 예를 들어 학술행사 VOD는 '학술행사(디지털도서관팀)'이라는 생산자를 선택하여 IE Entity Type을 AcadEvent로 선택한 '학술행사'라는 템플릿으로 업로드를 한다. 이렇게 생산자가 업로드한 콘텐츠는 운영 서버로 이동되고, 운영 서버 안에서 SIP 처리를 하게 되난. 문제가 있을 경우 콘텐츠를 반환하고 문제가 없을 경우 업로드된 콘텐츠를 승인하여 영구 저장하게 된다.



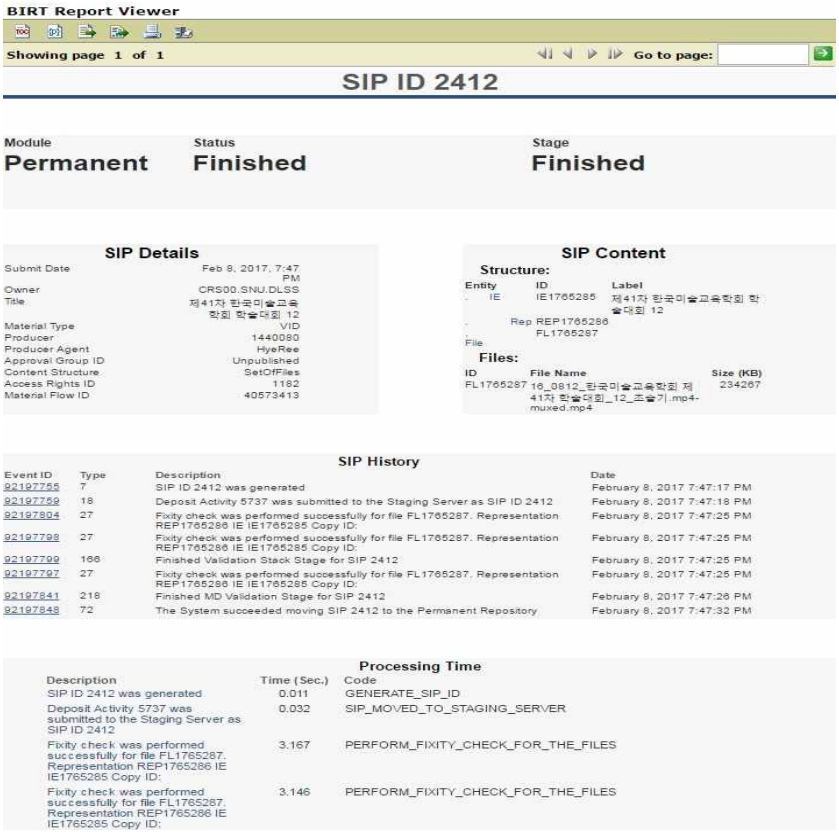
[그림 4] Deposit 페이지-학술행사 VOD 업로드 예시

Management는 로그인하면 Deposit, Submissions, Preserved 세 개의 칼럼으로 구분되어 있다. Deposit은 콘텐츠를 업로드하는 일련의 과정을 의미하고, Submissions에서는 수집된 콘텐츠 등록에 관한 업무를 처리하며, Preserved에서는 콘텐츠를 보존하는 업무를 수행할 수 있다. 각 칼럼과 관계된 통계, 차트 등 시각화된 위젯이 제공됭 정보를 확인하는데 용이하다.



[그림 5] Management 페이지 화면

Deposit 칼럼에서의 기능은 앞서 설명한 Deposit 페이지에서의 기능과 같다. Submissions 칼럼에서는 SIP와 관련된 작업을 할 수 있다. 진행 중이거나 완료된 SIP 작업 과정이나 그에 따른 기술 통계 등을 확인한 후 에러 발생 여부와 원인도 필터링하여 검색할 수 있다.


BIRT Report Viewer

Showing page 1 of 1 Go to page:

SIP ID 2412

Module	Status	Stage
Permanent	Finished	Finished

SIP Details

Submit Date	Feb 8, 2017, 7:47 PM
Owner	CRS00.SNU.DLSS
Title	제41차 한국미술교육학회 학술대회 12
Material Type	VID
Producer	1440080
Producer Agent	HyeRee
Approval Group ID	Unpublished
Content Structure	SetOfFiles
Access Rights ID	1182
Material Flow ID	40573413

SIP Content

Structure:

Entity	ID	Label
IE	IE1765285	제41차 한국미술교육학회 학술대회 12
Rep	REP1765286	
File	FL1765287	

Files:

ID	File Name	Size (KB)
FL1765287	16_0812_한국미술교육학회 제41차 학술대회_12_조슬기.mp4-muxed.mp4	234267

