발간등록번호
11-1371029-100022-01



디지털자원의 장기보존전략 및 인프라 구축 방안 연구

2025. 12.

국립중앙도서관

발간등록번호
11-1371029-100022-01

디지털자원의 장기보존전략 및 인프라 구축 방안 연구

2025. 12.

문화체육관광부
국립중앙도서관

제 출 문

국립중앙도서관장 귀하

본 보고서를 “디지털자원의 장기보존전략 및 인프라 구축 방안 연구” 최종 보고서로 제출합니다.

2025년 12월 15일

연 구 기 관 : ㈜디아이엔티
연구책임자 : 최성만 대표
참여연구원 : 이상헌 수석컨설턴트
박지혜 선임컨설턴트

.

이 연구는 2025년도 국립중앙도서관의 연구개발비로 수행되었으며,
본 연구에서 제시된 정책적 대안이나 의견 등은
국립중앙도서관의 공식 의견과 다를 수 있습니다.

차 례

요약문 ····· 1

제1장 연구개요

1.1. 연구의 배경 및 필요성 ····· 5
1.1.1. 연구의 배경 ····· 5
1.1.2. 디지털자원 장기보존의 필요성 ····· 5
1.2. 연구의 범위 ····· 6
1.3. 연구 수행 방법 ····· 7

제2장 현황 및 사례 분석

2.1. 국립중앙도서관 현황 분석 ····· 9
2.1.1. 디지털자원 및 보존 현황 분석 ····· 10
2.1.2. 국립중앙도서관 시스템 분석 ····· 17
2.2. 국내외 사례조사 및 분석 ····· 22
2.2.1. 해외 사례 분석 ····· 22
2.2.2. 국내 사례 분석 ····· 37
2.3. 시사점 ····· 42

제3장 보존 포맷 및 인프라 구축방안과 AI 기술 적용방안

3.1. 포맷 레지스트리 기술 ····· 45
3.1.1. 장기보존전략 유형 ····· 45
3.1.2. 장기보존전략(에뮬레이션) 적용 사례 ····· 54
3.1.3. 국립중앙도서관 디지털자원 보존전략(안) ····· 66
3.2. 스토리지 및 백업 인프라 구성 방안 ····· 72
3.2.1. 파일저장 스토리지 구성 방안 ····· 72
3.2.2. 백업 인프라 구성 방안 ····· 120
3.3. 디지털자원 보존을 위한 인공지능 기술 적용방안 ····· 127
3.3.1. 생성형 AI 기술 ····· 127

3.3.2. 디지털자원 인공지능 서비스 요소 ····· 129
3.3.3. 디지털자원 관리 및 보존을 위한 인공지능 서비스 제안 ···· 134
3.3.4. 디지털자원 복원을 위한 인공지능 서비스 제안 ····· 149

제4장 메타데이터 영역 기술동향 분석 및 설계

4.1. 현황분석 ····· 155
4.1.1. 선행 연구 조사 분석 ····· 155
4.1.2. 자료보존연구센터 현황 개선방향 ····· 165
4.2. 데이터 영역 전략 ····· 168
4.2.1. 전략 수립 프레임워크 ····· 168
4.2.2. 데이터 영역별 전략 ····· 169
4.3. 관련 기술동향 조사 분석 ····· 173
4.4. 데이터 목표모델 추진 방안 ····· 189

제5장 국립중앙도서관 디지털자원 보존전략 수립

5.1. 장기보존전략 ····· 213
5.2. 실행 우선순위 설정 ····· 214
5.3. 장기보존전략 흐름도 ····· 223
5.3.1. 실행 로드맵 수립 ····· 223

제6장 결론 및 제언

6.1. 결론 ····· 227
6.2. 제언 ····· 230

참 고 문 헌 ····· 233
[부록] 용어정리 ····· 237

표 차례

[표 1] 디지털자원의 종류 ····· 11
[표 2] 국립중앙도서관 장서현황 온라인자료 ····· 12
[표 3] 국립중앙도서관 장서구성 온라인자료 수집별 ····· 12
[표 4] 국립중앙도서관 장서구성 온라인자료 자료유형별 ····· 13
[표 5] 자료유형별 적용 규칙 ····· 16
[표 6] 통합자료관리시스템(KOLIS) 보존 PREMIS (시범)기능 데이터항목 ···· 18
[표 7] 통합자료관리시스템(KOLIS) 보존 영구보존 (시범)기능 데이터항목 ·· 18
[표 8] 해외 사례 비교표 ····· 36
[표 9] 국내 사례 비교표 ····· 41
[표 10] 포맷 표준 형식 ····· 48
[표 11] 디지털 장기보존전략 간 비교 분석 ····· 53
[표 12] Internet Archive 시스템 예약 필드 (System-Reserved) ····· 55
[표 13] Internet Archive 매직 필드 (Magic) ····· 57
[표 14] Internet Archive 매체유형별 메타데이터 항목 ····· 57
[표 15] 자료유형별 보존전략 선택 기준 및 적용 ····· 69
[표 16] 매체 유형별 스토리지 비교 ····· 77
[표 17] 디지털자원 보존 구조 1단계 Hot Tier ····· 92
[표 18] 디지털자원 보존 구조 2단계 Warm Tier ····· 94
[표 19] 디지털자원 보존 구조 3단계 Cold Tier ····· 96
[표 20] NDSA 레벨 모델의 요구사항 ····· 97
[표 21] NDSA 카테고리 기준 무결성 점검 근거 ····· 98
[표 22] NDSA 카테고리 기준 포맷 유효성 검사 근거 ····· 99
[표 23] NDSA 카테고리 기준 저장매체 상태 점검 근거 ····· 101
[표 24] NDSA 카테고리 기준 메타데이터 오류점검 근거 ····· 102
[표 25] 디지털자원 스토리지 점검 항목별 점검 주기 ····· 104
[표 26] 오브젝트 스토리지 기술요소 비교 ····· 115
[표 27] 오브젝트 스토리지 기술요소 선정 ····· 115
[표 28] SSD 스토리지 마이그레이션 주기 설정 ····· 118
[표 29] HDD 스토리지 마이그레이션 주기 설정 ····· 119
[표 30] 테이프 드라이브 스토리지 마이그레이션 주기 설정 ····· 119
[표 31] RAID 유형 ····· 121
[표 32] 국립중앙도서관 디지털자원 6단계 백업 아키텍처 ····· 126
[표 33] 디지털자원 보존을 위한 AI 서비스 요소 ····· 133

[표 34] 메타데이터 자동 생성을 위한 오픈소스 및 클라우드 ····· 139
[표 35] 데이터 영역 전략 수립 프레임워크 ····· 168
[표 36] 데이터 구조 영역 전략 ····· 171
[표 37] 메타데이터 표준 영역 전략 ····· 172
[표 38] DRI 관계 요소 사례 ····· 193
[표 39] 계층구조 설명 용어 ····· 195
[표 40] 계층구조 관계 유형 (Relationship Types) ····· 196
[표 41] Dcmitype:Collection에서 사용하는 Namespace ····· 198
[표 42] Dcmitype:Collection 입력 항목 ····· 198
[표 43] 테크니컬 메타데이터 CARARE 사례 ····· 202
[표 44] 포맷 레지스트리 정보 영역 구성 방안 ····· 205
[표 45] SIP 이관 필수 메타데이터 요소 ····· 211
[표 46] [전략 1] OAIS 기반 보존 책임 구조 확립 ····· 217
[표 47] [전략 2] 무결성 검증과 보존 이력 관리 체계 구축 ····· 218
[표 48] [전략 3] 보존용 메타데이터 및 포맷 정책의 제도화 ····· 219
[표 49] [전략 4] 다계층·분산 스토리지 확장 ····· 220
[표 50] [전략 5] 조직·거버넌스 재편 ····· 221
[표 51] [전략 6] AI 기반 고도화 전략 ····· 222

그림 차례

[그림 1] 연구 수행 방법 ····· 8
[그림 2] 부서별 자료 수집 및 디지털자원 관리 업무 현황 ····· 15
[그림 3] 미국의 Internet Archive(archive.org) ····· 59
[그림 4] 호주의 AusEaaSI(auseaasi.org) ····· 61
[그림 5] MAME 웹어셈블러 ····· 63
[그림 6] KVM-QEMU 모듈 ····· 64
[그림 7] HDD 스토리지 구조 ····· 73
[그림 8] SSD 스토리지 구조 ····· 75
[그림 9] LTO 테이프 스토리지 구조 ····· 76
[그림 10] 연결 방식에 따른 스토리지 유형 ····· 79
[그림 11] 오브젝트 스토리지의 구조 ····· 84
[그림 12] 데이터 접근 방식에 따른 스토리지 구분 ····· 86
[그림 13] 파일 스토리지와 오브젝트 스토리지 비교 ····· 87
[그림 14] 디지털자원 생애주기별 스토리지 기술 ····· 89
[그림 15] 무결성 검증(Fixity Check) 절차 ····· 108
[그림 16] 포맷 유효성 검사(File Format Validation) 절차 ····· 110
[그림 17] 저장매체 상태 점검(Storage Media Health Check) 절차 ····· 112
[그림 18] 메타데이터 오류점검(Metadata Validation) 절차 ····· 113
[그림 19] MinIO 아키텍처 ····· 116
[그림 20] MinIO + Kubernetes 기반 통합 스토리지 플랫폼 구조 ····· 117
[그림 21] 노드 내부 데이터 보호 RAID 종류 ····· 120
[그림 22] AI 개념도 ····· 128
[그림 23] AI 기반 디지털자원 심층 인덱싱 자동화 사례 ····· 130
[그림 24] 제목 자동 생성 단계 ····· 136
[그림 25] 선행 연구 조사 분석 ····· 155
[그림 26] 도서관 자료 생명주기 상의 국가문헌보존관의 위치 설정 모델링 ··· 156
[그림 27] 국가문헌보존관의 기능 구성 모델링 ····· 156
[그림 28] 국가문헌보존관 목표 시스템 구성도 ····· 158
[그림 29] 국가문헌보존관 목표 통합보존관리 시스템 구성도 ····· 159
[그림 30] 국가문헌보존관 목표 온라인 플랫폼 개념도 ····· 159
[그림 31] PREMIS 데이터모델 ver.2 ····· 173
[그림 32] PREMIS 데이터모델 ver.3 ····· 173
[그림 33] 기록과 매체의 관계 ····· 176

[그림 34] IFLA-LRM(Library Reference Model) ····· 181
[그림 35] DCCDAP의 데이터 모델 ····· 183
[그림 36] METS 문서 전체 스키마 구조 ····· 186
[그림 37] 엔티티 관계 모델 ····· 191
[그림 38] 파생기록물의 구조적 연계 및 기본 메타데이터 요소 ····· 192
[그림 39] 계층구조 개념도 ····· 195
[그림 40] 목표 데이터 구조 개념도 ····· 197
[그림 41] 코드 전환을 위한 매트릭스 사례 ····· 207
[그림 42] 국립중앙도서관 장기보존전략 ····· 213
[그림 43] 국립중앙도서관 디지털자원 보존 실행 로드맵 ····· 225

요 약 문

본 연구는 국립중앙도서관이 보유한 디지털자원의 장기보존 역량을 갖추기 위해서 체계적이고 지속 가능한 보존 인프라를 구축하고 보존자원의 품질을 유지하며, 보존 데이터 구조 및 메타데이터를 정의하기 위한 추진 전략과 구체적인 추진 방안을 제시하는 데 목적을 두었다. 이를 위해 국립중앙도서관의 디지털자원 현황과 보존 정책을 검토하고 KOLIS 및 OASIS(Online Archiving & Searching Internet Sources) 등 현행 시스템 구조를 분석하는 한편, 국가문헌보존관 운영계획 및 목표 시스템 아키텍처 분석을 병행하였다. 이러한 분석을 통해 국립중앙도서관의 역할과 기능에 적합하면서도, 내부 관련 시스템과 외부 서비스 시스템을 포함한 데이터 생명주기를 하나의 흐름으로 연계·운영할 수 있는 디지털자원 보존 시스템의 미래모델을 설정할 수 있었다. 이어서 이 미래모델의 실현을 위한 근거와 실행 가능성을 확보하기 위하여 국내외 주요기관 및 선행사례의 디지털 장기보존 적용 기술, 보존·복원 정책 및 운영사례를 분석하고 디지털 보존 분야의 시스템 인프라 및 데이터 영역의 최신 기술동향과 기술적용성을 검토함으로써 목표 기술 구성 전략과 방안을 제시하였다.

본 연구는 이러한 목표를 수행하기 위해 디지털 장기보존 기술을 포함한 시스템 인프라 영역, 인공지능 응용 영역, 데이터 구조 및 메타데이터 영역으로 구분하여 진행하였다. 먼저 시스템 인프라 영역에서는 보존 단계에서 필수적인 디지털 자원의 보존·재현·구현을 가능하게 하는 다계층 분산 구조의 스토리지 및 백업 인프라 구성, 무결성 검증(Fixity Check)을 중심으로 한 무결성 검증 기술 및 프로세스, 자료유형별 디지털 보존전략, 보존 이벤트의 PREMIS(Preservation Metadata: Implementation Strategies) 기반 체계적 관리, 장기보존 포맷과 기술적 처리에 필요한 포맷 레지스트리의 기술적용성 분석 및 전략적 추진 방안을 제시하였다. 이를 구체화하기 위하여 기술환경 변화 속에서도 접근성과 진본성을 유지하기 위한 단계별 접근전략으로서 디지털 장기보존전략을 매체전환(Refreshment), 마이그레이션(Migration), 포맷정규화(Normalization), 캡슐화(Encapsulation), 에뮬레이션(Emulation) 단계로 구분하고 보존 개입 수준, 보존 대상 관점, 진본성·원본성 유지 관점, 비용 및 기술 난이도 관점에서 비교 분석 결과를 제시하였다. 아울러 테크니컬 메타데이터 영역에서 포맷 레지스트리의 필요성과 구성 정보 영역 및 메타데이터 요소, 단계별 적용방안을 함께 제시하였다. 그 결과 매체전환(Refreshment), 마이그레이션(Migration), 포맷정규화(Normalization)은 디지털자원이 데이터 자체의

진본성과 유용성을 유지하면서 살아있는 상태로 운영될 수 있게 하는 보존전략으로서 사용 단계에서도 필수적 절차에 해당함을 확인하였다. 또한 장기보존전략 중 에뮬레이션(Emulation) 적용 사례로 Internet Archive(archive.org), 호주의 AusEaaSI (auseaasi.org), MAME 프로젝트(mamedev.org), KVM+QEMU 기반 EaaS, 영국 TNA 디지털 기록 재현(Replay) 프로젝트, 독일 DNB EaaS 프로젝트 등을 분석하여 참조하였으며, 국립중앙도서관 자료유형별 보존전략 수립을 위해 전자책・전자저널・학위논문(PDF, ePub 등), 디지털화 인쇄자료 이미지(TIFF, JPEG 등), 웹 자원(OASIS 수집 웹사이트, WARC 등), 시청각자료(영상, 음원), 데이터세트・연구데이터(CSV, XML, DB 등), 특수 디지털 객체(인터랙티브 콘텐츠, 구형 소프트웨어 등)별 보존전략의 선택 기준과 적용방안을 제시하였다.