SIP History

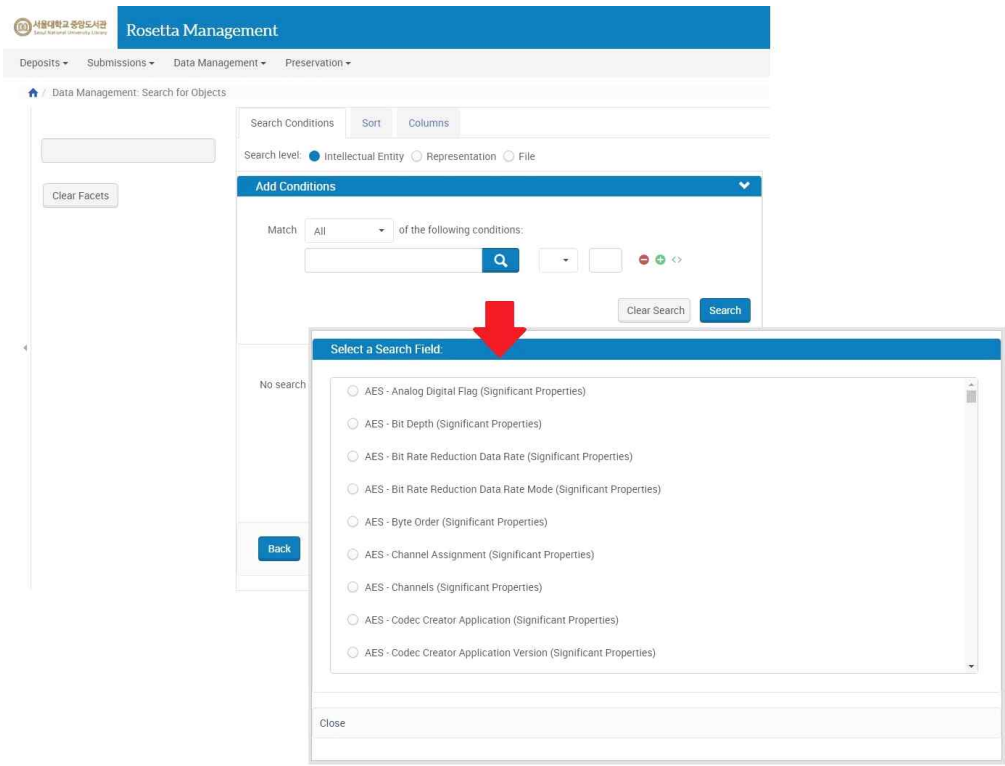
Event ID	Type	Description	Date
92197755	7	SIP ID 2412 was generated	February 8, 2017 7:47:17 PM
92197759	18	Deposit Activity 5737 was submitted to the Staging Server as SIP ID 2412	February 8, 2017 7:47:18 PM
92197804	27	Fixity check was performed successfully for file FL1765287. Representation REP1765286 IE IE1765285 Copy ID:	February 8, 2017 7:47:25 PM
92197798	27	Fixity check was performed successfully for file FL1765287. Representation REP1765286 IE IE1765285 Copy ID:	February 8, 2017 7:47:25 PM
92197799	166	Finished Validation Stack Stage for SIP 2412	February 8, 2017 7:47:25 PM
92197797	27	Fixity check was performed successfully for file FL1765287. Representation REP1765286 IE IE1765285 Copy ID:	February 8, 2017 7:47:25 PM
92197841	218	Finished MD Validation Stage for SIP 2412	February 8, 2017 7:47:26 PM
92197848	72	The System succeeded moving SIP 2412 to the Permanent Repository	February 8, 2017 7:47:32 PM

Processing Time

Description	Time (Sec.)	Code
SIP ID 2412 was generated	0.011	GENERATE_SIP_ID
Deposit Activity 5737 was submitted to the Staging Server as SIP ID 2412	0.032	SIP_MOVED_TO_STAGING_SERVER
Fixity check was performed successfully for file FL1765287. Representation REP1765286 IE IE1765285 Copy ID:	3.167	PERFORM_FIXITY_CHECK_FOR_THE_FILES
Fixity check was performed successfully for file FL1765287. Representation REP1765286 IE IE1765285 Copy ID:	3.146	PERFORM_FIXITY_CHECK_FOR_THE_FILES

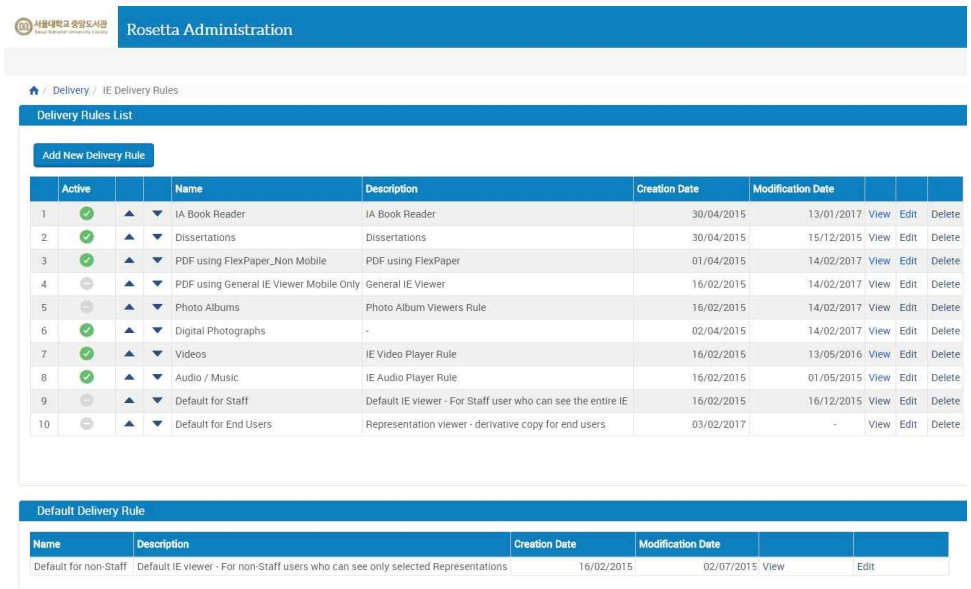
[그림 6] Submissions 칼럼 화면

Preserved 칼럼에서는 영구 보존된 IE를 검색할 수 있다. 고급 통계 기능을 활용하여 각종 조건에 따른 쿼리를 만들어 원하는 IE를 편리하게 검색할 수 있고, 해당 IE 메타데이터 등을 수정할 수도 있다.