이와 함께 안정적인 디지털자원 보존을 위해 보존 파일, 매체, 메타데이터 등에 대한 정기적 점검체계를 NDSA 레벨 모델의 요구사항을 기준으로 제시하고 무결성 검증(Fixity Check), 파일 포맷 유효성 점검, 저장매체 상태 점검, 메타데이터 오류점검 방안을 정리하여 디지털자원 스토리지의 단계별 점검에 적용하는 방안과 절차를 제시하였다. 디지털자원의 저장과 관리를 위한 파일저장 스토리지 구성 방안은 매체 유형에 따른 적용성, Local Storage・DAS・NAS・SAN 등 연결방식에 따른 적용성, 파일 스토리지(File storage)・블록 스토리지(Block storage)・오브젝트 스토리지(Object storage) 등 접근방식에 따른 적용성 검토를 통해 도출하였으며, 국립중앙도서관 적용방안으로 계층형 스토리지 구조 설계 필요성, 파일 스토리지 방식 적용방안, 그리고 디지털자원 생애주기에 적합한 단계별 스토리지 적용방안(입수단계 NAS, 관리단계 SAN/DAS, 보존단계 Object)을 제시하였다. 또한 백업 인프라 구성 방안으로 RAID 구성, Disk-to-Disk, Virtual-Tape-Library, LTO(Linear Tape-Open) Tape-backup, Off-site-backup, 원격지 백업 방안을 분석하여 국립중앙도서관에 적합한 6단계 다계층 백업 구조를 제시하였다. 이러한 기술 중심의 분석 결과를 바탕으로 시스템 인프라 영역의 추진 방안을 5개의 장기보존전략 추진과제로 재구성하여, 1) OAIS(Open Archival Information System, ISO 14721) 기반 보존 구조 확립, 2) 무결성 검증과 보존 이력 관리 체계 구축, 3) 다계층 분산 스토리지 확장, 4) 장기보존 포맷 정책의 제도화, 5) 보존 전담 조직 및 거버넌스 재편으로 제시하였다. 이 가운데 OAIS(Open Archival Information System) 기반 보존 구조는 장기보존 단계, 즉 영구보존 단계에서 추진하는 것이 적절하며, 나머지 과제는 사용과 보존 단계에서도 점진적으로 적용할 필요성이 높다는 점을 확인하였다.

인공지능 응용 영역에서는 생성형 인공지능 기술을 활용한 자동분류, 자동색인, 문서요약 및 메타데이터 자동 생성 등 데이터 처리 방안과, 포맷 자동

식별·분석 및 노후화 예측, 손상 위험 진단·탐지 및 자동식별, 중복·유사 디지털자원 식별, 문서·이미지·영상·음원 복원 등 복원 처리 방안을 제시하였다. 이는 시스템 인프라 영역에서 제시된 보존 프로세스의 자동화·고도화 요구와 연결되며, 보존 운영의 효율성과 품질 향상을 뒷받침하는 기술적 대안으로서 위치한다.

데이터 구조 및 메타데이터 영역에서는 국가문헌보존관 건립 계획 및 운영체계 관련 연구결과와 현황분석을 통해 국립중앙도서관의 디지털자원 보존 업무를 '사용과 보존이 공존하는 단계'와 '완전한 영구보존 단계'로 구분할 수 있음을 제시하였다. 완전한 영구보존 단계는 보존대상 자원이 변화가 없는 안정적 상태를 유지한다는 전제 아래 OAIS(Open Archival Information System) 기반 PREMIS(Preservation Metadata: Implementation Strategies) 표준의 사용을 전제로 하는 반면, 사용과 보존이 공존하는 단계는 실시간으로 변화하는 데이터 환경을 전제하며 관리와 서비스를 위해 영구보존과는 다른 환경을 필요로 한다는 점을 확인하였다. 또한 국립중앙도서관의 시청각기록물 디지털화 보존 연구와 PREMIS(Preservation Metadata: Implementation Strategies) 보존 데이터 시범 개발 등을 분석하여, KORMARC 계열 기술규칙을 활용한 자원의 구조적 분석 및 자원 간 상관관계 표현의 한계와, 실세계에서 발생하는 다양한 파생기록물 모델 적용의 필요성을 발견하고 국가기록원 등 동일 분야 선행연구를 참조하여 이를 데이터 구조와 메타데이터 요소로 해결하는 방안을 제시하였다. 이에 따라 사용과 보존이 공존하는 단계에서 요구되는 DC Collection Application Profile 기반 데이터 집합체의 계층적 구조, MARC21·MODS(Metadata Object Description Schema)와 호환성을 확보한 DC 메타데이터 Relation 요소를 통한 개체 유일 식별 및 파생기록물 연계 구조, DCMI 기반 컬렉션 메타데이터 요소와 자료유형별(Type of Resource) 테크니컬 메타데이터 상세화, 보존 시스템 외부에 존재하는 다양한 메타데이터 표준으로 구축된 아이템 메타데이터를 외부 스키마로 이용하는 방안, 포맷 레지스트리 메타데이터 구성 방안, 그리고 사용 단계 데이터를 영구보존 단계로 이관·전환하기 위한 OAIS(Open Archival Information System) SIP(Submission Information Package) 및 PREMIS (Preservation Metadata: Implementation Strategies) 환경과의 상호운용성 확보를 위해 사용 단계 데이터 아키텍처 설계 시 고려사항을 제시하였다.

종합하면 본 연구는 납본제도와 정보·지식자원의 서비스에 중점을 두는 국립중앙도서관의 특성을 전제로, 디지털자원 장기보존에 적합한 시스템 인프라와 데이터 아키텍처의 목표모델 및 단계별 추진 전략을 구체화하였다.

아울러 사용과 보존이 공존하는 단계에서의 시스템 구축과 데이터 아키텍처 설계에 필요한 기초 데이터로 활용될 수 있도록 연구 결과를 구조화함으로써, 국립중앙도서관의 디지털자원 보존 역량 강화를 위한 전략적 추진 기반을 제시하였다.

제1장 연구개요

1.1. 연구의 배경 및 필요성

1.1.1. 연구의 배경

21세기 사회의 거의 모든 정보는 디지털 형태로 생성되고 관리된다. 행정 문서, 과학 연구 데이터, 문화유산, SNS 콘텐츠에 이르기까지 인류의 지식과 활동의 대부분이 전자적 형태로 존재한다.

그러나 디지털 정보는 매체의 물리적 수명, 포맷의 급속한 변화, 소프트웨어 의존성 등으로 인해 장기적으로는 손실될 위험이 매우 높다. 이러한 상황에서 디지털자원의 장기보존은 단순한 기술적 선택이 아니라 국가의 지식 자산과 문화유산을 지속시키기 위한 필수 과제로 부상하고 있다.

이에 따라, 국가적으로 중요한 데이터가 체계적으로 보존되고 미래에 재구현되어 활용될 수 있도록 HW, SW 및 데이터 등의 정보자원을 체계적으로 보존할 방안 마련이 필요하다.

1.1.2. 디지털자원 장기보존의 필요성

디지털자원의 장기보존은 기술적·사회문화적 환경 변화 속에서 필수적인 정책 영역으로 자리 잡고 있다. 기술적 측면에서 가장 먼저 고려해야 할 요소는 저장매체의 불안정성이다. 현재 널리 사용되는 하드디스크나 광디스크 기반 저장매체는 일반적으로 3년에서 길게는 10년 수준의 제한된 수명을 지니고 있어 자연적 노후화 과정에서 데이터 손실이 발생할 가능성이 크다. 또한 운영체제, 하드웨어 아키텍처, 파일포맷 등 디지털 생태계를 구성하는 기술적 요소들이 빠르게 진화함에 따라 과거 환경에서 생성된 데이터가 최신 시스템에서 더 이상 접근되지 않는 기술적 단절 현상이 발생할 위험 또한 상존한다. 여기에 데이터 생산량이 기하급수적으로 증가하는 ‘데이터 폭증’ 환경이 더해지면서 대규모 자원을 지속적으로 관리·검증·보존하기 위해서는 자동화된 업무절차 고효율적인 보존 기술의 도입 필요성이 더욱 증대되고 있다.

사회·문화적 측면에서도 디지털 장기보존의 필요성은 계속해서 강화되고 있다. 오늘날 웹사이트, 디지털 예술 콘텐츠, 사회관계망서비스(SNS) 기록 등은 현대 사회의 문화를 구성하는 핵심적 요소로 간주되며 이들 자원의 체계적 보존은 미래

세대에게 전달해야 할 디지털 문화유산을 축적한다는 점에서 중요한 의미를 갖는다. 또한 행정 자료, 법률 기록, 정책 문서 등 공공 기록은 민주적 정당성과 투명성을 뒷받침하는 근거로 기능하기 때문에 이를 장기간 안정적으로 보존하는 것은 사회적 책임성 확보의 기반이 된다. 연구 영역에서도 데이터 보존의 중요성은 점차 강조되고 있다. 연구데이터는 과학적 재현성 검증과 후속 연구의 연속성을 보장하는 핵심 자료이기 때문에 장기보존체계의 부재는 학문적 진전 자체를 저해하는 요인이 될 수 있다. 결국 디지털 장기보존은 기술적 지속 가능성뿐 아니라 사회·문화적 가치의 전승과 공적 책임성을 실현하기 위한 필수 인프라라고 할 수 있다.

1.2. 연구의 범위

본 연구는 국립중앙도서관이 보유한 디지털자원의 장기보존전략을 체계적으로 수립하고 이를 뒷받침할 인프라 구축 방안을 마련하기 위한 목적으로 수행되었다. 이를 위해 연구 범위는 크게 다섯 가지 영역으로 구분된다.

첫째, 국내외 디지털자원의 최신 동향과 장기보존 시스템의 현황을 조사·분석하는 작업을 수행하였다. 여기에는 주요 국가 및 기관의 최신 보존정책과 제도적 기반에 대한 사례 분석을 포함하며, 이를 통해 국내 환경에 적합한 최적의 정책 요소를 도출하고자 하였다. 동시에 해외 도서관·기록관이 구축한 장기보존 시스템의 구조와 스토리지 인프라, 운영 방식 등을 면밀히 검토함으로써 국립중앙도서관의 현행 보존 시스템 구조와 대비·분석하였다. 이러한 분석을 바탕으로 향후 구축되어야 할 보존 시스템의 필수 기능 모듈과 모듈 간 인터페이스의 기본 설계 방향을 도출하였다.

둘째, 포맷 레지스트리 및 보존 메타데이터 저장체계의 기초안을 마련하기 위한 연구를 수행하였다. 이를 위해 국제적으로 활용되고 있는 포맷 레지스트리 기술을 조사하여 국내 도입 가능성과 적용 방식에 대해 검토하였으며, 국립중앙도서관의 조직·기술 환경에 적합한 포맷 등록·관리 체계의 구조를 설계하였다.

셋째, 장기보존을 위한 스토리지 및 백업 인프라 구성 방안을 마련하였다. 장기저장에 적합한 스토리지 유형을 기술적 안정성, 경제성, 확장성 측면에서 구분·비교하였으며, 저장매체의 교체 주기 설정 기준과 포맷 무결성을 지속적으로 검증할 수 있는 체계적인 진단·검증 모델을 설계하였다. 이를 통해 디지털자원의 장기 안정성을 보장하는 기술적 토대를 강화하고자 하였다.

넷째, 국립중앙도서관이 보유한 디지털자원의 특성을 세분화하여 유형별 보존 전략을 수립하였다. 콘텐츠의 특성(소프트웨어 기반 콘텐츠, 특정 파일 형식, 플랫폼 의존형 데이터 등)을 고려하여 에뮬레이션 기반 전략과 마이그레이션 기반 전략을 구분하여 제안하였다. 또한 콘텐츠 유형, 실행환경, 포맷 구조 등에 따른 세부 적용 기준과 전환 시나리오를 설계하였으며, 각 전략의 장단점과 기술적 지속가능성을 종합적으로 분석하였다. 더불어 국립중앙도서관이 운영 중인 기존 정보시스템과의 연계성을 확보하기 위한 통합 운영 방안도 함께 도출하였다.

다섯째, 디지털 보존 분야에서 활용도가 증가하고 있는 AI 기반 기술의 적용 가능성을 검토하였다. 인공지능을 활용한 포맷 식별 자동화, 보존 우선순위의 자동 분류, 파일 상태 진단 등의 기능은 대규모 자원 환경에서 높은 잠재적 효율성을 갖고 있다. 이에 따라 AI 기반 자동 태깅, 의미 기반 색인화 기술을 포함하여 대용량 보존 대상의 관리 효율성을 획기적으로 향상시킬 수 있는 방안을 제안하였다.

마지막으로 국립중앙도서관의 보존 시스템의 특징을 사용과 보존이 병행하는 시스템으로 정의하고 이에 합당하게 계층적 구조, 유일식별체계, 파생기록물 연계구조, 영역별 메타데이터 표준 지침을 제시하고 영구보존 단계로 진행될 경우에 OAIS(Open Archival Information System) 모델 및 PREMIS(Preservation Metadata: Implementation Strategies) 메타데이터 표준과 상호운용성을 확보할 수 있는 방안을 제시하였다.

이와 같은 일련의 연구 범위는 국립중앙도서관이 디지털 시대에 요구되는 장기보존 역량을 확립하고 체계적·지속 가능한 보존 인프라를 구축하기 위한 전략적 기반을 마련하는 데 목적을 두고 수행되었다.