[그림 7] Preserved 칼럼 화면

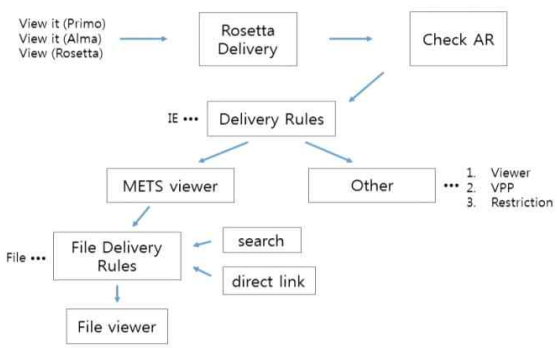
Administration페이지에서는 Rosetta의 제반 환경설정을 할 수 있다. Delivery 항목에서 업로드된 콘텐츠의 이용과 관련한 사항 설정, IE Delivery 규칙 설정, 뷰어와 데이터가 저장된 리포지터리의 설정 등을 할 수 있다. 이외에도 Rosetta 일반 설정과 개별 세부 항목 설정을 하여 디지털 콘텐츠의 특성에 맞는 뷰어와 환경을 정할 수 있다.



[그림 8] Administration 페이지-Delivery 규칙 설정

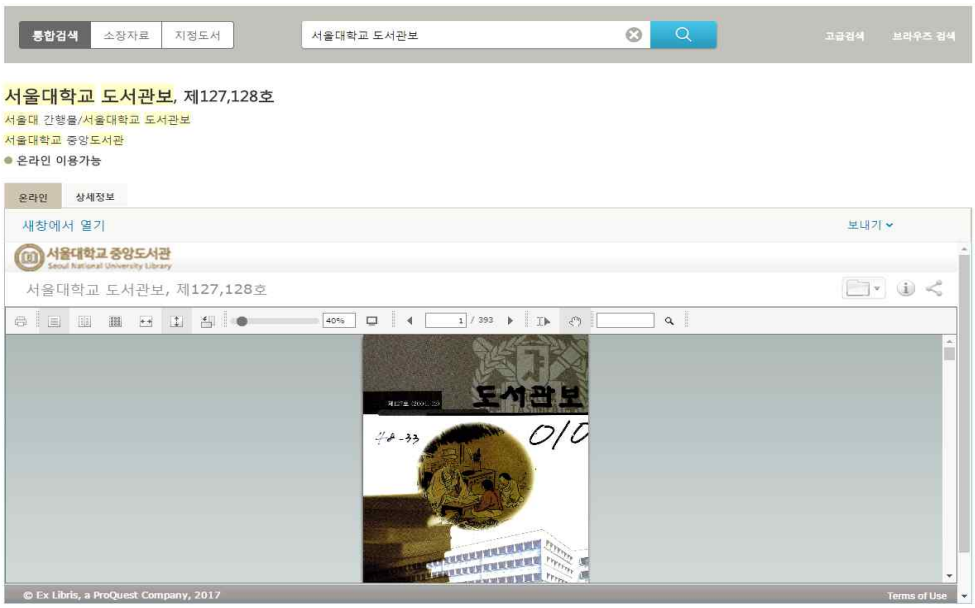
5. 디지털 콘텐츠의 이용

승인된 콘텐츠는 Publishing 모듈을 통해 Primo로 전송된다. 이때 Rosetta Delivery 관리 모듈에서는 이용자가 요청한 IE의 PID를 기준으로 검색 작업을 처리하게 된다. 이용자가 해당 콘텐츠 공개범위에 적절한 권한을 갖고 있는지 검토하고, Delivery 규칙 관리에 따라 콘텐츠의 파라미터를 확인하고 상세정보, 뷰어를 선정하여 제공한다.

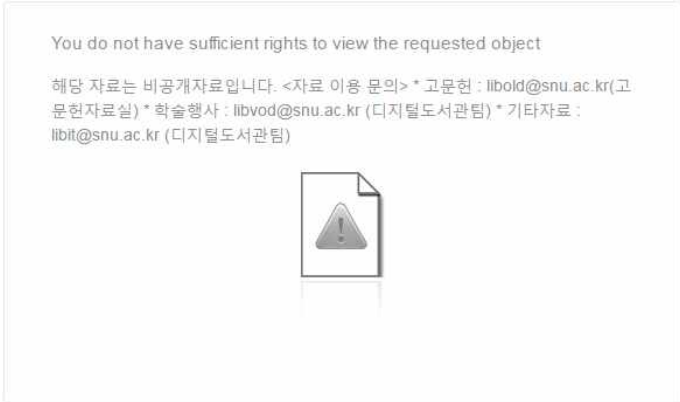


[그림 9] Rosetta Delivery 모델링

예를 들어 이용자가 Primo에서 ‘서울대학교 도서관보’라는 자료를 검색하여 이용할 때, 이 콘텐츠는 전체공개 자료이므로 누구나 볼 수 있으며 Delivery 규칙에 따라 ‘서울대 간행물’인 자료는 IA Book Reader라는 뷰어를 사용한다. 학술행사 VOD는 JWPlayer라는 뷰어를 통해 제공되는데, 각각의 뷰어에서는 공개된 자료와 비공개된 자료의 뷰어 화면이 다르게 제공된다.