1.3. 연구 수행 방법

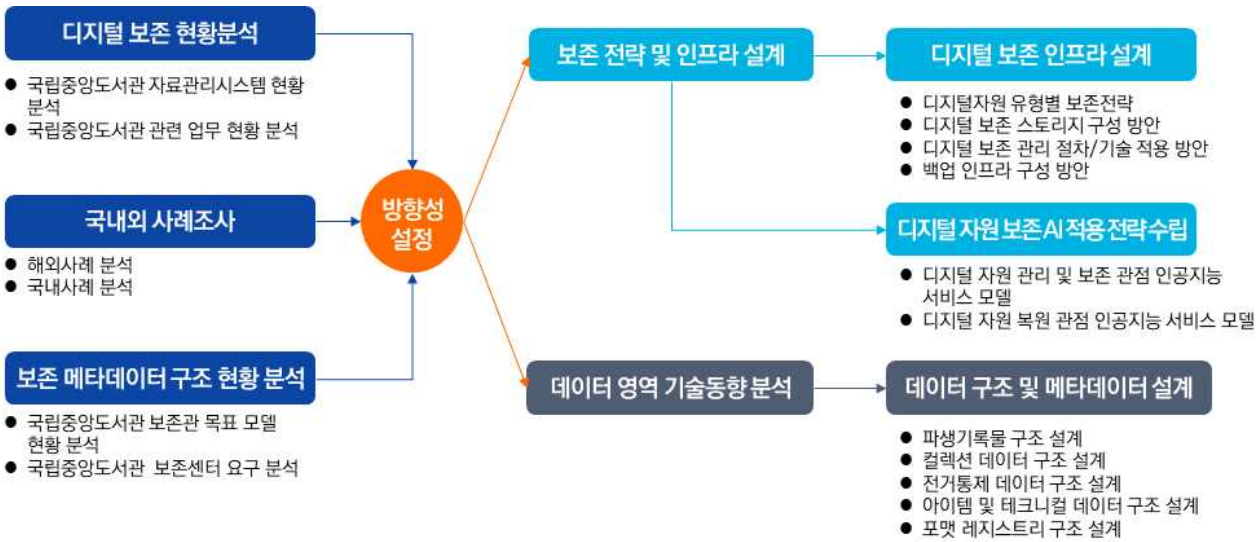
본 연구의 범위는 디지털자원의 장기보존전략 및 인프라 구축 방안을 제시하는 것이며 보존체계 구축을 위한 단계적·연계적 분석과 설계로 구성하였다.

먼저 연구는 디지털 보존 현황분석, 국내외 사례조사, 보존 메타데이터 구조 현황 분석이라는 세 가지 기초 분석 영역을 수행하는 데에서 출발한다. 디지털 보존 현황분석은 현재 운영 중인 보존 체계의 기능, 구조, 문제점을 규명하기 위한 기초 진단 단계이며, 국내외 사례조사는 주요 기관들의 보존 정책과 운영 방식을 확인함으로써 비교 기준을 마련하는 과정이다. 보존 메타데이터 구조 현황 분석은 국립중앙도서관의 보존관과 관련하여 보존용 메타데이터 모델, 적용 구조, 요구 요소 등을 검토함으로써 장기보존 체계의 정보 구조적 기반을 점검하는 단계이다.

이들 세 가지 분석 결과는 연구의 중심축인 디지털자원 보존의 방향성 설정으로 수렴된다. 방향성 설정 단계는 앞선 분석에서 드러난 핵심 과제와 개선 필요성을 근거로 하여 향후 구축되어야 할 보존체계의 기본 원칙과 설계 방향을 정립하는 과정이다. 이러한 방향성은 연구의 후반부에 제시될 구체적 설계 단계로 연결되며, 연구 전체의 구조적 기준을 제공한다.

설정된 방향성은 두 개의 주요 연구 축으로 확장된다. 첫 번째 축은 보존전략 및 인프라 설계이다. 이 축은 장기보존을 위한 구조적·기술적 기반을 마련하는 데 초점을 두며, 여기에는 디지털 보존 인프라 설계가 포함된다. 동시에, 보존 환경에서의 기술적 진전 가능성을 고려하여 디지털자원 보존 AI 적용 전략 수립도 병행된다. 그러므로 보존 인프라와 AI 전략은 동일한 방향성 아래에서 상호 보완적으로 설계된다.

두 번째 연구 축은 데이터 영역 기술동향 분석으로 시작된다. 데이터 기술의 변화와 요구사항을 분석함으로써 사용과 보존이 공존하는 환경에서 필요한 기술적 고려 요소를 명확히 한다. 이러한 분석 결과는 데이터 구조 및 메타데이터 설계로 이어지며, 이는 사용과 보존이 공존하는 단계에서 이후 영구보존 단계로 이관하는 단계까지 연속성을 가지고 실제로 관리될 데이터 구조의 형태와 메타데이터 구성 방식에 대한 구체적 설계를 완성하는 단계이다.



[그림 1] 연구 수행 방법

제2장 현황 및 사례 분석

2.1. 국립중앙도서관 현황 분석

❑ 분석 목적 및 필요성

본 연구는 국립중앙도서관이 보유·관리하고 있는 디지털자원의 장기보존 체계와 인프라 구축 방안을 마련하기 위해 수행되었다. 이를 위해 국립중앙도서관의 디지털자원 현황, 관련 정보시스템 구조, 업무 프로세스 및 담당자 요구사항을 다각도로 분석함으로써, 현행 체계의 문제점과 개선 필요 요소를 도출하는 데 목적이 있다.

특히 분석 결과는 향후 디지털자원의 지속적 보존을 위한 저장체계 구조, 메타데이터 관리 방안, 자동화 기반의 보존 인프라 설계, 그리고 이를 구현하기 위한 기술·정책적 보존 전략 수립의 기초 자료로 활용된다. 그러므로 본 연구는 현황 분석을 통해 국립중앙도서관 디지털 보존체계의 미래지향적 설계 원칙을 제시하는 데 그 필요성을 둔다.

❑ 분석 대상 및 내용

분석은 크게 디지털자원 분석, 시스템 분석, 데이터 분석으로 구분하여 수행하였다.

첫째, 디지털자원 분석에서는 국립중앙도서관이 생산·수집·관리하고 있는 디지털자원의 정의, 유형, 특성 및 보유 현황을 파악하였다. 자료의 종류를 기반으로 보존 수준, 서비스 방식, 포맷 특성, 기술적 제약 등을 종합적으로 검토하였다.

둘째, 시스템 분석에서는 국립중앙도서관의 통합자료관리시스템(KOLIS)과 웹자원 아카이브 시스템(OASIS)을 중심으로 디지털자원이 실제로 관리·저장되고 있는 구조를 분석하였다. 또한 각 시스템이 제공하는 보존 관련 기능의 현황과 함께 장기보존 관점에서의 한계·이슈를 도출하였다.

셋째, 데이터 분석 대상 및 내용은 4장에서 별도로 기술하였다.

2.1.1. 디지털자원 및 보존 현황 분석

❑ 디지털 보존의 정의[1)]

디지털 보존은 필요한 만큼 디지털자료에 대한 지속적인 액세스를 보장하기 위해 필요한 일련의 관리 활동을 나타낸다. 따라서 디지털 보존은 그 디지털 자원의 보존 기간에 따라 그 목표와 기능을 달리한다.

- ㅇ단기보존: 사용이 예측되는 일정 기간 동안 디지털자료에 접근할 수 있지만 예측 가능한 미래를 넘어서지 않으며, 기술의 변화로 인해 접근이 불가능해질 때까지 접근할 수 있다.
- ㅇ중기보존: 지정된 기간 동안 기술 변화와 상관없이 디지털자료에 계속 접근할 수 있지만 무기한은 아니다.
- ㅇ장기보존: 디지털자료 또는 디지털자료에 포함된 정보에 계속 접근할 수 있다.

디지털 보존이란 미래의 후손들이 용이하게 볼 수 있고 이해할 수 있는 방식으로 디지털자원을 보존하는 것을 말한다.

❑ 디지털 보존의 필요성

디지털자원의 휘발성 및 기술의존성과 같은 심각한 약점과 디지털 환경이 가지는 특성으로 인하여 디지털자원의 유지관리는 인쇄자료의 유지관리보다는 훨씬 어렵고 비용도 많이 든다. 그러나 이러한 문제점과 어려운 점을 적절히 대응하고 해결해야 현 시대에서 생산되고 활용되는 디지털자원을 영구적으로 보존할 수 있다.

디지털 보존은 디지털자원의 무결성을 상실하지 않고 저장매체의 손상 및 퇴화, 기술적 퇴화의 위협으로부터 디지털정보를 안전하게 보호하여 유용한 형태로 존속하게 하는 일련의 관리 활동이다. 따라서 미래 세대에 우리의 지적 유산을 계승하기 위해서는 지속적이고 체계적인 디지털 보존이 수행되어야 하고 디지털콘텐츠를 보유하고 있는 모든 기관에서는 디지털보존의 필요성과 중요성을 인지해야 한다.

1) 국립중앙도서관 디지털자원 보존전략 수립 연구, 국립중앙도서관, 2018.12.

❑ 디지털 보존 대상의 범위

디지털자원이란 정보를 디지털로 인코드하고 컴퓨터를 통해 접속·인지하도록 하는 방식으로 제작하여, 일반 공중의 이용에 제공할 목적으로 배포된 자료를 말한다. 디지털자원의 종류는 매우 다양하지만 대체로 오프라인 자료와 온라인 자료로 나눌 수 있으며, 온라인 자료는 다시 정적(Static) 온라인 자료와 동적(dynamic) 온라인 자료로 나눌 수 있다.

오프라인 자료		• 오프라인 자료는 손으로 잡을 수 있는 물리적 형태가 있는 매체에 고정된 것 • 패키지형 전자출판물 • CD-ROM, 컴퓨터디스켓, DVD 등
온라인 자료		• 인터넷, PC통신 등 네트워크를 통해 생산되고 이용되는 것 • 손으로 잡을 수 있는 물리적 형태가 없음 • 네트워크형 전자출판물 • 인터넷 상에서 제공되는 각종 웹정보, 데이터베이스 등
	정적 자료	• 파일의 내용이 일정하게 고정된 것
	동적 자료	• 이용자와의 상호작용 속에 지속적으로 변화하는 것

[표 1] 디지털자원의 종류

❑ 「도서관법」에서 본 디지털자원 범위

국내에서는 「도서관법」 제21조(도서관자료의 납본)[2]에 따라 납본의 대상을 국제표준자료번호를 부여받은 온라인 자료와 국가, 지방자치단체 및 그 밖에 대통령령으로 정하는 공공기관이 발간하는 공공간행물의 디지털파일로 정의하고 있다. 또한 제22조(온라인 자료의 수집)에 따라 우리나라에서 서비스되는 온라인 자료 중에서 보존가치가 높은 자료에 대하여 수집·보존하여야 한다는 조항을 명시하였다.

제21조(도서관자료의 납본)
① 누구든지 도서관자료(온라인 자료는 제외한다. 다만, 온라인 자료 중 제23조에 따라 국제표준자료번호를 부여받은 온라인 자료는 포함한다. 이하 이 조에서 같다)를 발행 또는 제작한 경우 그 발행일 또는 제작일부터 30일 이내에 그 도서관자료를 국립중앙도서관에 납본하여야 한다. 수정증보판인 경우에도 또한 같다. <개정 2023. 8. 8.>
② 국가, 지방자치단체 및 그 밖에 대통령령으로 정하는 공공기관이 제1항에 따라 도서관자료를 국립중앙도서관에 납본하는 경우에는 대통령령으로 정하는 바에 따라 디지털파일 형태로도 납본하여야 한다.

제22조(온라인 자료의 수집)
① 국립중앙도서관은 대한민국에서 서비스되는 온라인 자료 중에서 보존가치가 높은 온라인 자료를 선정하여 수집·보존하여야 한다.

2) 도서관법(법률 제21090호, 시행 2025. 11. 11.)

□ 국립중앙도서관 온라인 자료 현황[3)]

국립중앙도서관의 디지털자원 입수 유형은 크게 납본(단행・연속)과 자체수집(웹사이트, 웹자료)으로 구분되어 운영되고 있다. 납본자료는 발행자가 디지털파일을 국립중앙도서관 ISBN・ISSN・I납본시스템에 직접 업로드하거나 저장매체에 저장하여 송부하는 방법으로 입수되며, 웹사이트는 오아시스(Online Archiving & Searching Internet Sources, OASIS) 시스템을 기반으로 한 크롤링 방식으로 수집된다.

국립중앙도서관은 「도서관법」 제21조(도서관자료의 납본) 및 같은 법 시행령에 따라 ISBN을 부여받는 온라인 자료, 공공간행물의 디지털파일을 납본 대상으로 수집하고 있다. 2024년 기준 납본을 통해 수집한 온라인 자료는 전자저널, 전자책, 공공간행물 등으로 구성되어 있으며, 비교적 표준화된 파일 포맷과 메타데이터 구조를 갖춘 자료가 다수를 차지하고 있다.

한편, 국립중앙도서관은 「도서관법」 제22조(온라인 자료의 수집)에 따라 납본 대상 외 온라인 자료에 대해서도 자체 수집 체계를 통해 적극적으로 확보하고 있다. 자체 수집 자료는 웹사이트, 이미지, 영상, 음성 등 비정형・멀티미디어 자료 비중이 높고, 생성・변경・소멸 주기가 짧은 특성을 보인다. 이러한 자체 수집 온라인 자료는 납본 자료에 비해 자료의 기술적 특성과 관리 방식이 다양하여, 장기보존 관점에서 보다 복합적인 관리 전략이 요구된다.

국립중앙도서관의 온라인 자료 장서 현황을 지역별로 살펴보면 전체 온라인 자료 20,478,575건 가운데 국내 자료가 20,417,099건, 국외 자료는 61,476건으로 나타난다. 온라인 자료는 국내 자료를 중심으로 구축・관리되고 있다.

(단위: 건, 2024. 12. 31. 기준)

구분	국내	국외	계
장서	20,417,099	61,476	20,478,575

[표 2] 국립중앙도서관 장서현황 온라인자료

수집 유형별로는 자체생산 자료가 11,272,953건, 자체수집 자료가 4,800,318건(23.4%), 기증 자료가 3,174,066건(15.5%), 납본 자료가 1,231,238건(6.0%) 순으로 나타났다. 자체생산과 자체수집의 합이 전체의 약 78%에 달하고 있다.