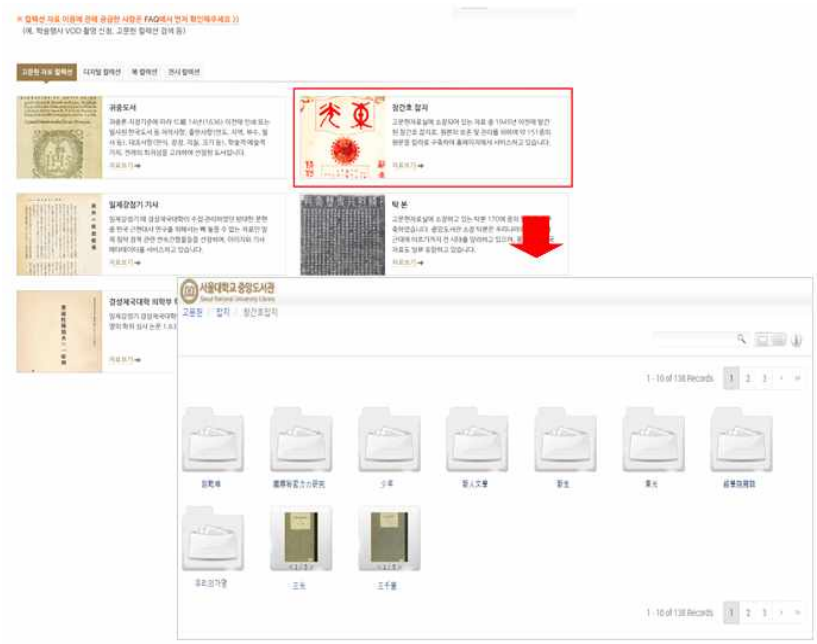


[그림 10] Primo에서 Rosetta 콘텐츠 이용 예시



[그림 11] 비공개 디지털 콘텐츠 제공 화면 예시

또한 Rosetta에서는 컬렉션 계층구조를 생성하여 관리할 수 잇다. 콘텐츠의 특성에 따라 계층관계를 세분화하여 IE를 제공할 수 있도록 해준다. 도서관 홈페이지에서는 디지털 콘텐츠를 컬렉션으로 구분하여 리스트를 제공하고 있는데, Rosetta에서 생성한 계층구조에 따라 디스플레이 된다.



[그림 12] 디지털 컬렉션 구성 및 Rosetta 계층구조

Ⅲ. 디지털 콘텐츠 마이그레이션

1. 추진 배경

차세대학술정보시스템 Alma 도입과 함께 기존 SOLARS 서버에 잇었던 디지털 콘텐츠를 Rosetta 서버로 이관하는 작업을 진행하였고, 그동안 디지털 콘텐츠 별로 특성화 페이지를 운영하여 관리가 이원화되었던 부분을 일원화하고자 하였다. 그러나 디지털 콘텐츠의 양이 방대하여 한 번에 이관하기는 불가능하였으므로, 2015년부터 디지털 콘텐츠 마이그레이션 사업을 추진하였다. 수차례 회의를 거쳐 결정된 순위에 따라

단계적으로 컬렉션을 이전하기로 하였으며, 이관 전까지 디지털 콘텐츠는 기존 특성화 페이지에서 제공하기로 하였다. 디지털 콘텐츠를 마이그레이션하는 과정에서 데이터클리닝 작업을 선행하여 기존의 데이터를 재점검하고 정제된 데이터로 구축하는 것을 원칙으로 하였다.

2. 추진 경과

디지털 콘텐츠 마이그레이션 추진 경과는 다음과 같다.

○ 디지털 콘텐츠 마이그레이션을 위한 사전 준비: 2015. 5. 12. ~ 25.
○ 디지털 콘텐츠 마이그레이션 사업 추진 기본 계획 수립: 2015. 6. 4.
○ 디지털 콘텐츠 마이그레이션 사업 추진 기본 계획 변경: 2015. 6. 19.
- 주요 컬렉션(고문헌, 학술행사 VOD) 중심으로 1차 사업 추진
- 학위논문 원문은 dCollection으로 이관
- 기타 컬렉션은 이후 단계별로 사업 진행
대상 파일 경로 오류 및 누락 페이지 재정비
○ 제1차 디지털 콘텐츠 마이그레이션 사업: 2015. 7. 8. ~ 2015. 8. 31.
- 고문헌 컬렉션 마이그레이션: 파일 518,265건, 메타 142,982건
- 학술행사 VOD 컬렉션 마이그레이션: 파일 6,307건, 메타 6,307건
○ 제2차 디지털 콘텐츠 마이그레이션 사업 : 2015. 12. 7. ~ 2016. 2. 12.
- 서울대/기타 간행물 컬렉션 마이그레이션: 국내・외 단행본, 국내・외 연간물 파일 58,773건, 메타 4,103건
○ 제3차 디지털 콘텐츠 마이그레이션 사업: 2016. 11. 14. ~ 2017. 2. 10.
- 음악자료 컬렉션 마이그레이션: 파일 6,914건, 메타 5,607건
- 디지털사진관 컬렉션 마이그레이션: 파일 3,354건, 메타 3,357건
- 대학신문 컬렉션 마이그레이션: 파일 84,976건, 메타 84,895건
- 대학사료 컬렉션 마이그레이션: 파일 360,770건, 메타 25,349건

3. 준비 작업

디지털 콘텐츠 마이그레이션을 위해 SOLARS에 구축되어 있는 디지털 콘텐츠를 분석하여 다음과 같은 과정으로 이관 준비 작업이 진행되었다. 이관 준비 작업은 크게 두 가지로 나눌 수 있다. 첫째, SOLARS 디지털 콘텐츠를 분석하고 이중 관리되고 있는 파일 형식을 통합하여 데이터 파일을 변환할 프로그램을 개발하였다. 둘째, Rosetta 시스템의 환경을 설정하고 뷰어를 연동하였다.

디지털 콘텐츠의 효율적 이관을 위해 먼저 고문헌, 학위논문, 학술행사 등의 원문 분석 및 데이터클리닝 작업을 진행하였다. 2002년 전자도서관 운영에 이어 2006년 SOLARS에서 통합 운영되었던 디지털 콘텐츠는 시스템이 변경되면서 복잡한 경로로 구축되었고 파일 정보가 누락되거나 원본 파일 없이 서비스용 파일만 있는 콘텐츠가 존재하는 경우가 많았다. 복잡한 디지털 콘텐츠 형식을 재정비하고, SOLARS에 구축되어 있는 MARC 데이터와 서울대학교에서 확장 개발한 XML 데이터를 더블린 코어 데이터로 변환하였다.

뷰어와 관련하여 다음과 같은 작업을 하였다. 뷰어 커스터마이징이 이루어질 경우 Alma-D에서도 뷰어를 커스터마이징하게 되어 비용이 이중으로 발생할 수 있었다. 이에 기존 뷰어는 사용하지 않는 대신 Rosetta에서 기본으로 제공하는 자체 뷰어를 원칙적으로 사용하고, 예외적으로 일부 컬렉션을 위해 외부 뷰어를 사용할 경우 별도의 어플리케이션 다운이 필요 없는 웹 베이스 오픈소스 기반의 무료 뷰어를 선택하였다.

이미지 및 책자 뷰어 가운데 DjVu 뷰어와 ezPDF 뷰어의 경우 기존 프로그램을 유지하기 위해서는 별도의 개발이 필요했기 때문에 모두 Rosetta 자체 뷰어 중 IA Book Reader 뷰어로 대체하였다. 특히 DjVu 뷰어로 서비스되던 파일의 경우 보존용 파일이 누락된 채로 서비스용 파일만 남아있는 경우도 많았으므로 보존용이 없을 경우 서비스용 파일을 마이그레이션하였다. IA Book Reader 뷰어는 JPG 포맷에 최적화되어 있었으므로 DjVu 포맷 형식을 JPG 포맷으로 변환하였다. 학술행사 VOD DRM의 경우 Windows 환경에서만 지원되고 2015년 7월 이후 Microsoft사에서 Window Media Player 뷰어의 DRM 시스템 서비스를 중단하여 새로운 뷰어가 필요하게 되었다. 그러나 자체 뷰어 중 적절한 뷰어가 없어 동영상과 음성을 동시에 처리할 수 있는 JWPlayer 뷰어를 선택하였다.

4. 추진 내용

디지털 콘텐츠 마이그레이션의 대상 파일과 추진현황은 아래와 같다. 마이그레이션 회의를 통해 중요도 순서대로 이관 순위를 결정하였다.

[표 1] 전체 디지털 콘텐츠 마이그레이션 대상파일

(2016. 12. 31. 기준)

순위	콘텐츠유형	이전 파일포맷		현재 파일포맷	원본자료	데이터 형식	건수 (메타)	사업구분 (저장위치)
		보존용	서비스용					
1	고문헌	JPG	DjVu	JPG	책자	XML	142,994	1차 완료 (Rosetta)
2	학술행사	WAV MPEG2	ASF	MP4	테이프 파일	XML	6,880	1차 완료 (Rosetta)
3	학위논문	TIFF 400dpi	PDF	PDF	책자	MARC	113,765	1차 완료 (dCollection)
4	서울대/기타 간행물	JPG TIFF	PDF DjVu	PDF	책자 파일	MARC XML	4,100	2차 완료 (Rosetta)
5	디지털사진관	TIFF 600dpi TIFF 4,000dpi	JPG 600dpi	JPG	사진	XML	3,357	3차 진행
6	음악작품	TIFF JPG WAV MPEG2	PDF ASF	JPG MP4	악보 테이프	XML	11,011	3차 진행
7	대학신문 (기사,사진)	JPG 600dpi	DjVu 600dpi	JPG	책자 사진	XML	86,606	3차 진행
8	대학사료	JPG 400dpi WAV MPEG2	DjVu 400dpi ASF	JPG	책자	XML	25,557	3차 진행
9	학내간행물	TIFF 400dpi	PDF 400dpi		책자	MARC XML	1,979	2017년도 추진 예정
10	농학자료	JPG 600dpi JPG 4,000dpi	DjVu 600dpi		슬라이드	XML	44,481	
11	지리학자료	JPG 600dpi JPG 2,000dpi	DjVu 600dpi		슬라이드	XML	3,105	
12	미술작품	TIFF 400dpi	PDF 300dpi		책자	XML	14,549	

		JPG 600dpi WAV MPEG2	DjVu 600dpi ASF		필름			
13	**한국병합사료**	JPG 600dpi WAV MPEG2	DjVu 600dpi ASF		책자 테이프	XML	452	
14	**한국교육사고**	JPG 400dpi WAV MPEG2	PDF 200dpi DjVu 400dpi ASF		책자 테이프	XML	513	
15	**의학자료**	JPG 600dpi JPG 4,000dpi	DjVu 600dpi		슬라이드	XML	50,654	
16	**곤충자료**	JPG 72dpi	JPG 72dpi		표본	XML	4,213	
합 계							514,216	

1차 디지털 콘텐츠 마이그레이션은 예산을 확보하여 계속적으로 사업을 진행하고 있는 고문헌 및 학술행사 VOD, 학위논문을 대상으로 하였다. 최종 변환된 디지털 파일은 중앙도서관에서 요청한 Network File System[3]으로 연결하였으며, 지정된 하위 디렉토리와 컬렉션에 매칭될 수 있도록 하였다.