(단위: 건, 2024. 12. 31. 기준)

구분	납본	기증	자체생산	자체수집	계
수량	1,231,238	3,174,066	11,272,953	4,800,318	20,478,575
비율(%)	6.0	15.5	55.0	23.4	100

[표 3] 국립중앙도서관 장서구성 온라인자료 수집별

3) 2024 국립중앙도서관 연보, 국립중앙도서관

자료 유형별로는 디지털화 자료가 11,162,377건으로 54.5%를 차지하여 가장 큰 비중을 보였으며, 전자저널은 5,594,310건, 전자책은 1,555,101건으로 나타나고 있다. 전자신문은 6,075건, 영상 70,652건과 이미지 244,499건, 음성·음향 자료 593,043건 등 다양한 비텍스트 자료도 다수 존재하는 것으로 보여진다. 또한, 웹사이트 자료는 1,117,389건으로 전체의 5.5%를 차지하고 있으며, 해외기록물(43,342건, 0.2%)과 국외고문헌(18,134건, 0.1%), 장애인대체자료(73,653건)는 각각 소규모이지만 특수 유형 자료로 구성되어 있다.

(단위: 건, 2024. 12. 31. 기준)

구분	전자책	전자저널	전자신문	음성 · 음향	영상	이미지
수량	1,555,101	5,594,310	6,075	593,043	70,652	244,499
비율(%)	7.6	27.3	0.0	2.9	0.3	1.2
구분	**웹사이트**	**디지털화 자료**	**해외기록물**	**국외고문헌**	**장애인 대체자료**	**계**
수량	1,117,389	11,162,377	43,342	18,134	73,653	20,478,575
비율(%)	5.5	54.5	0.2	0.1	0.4	100

[표 4] 국립중앙도서관 장서구성 온라인자료 자료유형별

국립중앙도서관의 온라인 자료 장서는 국내 중심, 자체생산·자체수집 중심, 디지털화 및 전자출판물 중심의 구조를 형성하고 있으며, 국립중앙도서관이 지식자원의 디지털 축적과 서비스 제공을 하고자 하는 점을 보여주고 있다. 동시에 이러한 장서 구성은 향후 디지털 장기보존 전략 수립 시, 자체생산 디지털 자료·다양한 유형의 자료를 안정적으로 보존·관리하기 위한 전담 보존 체계의 필요성이 있다.

포맷 관리 측면에서는 발행자가 제공한 원본 파일 형식을 최대한 그대로 유지하는 방식이 기본 원칙으로 설정되어 있으며, 디지털 파일 납본 시에도 포맷 변환은 최소화하는 방향으로 운영되고 있다. 다만 저용량 변환이나 뷰어 미지원 형식에 한해서는 예외적으로 변환이 이루어지고 있으며, 텍스트·이미지·영상·음성을 통합적으로 확인할 수 있는 통합 뷰어 환경이 제공되고 있다.

저작권 및 이용 동의 절차와 관련해서는 납본 및 일반 수집 자료의 경우 이용동의서를 확보하고 있으나, 웹 수집 자료는 개별 동의 확보가 현실적으로 어렵기 때문에 관내 이용을 중심으로 제한적인 서비스가 이루어지고 있다. 또한 일부 자료에는 DRM이 적용되어, 보존 및 서비스 과정에서 추가적인 제약이 발생하는 것으로 나타났다.

❑ 부서별 자료 수집 및 디지털자원 관리 업무 현황

「도서관법」[4]이 정하는 자료의 납본 및 온라인 자료 수집 근거에 따라 국립중앙도서관에서는 자료 수집 및 디지털자원 관리 업무를 수행하고 있다. 국립중앙도서관 조직은 기획연수부, 지식정보관리부 그리고 자료보존연구센터로 구성되어 있다. 이 중 디지털자원의 직접적인 납본·수집·관리 기능은 주로 지식정보관리부 소속 부서들에서 수행하고 기획연수부는 디지털자원 정책 기획과 시스템 구축을 담당하는 구조를 형성하고 있으며, 자료보존연구센터는 디지털자원의 보존을 담당하고 있다.

디지털자원과 직접적으로 관련된 부서는 기획연수부의 디지털정보기획과, 지식정보관리부의 온라인자료과, 지식정보서비스과 그리고 자료보존연구센터 등이다.

디지털정보기획과는 국립중앙도서관의 디지털자원 정책과 플랫폼 구축을 총괄하는 핵심 부서로, 디지털자원 공유, OAK(Open Access Korea), 소장자료 디지털화, 문화예술자료 디지털화, 통합자료관리시스템, 국가시지식유통플랫폼 구축 등을 담당하고 있다. 이는 온라인 자료의 생산 환경을 조성하고 디지털 콘텐츠의 국가적 유통 기반을 구축하는 기능을 수행하는 부서로서, 실질적인 온라인 자료 생산 인프라의 기획 기능을 담당한다고 볼 수 있다.

온라인자료과는 온라인 자료 납본의 핵심 부서로서, 온라인 자료(단행, 연속)의 납본 및 수집, 웹자원(웹사이트, 웹자료) 수집 및 웹자원 아카이브 운영, 온라인 자료 서지 구축(단행, 연속, 멀티미디어, 해외기록물 등), 웹콘텐츠 UCI 표준식별체계 발급 및 운영 관리, 해외 한국 관련 자료 수집 업무 등을 담당하고 있다. 이 부서는 국내에서 서비스 중인 웹 기반 콘텐츠와 디지털 원본 자료의 수집 등록 정리를 담당함으로써 실질적인 웹아카이빙과 디지털콘텐츠 축적이 집중되는 부서이다.

지식정보서비스과는 디지털도서관 관리 및 운영, 실감콘텐츠 제작 및 서비스 등을 담당하며, 온라인 자료를 이용자에게 실제로 제공하는 부서이다.

자료보존연구센터는 수집된 디지털자료의 장기보존 기술 연구와 보존 인프라 고도화 기능을 담당하는 연구 조직으로 향후 온라인 자료를 포함한 디지털자원의 장기보존 전략 수립과 기술 표준 적용의 핵심 거점으로 기능하는 구조를 갖는다.

---

4) 도서관법(법률 제21090호, 시행 2025. 11. 11.)



[그림 2] 부서별 자료 수집 및 디지털자원 관리 업무 현황

정리하면 국립중앙도서관에서는 문화예술자료, 소장자료(일반도서, 연속간행물, 공공간행물, 학위논문, 비도서, 외국자료 등). 웹자원(웹사이트, 웹자료), 해외 한국 관련 자료, 실감콘텐츠, 고문헌 자료 등 다양한 유형의 디지털자원을 수집하고 있으며, 이러한 자원들은 국립중앙도서관이 보존해야 할 보존 대상으로 보여진다.

운영 체계 측면에서는 수집, 제작, 보존 업무가 부서별로 분절되어 수행되고 있어, 디지털자원의 전주기적 통합 관리가 어렵다는 문제가 지속적으로 제기되고 있다. 이러한 업무 분절 구조는 입수-정제-등록-보존-서비스로 이어지는 데이터 흐름을 단절시키는 요인으로 작용하고 있으며, 결과적으로 장기보존 체계 구축의 가장 큰 구조적 장애 요인 중 하나로 분석된다.

이에 대해 자료보존연구센터를 중심으로 한 통합 거버넌스 구축의 필요성이 인식되고 있으며, 수집 지침, 포맷 정책, 수집 주기(Depth 및 주기 설정)에 대한 표준화, 그리고 부서 간 데이터 흐름에 대한 명확한 역할 정의가 필요하다.