고문헌은 원문 518,265건, 메타 142,982건, 학술행사 VOD는 원문 6,307건, 메타 6,307건을 변환하였는데, 파일 변환은 자동 변환 프로그램을 활용하였고으며 변환 완료된 파일은 원본은 변환 전 해상도와 동일한 수준을 유지하였다.

3) 네트워크 파일 체계. 클라이언트 컴퓨터의 사용자가 네트워크 상의 파일을 직접 연결된 스토리지에 접근하는 방식과 비슷한 방식으로 접근하도록 도와주는 프로토콜

[표 2] 1차 마이그레이션 대상 콘텐츠

구분	메타데이터		변환전원문		변환후원문		
	형식	건수	형식	건수	형식	건수	사이즈(GB)
고문헌	MARC	1,132	DjVu	3,856	JPG	317,198	295.06
	XML	139,147	DjVu	139,574	JPG	249,135	143.63
학술행사 VOD	XML	6,694	ASF	6,634	MP4	6,425	728.71
					MP3	202	13.66
			WMV	60	MP4	60	4.67

고문헌의 경우 최종 산출물은 JPG 형식이며, 하나의 콘텐츠에 다수의 페이지가 있는 경우 각 페이지 별로 JPG 파일을 생성하였다. 디지털 파일은 각 콘텐츠별로 나누어 파일 이름을 붙였는데, SOLARS 콘텐츠번호와 MOI번호를 준용하였다.

학술행사 VOD의 경우 기본적인 항목은 고문헌 콘텐츠와 동일하며 최종 산출물은 VOD(영상)인 경우 MP4, AOD(음성)인 경우 MP3 형식으로 하였고, 만약 하나의 콘텐츠에 다수의 AOD와 VOD 파일이 있는 경우 개별 파일로 구분해서 변환하였다.

학위논문의 경우 관리의 일원화를 위해 학위논문 온라인 제출 및 서비스 시스템인 dCollection으로 마이그레이션하였다.

[표 3] 1차 마이그레이션 변환 완료 파일 세부 내용

구축유형	원본자료	원문유형	변환 전	변환 후	
고문헌	책자	이미지	TIFF 400dpi	JPG 400dpi	
			DjVu A4 600dpi A3 400dpi A2 300dpi A1 200dpi (경성제대) 300 dpi	JPG A4 600dpi A3 400dpi A2 300dpi A1 200dpi (경성제대) 300 dpi	
학술행사 VOD	동영상	VOD	ASF, WMV	MP4	◦ Bitrate : 528kbps 이상 ◦ Size : 320×240(4:3) 576×320(16:9) ◦ Video Streaming Output
		AOD	ASF, WMV	MP3	◦ Bitrate : 128kbps 이상 ◦ Audio Streaming Output

2차 디지털 콘텐츠 마이그레이션은 서울대/기타 간행물 원문 58,773건, 메타 4,103건을 대상으로 진행하였다. 기본적인 방식은 1차 마이그레이션과 동일하였으나, 데이터클리닝 과정에서 IA Book Reader 뷰어와 맞지 않는 가로 형태의 자료가 많았으며 경로가 유실되어 재연결해야 하는 파일이 많아 예상보다 오랜 시간이 소요되었다. 이에 따라 보존 인력을 확보한 후 자동 측정 프로그램을 이용하여 가로 형태 자료를 별도로 추출한 후 단면 분할하여 세로 형태의 자료로 만들고 경로가 유실된 파일을 보완하였다.

[표 4] 2차 마이그레이션 대상 콘텐츠

구분	메타데이터		변환 전 파일		변환 후 파일		
	형식	건수	형식	건수	형식	건수	사이즈
서울대/기타 간행물	MARC	1,537	DjVu	2,992	JPG	451,047	329.54 GB
			PDF	1,111	PDF	1,111	37.15 GB
			JPG	1	JPG	중복변환대상 삭제	

[표 5] 2차 마이그레이션 변환 완료 파일 기준

구축유형	원본자료	원문유형	변환 전	변환 후
서울대/기타 간행물	책자	이미지	TIFF 100~400dpi	JPG 100~400dpi
			DjVu 100~400dpi	
			DjVu 100~400dpi	
			PDF 100~400dpi	

3차 디지털 콘텐츠 마이그레이션은 음악자료, 디지털사진관, 대학신문, 대학사료 컬렉션을 대상으로 진행하였다. 기본적인 방식은 1차 및 2차 마이그레이션과 동일하며, 음악자료 파일 6,914건, 메타 5,607건, 디지털사진관 파일 3,354건, 메타 3,357건, 대학신문 파일 84,976건, 메타 84,895건, 대학사료 파일 360,770건, 메타 25,349건이 마이그레이션되었다.

[표 6] 3차 마이그레이션 대상 콘텐츠

구분	메타데이터		변환 전 파일		변환 후 파일		
	형식	건수	형식	건수	형식	건수	사이즈
음악자료	XML	5,607	ASF	4,631	MP4	272	23.47 GB
					MP3	4,339	36.80 GB
			PDF	2,303	PDF	2,303	61.81 GB
디지털사진관	XML	3,357	CTL_JPG	3,357	JPG	3,354	6.01 GB
대학신문	XML	84,895	DjVu	9,275	JPG	9,275	318.21 GB
			JPG	75,710	JPG	75,701	
			PDF	1	PDF	0	*중복변환대상
대학사료	XML	25,349	DjVu	25,367	JPG	360,770	254.08 GB

[표 7] 3차 마이그레이션 변환 완료 파일 세부 내용

구축유형	원본자료	원문유형	변환 전	변환 후	
음악자료	동영상	VOD	ASF	MP4	◦ Bitrate : 528kbps 이상 ◦ Size : 320×240(4:3) 576×320(16:9) ◦ Video Streaming Output
		AOD	ASF	MP3	◦ Bitrate : 128kbps 이상 ◦ Audio Streaming Output
디지털사진관 대학신문 대학사료	이미지	JPG Djvu	TIFF 400dpi	JPG 400dpi	
			DjVu A4 600dpi A3 400dpi A2 300dpi A1 200dpi	JPG A4 600dpi A3 400dpi A2 300dpi A1 200dpi	

마이그레이션 작업과정은 다음과 같다. 각 디지털 콘텐츠의 메타데이터 다운로드 및 반출은 XML, MARC 데이터 파싱을 통한 Rosetta 시스템 정의 Element를 추출하는 작업을 우선 진행한다. 이 과정에서 매핑룰 작성 및 이미지 서버 디렉토리 룰을 지정한다. 이후 데이터클리닝을 통해 정제된 데이터의 메타데이터를 Rosetta 시스템 업로드를 위한 CSV 파일로 변환 생성한 후, 필요한 정보를 추가 입력하여 업로드용 파일을 만들어 지정된 위치의 공용 서버에 업로드한다. 지정된 서버의 디렉토리 아래 위치해 있는 파일을 자동으로 업로드해주는 Rosetta 시스템의 NFS 템플릿을 통해 Rosetta 시

스템 서버에 업로드를 진행하며, 업로드가 끝났을 경우 Rosetta 시스템 검수 툴을 통해 에러가 있는지 여부를 검수한다. 이후 Rosetta 시스템 Data Management에서 컬렉션 수정 및 메타데이터 교정을 거쳐 통합검색시스템 Primo로 퍼블리싱하게 된다.

Ⅳ. 향후 계획

서울대학교 중앙도서관에서는 개교 70주년 기념 전시회 및 책자 발간 등을 통해 사진 자료의 중요성을 재인식하고 사진자료 수집을 원활히 하고자 기존 디지털사진관 홈페이지를 개편하였다. 또한 디지털 콘텐츠의 유기적인 구축 및 이용을 위하여 전 기관에 공문을 발송하여 서울대학교 학내 기관 발간자료 원문을 수집하고, 특히 2016년도 서울대학교 개교 70주년 기념으로 추진하였던 학내 발간자료를 수집하여 서울대학교의 역사와 정신을 공유하는 기념 컬렉션을 만듦으로써 디지털 콘텐츠 구축 및 관리의 주체로서의 중앙도서관 역할을 다할 것이다.

[그림 14] 개편된 디지털사진자료관 홈페이지 모습

현재 서울대학교 중앙도서관에서 사용하고 있는 Rosetta 시스템은 Alma-D의 개발 이전까지 한시적으로 제공되는 시스템이니만큼 한정된 내장 뷰어, 생성하기 까다롭고 콘텐츠 포맷마다 필요한 템플릿 등의 한계점을 가지고 있다. Alma-D가 정식 출시하게 되면 시스템의 기능을 면밀히 분석한 후 이관할 계획이다.

데이터클리닝을 통해 문제가 있는 파일을 수정하고 보완하여 디지털 콘텐츠 컬렉션의 서비스의 질을 높이며 디지털 콘텐츠 마이그레이션으로 기존 서비스와 동일한 콘텐츠 이용 환경을 조성하도록 연동하여 하며, 향후 이용자들의 요구를 파악하여 컬렉션 정비를 통해 끊임없이 소통하는 디지털 콘텐츠 정책을 운용하여야 할 것이다.

앞으로도 곤충자료, 의학자료 등 디지털 콘텐츠를 Rosetta 시스템으로 모두 마이그레이션할 계획이다. 향후 디지털사진관 홈페이지 등의 개편으로 디지털 콘텐츠 구축 및 마이그레이션을 동시에 진행할 수 있도록 Rosetta 시스템 템플릿을 기반으로 하여 메타데이터를 작성하는 표준 포맷을 만들어 적용할 예정이다. 향후 서울대학교 내에서 기관간의 프로토콜을 적용하여 소장기간이나 생산기관에 상관없이 디지털 콘텐츠를 일괄적으로 구축 및 관리함으로써 역사적, 기록적으로 중요한 의미를 갖는 컬렉션을 이루어나가도록 할 계획이다. 또한 지속적인 데이터클리닝을 통해 유실된 원본 파일의 수정 및 보완이 적시에 이루어지도록 하여, 수평적 구조의 Dublin Core를 기반으로 하는 디지털 콘텐츠 관리를 정기적으로 함으로써 디지털 콘텐츠의 유실을 방지할 예정이다.

Ⅴ. 맺음말

21세기에 들어 IT 기술의 급변화로 인해 이용자의 다양한 요구가 증가함에 따라, 서울대학교 중앙도서관에서도 디지털 콘텐츠의 관리 필요성과 막중한 책임감을 느끼지 않을 수 없게 되었다. 예를 들어 3D 영상 등의 디지털 콘텐츠는 최적의 뷰어를 찾기 위해 범용성 및 호환성 등을 고려하여 많은 탐색이 이루어지고 있으며, VR 콘텐츠 등 새롭게 나올 매체와 콘텐츠를 수용할 수 있는 포맷을 찾기 위해 노력하고 있다.

사업을 진행하면서 과거 디지털 콘텐츠에 대한 작업이 특히 수월하지 않았다. 디지털 콘텐츠 마이그레이션을 하면서 데이터클리닝을 기반작업으로 진행하면서 다양한 목록 규칙 및 기술 방식을 확인할 수 있었다. 또한 많은 메타데이터들이 목록자의 성향에 따라 나열되어 있어 데이터를 정비하는데 많은 시간과 비용, 노력이 들어갔다.

SOLARS 시스템에서는 단행본 등의 서지 레코드와 결합하여 디지털 콘텐츠를 제공하였기 때문에 따로 메타데이터를 작성할 필요가 없었고, 서지 레코드의 관리만으로도 디지털 콘텐츠의 데이터를 함께 관리할 수 있었다. 그러나 Rosetta 시스템을 사용하면서 디지털 콘텐츠가 서지 레코드와 분리되어 따로 관리 및 운영되면서, 우리 도서관은 단행본 중심의 MARC 21 또는 KORMARC 등의 기술방식과는 다른 디지털 콘텐츠 중심의 Dublin Core라는 메타데이터 기술방식을 사용하고 있다.

Dublin Core가 디지털 콘텐츠 기술방식으로 주목되고 있지만 완전한 목록의 대체 수단이 되기 위해서는 보다 완벽한 규칙의 정비와 함께 폭넓은 합의가 있어야 할 것이다. 원래 메타데이터는 자료의 상세한 기술보다는 신속한 탐색을 목적으로 한 것이므로 목록에 비해 식별요소가 부족하고, 전거파일과 같은 제어수단이 사용되지 않아 도서관의 서지데이터와는 질적으로 차이가 있기 때문이다. 또한 Dublin Core의 특성상 단순성과 범용성을 전제로 하기 때문에 자유로운 기술이 가능하다는 장점이 있는 반면, 기술자의 개성이 강하게 드러남으로써 일관성을 잃어버리기 쉽다는 점이 있다.

이러한 점으로 말미암아, 마이그레이션 사업이 종료된 후 정리된 메타데이터를 기반으로 디지털 콘텐츠 목록 규칙을 통일하여 디지털 콘텐츠의 표준화된 구조를 갖추어야 할 것으로 생각한다. 또한 디지털 콘텐츠의 특성에 부합하는 적합한 형태의 뷰어 및 포맷을 지속적으로 연구하여 보다 안정적인 관리와 보존을 위한 노력을 하여야 할 것이다.

[붙임]

디지털 콘텐츠 마이그레이션 중 국내 고서 매핑룰 예시

	Rosetta DC	SNU MARC			비고
SIP	Title (DC)	1-9 국내고서_01			
IE	IE Entity Type	OldBook			
	Type (DC)	고문헌			
	Identifier (DC)	001			
	Language (DC)	008	35		필요하면 코드테이블 따로 첨부
	Subject - DDC (DC)	082	$a		
	Contributor (DC)	100	$a	$d	$a와 $d추출시 구분기호 그대로 출력
	Contributor (DC)	110	$a		
	Contributor (DC)	130	$a		
	Title (DC)	245	$a	$b	$a와 $b추출시 구분기호 그대로 출력
	Title - Alternative (DC)	245	$x		
	Title - Alternative (DC)	246	$a		
	Relation - IsVersion of (DC)	250	$a		
	Publisher (DC)	260	$b		
	Date (DC)	260	$c		
	Format (DC)	300	$a		
	Subject (DC)	600	$a	$d	$a와 $d추출시 구분기호 그대로 출력
	Subject (DC)	650	$a		
	Subject (DC)	653	$a		
REP	Preservation Type	PRESERVATION_MASTER			
REP	Access Rights Policy ID (REP)	O -> 1182			mo_info테이블에서 open_type값 참조하여 치환(전체공개)
REP	Access Rights Policy ID (REP)	N -> 1181			mo_info테이블에서 open_type값 참조하여 치환(비공개)
REP	Access Rights Policy ID (REP)	C -> 1182			mo_info테이블에서open_type값이 C인경우-mo_copyright테이블에서 service_ restrict 참조하여 치환(제한공개) 1)1~3자리에Y값이하나라도있으면 -1184(관내공개) 2)4~6자리에Y값이하나라도있으면 -1183(학내공개) 3)7~9자리에Y값이하나라도있으면 -1182(전체공개)
REP	Access Rights Policy ID (REP)	C -> 1183			
REP	Access Rights Policy ID (REP)	C -> 1184			
File	File Original Path				NFS로 올릴 경우 Path정보는 필수아님
File	File Original Name				