❑ 자료 유형별 데이터 적용 규칙[5)]

국립중앙도서관 온라인자료과에서는 자료 유형별로 다음과 같은 규칙을 적용하여 데이터를 구성할 수 있도록 지침을 내리고 있다.

적용 범위	**단행 자료**	• 전자책, 공공간행물(단행), 학위논문, 동영상, 사운드(음악, 오디오북 등), 이미지, 웹툰, 해외기록물, 웹사이트 등
	연속 자료	• 전자저널(학술지), 잡지, 전자신문, 공공간행물(연속), 학술기사 등
적용형식 (목록규칙, 기술규칙, 분류법)		• 온라인 메타데이터 포맷 MODS(Metadata Object Description Schema) • 국립중앙도서관 온라인자료 정리 지침 • 한국목록규칙 제4판(KCR 4) 및 보유편(전자책 · 전자저널 기술규칙) • 한국문헌자동화목록형식(KORMARC) • 한국십진분류법(KDC) 제6판 및 듀이십진분류법(DDC) 23판(웹듀이 적용)
메타데이터 입력 항목		• 표제, 표제관련정보, 권차, 저자, 발행지, 발행처, 발행일, 식별기호(ISBN, ISSN 등), 언어, 형태기술정보, 요약정보, 목차정보, 이용대상자, 주기사항, 주제명, 분류기호, 연관정보 등

[표 5] 자료유형별 적용 규칙

5) 2024 국립중앙도서관 온라인자료 정리 지침(2025.04 개정)

2.1.2. 국립중앙도서관 시스템 분석

본 파트에서는 국립중앙도서관이 보유하고 있는 디지털자원과 이를 관리하는 내부 시스템의 현황을 분석한다. 분석 대상은 디지털자원의 종류, 메타데이터 구조, 수집・서비스 방식, 그리고 통합자료관리시스템(이하 KOLIS로 지칭함)・웹자원 아카이브(이하 OASIS로 지칭) 등 핵심 시스템의 기능적 구성으로 설정하였다.

❑ 통합자료관리시스템(KOLIS)[6)]

KOLIS는 국립중앙도서관이 수행하는 국가 납본・수집・등록・관리・서비스 제공 업무 전반을 통합적으로 지원하는 업무시스템이다. KOLIS는 단행・연속 간행물의 납본 접수, 미납본 관리 등 「도서관법」에 근거한 법정 업무 수행을 지원한다. 또한, 온라인자료, 전자책・전자저널・디지털화 자료에 대해 수집 → 검수 → 등록 → 배가 → 서비스까지 이어지는 디지털 콘텐츠의 수집・등록・서비스 관리를 수행한다.

KOLIS는 장기적으로 보존이 필요한 대상 자료를 선정하고 보존 정책에 따른 관리 기반을 마련하고자 하였다. KOLIS에서는 보존 업무와 관련하여 영구보존 대상 지정, PREMIS(Preservation Metadata: Implementation Strategies) 메타데이터 입력・수정・현황 관리를 할 수 있는 기능이 시범적으로 구현되어 있다.

먼저 영구보존 관리 기능은 KOLIS를 통해 단순 수집 자료와 장기보존 대상 자료를 분리하고 보존 정책에 따른 관리 기반을 마련하고 있다.

또한 PREMIS(Preservation Metadata: Implementation Strategies) 기반 관리 기능을 통해 대상콘텐츠 관리, 수정콘텐츠 관리, 완료콘텐츠 관리의 단계적 절차가 운영된다. 이는 디지털 객체가 최초 보존 대상으로 지정된 이후, 보존 과정에서 발생하는 수정 이력과 최종 보존 완료 상태까지의 전 과정을 PREMIS(Preservation Metadata: Implementation Strategies) 메타데이터로 기록・관리하는 구조를 보여주고 있다.

6) KOLISIII 온라인 – 사용자지침서, 국립중앙도서관, 2022.

보존 기능	구분	데이터 항목
대상콘텐츠 관리	조회 조건	자료유형, 등록일자, 수집유형, 항목, PREMIS 상태, 등록구분, 수입년도, 원부번호
	조회 항목	NO, 콘텐츠 제어번호, 콘텐츠유형, 장르, 서비스 상태, 표제, 수집유형, 저자, 발행처, 발행일, 원문, PREMIS 상태, 등록일, 수정일, PREMIS 등록일, PREMIS 수정일
	기능	[PREMIS 반입] : 해당되는 제어번호에 PREMIS를 반입 [PREMIS 등록] : 선택된 자료의 PREMIS를 등록 [PREMIS 수정] : 선택된 자료의 PREMIS를 수정 [PREMIS 삭제] : 선택된 자료의 PREMIS를 삭제
보존콘텐츠 관리	조회 조건	자료유형, 등록일자, 수집유형, 항목, PREMIS 상태
	조회 항목	NO, 콘텐츠 제어번호, 콘텐츠유형, 장르, 서비스 상태, 표제, 수집유형, 저자, 발행처, 발행일, 원문, PREMIS 상태, 등록일, 수정일, PREMIS 등록일, PREMIS 수정일
	기능	[PREMIS 반입] : 해당되는 제어번호에 PREMIS를 반입 [PREMIS 등록] : 선택된 자료의 PREMIS를 등록 [PREMIS 수정] : 선택된 자료의 PREMIS를 수정 [PREMIS 삭제] : 선택된 자료의 PREMIS를 삭제 [중요속성추가] : 파일을 통해 중요속성을 일괄 추가하나 일괄 추가하거나 일괄반입 [완료요청] : 선택된 자료를 완료 처리
완료콘텐츠 관리	조회 조건	자료유형, 등록일자, 수집유형, 항목, PREMIS 상태
	조회 항목	NO, 콘텐츠 제어번호, 콘텐츠유형, 장르, 서비스 상태, 표제, 수집유형, 저자, 발행처, 발행일, 원문, PREMIS 상태, 등록일, 수정일, PREMIS 등록일, PREMIS 수정일
	기능	[수정요청] : 선택된 자료를 수정요청

[표 6] 통합자료관리시스템(KOLIS) 보존 PREMIS (시범)기능 데이터항목

보존 기능	구분	데이터 항목
대상콘텐츠 관리	조회 조건	자료유형, 등록일자, 수집유형, 보존여부, 항목, 콘텐츠 상태, 등록구분, 수입년도, 원부번호
	조회 항목	NO, 콘텐츠 제어번호, 콘텐츠유형, 장르, 서비스 상태, 표제, 수집유형, 저자, 발행처, 발행일, 원문, 보존여부, 등록일, 수정일
	기능	[보존이관] : 선정된 자료를 보존이관 처리
보존콘텐츠 관리	조회 조건	자료유형, 등록일자, 수집유형, 보존여부, 보존상태
	조회 항목	NO, 콘텐츠 제어번호, 콘텐츠유형, 장르, 서비스 상태, 표제, 수집유형, 저자, 발행처, 발행일, 원문, 보존여부, 요청일, 완료일
	기능	[삭제] : 선정된 자료를 삭제 [다운로드] : 선정된 자료를 다운로드 [METS보기] : 선정된 자료의 메타데이터 보기

[표 7] 통합자료관리시스템(KOLIS) 보존 영구보존 (시범)기능 데이터항목

❑ 메타데이터 체계

메타데이터 표준 체계는 현재 MARC와 MODS(Metadata Object Description Schema)를 병행하는 통합 구조로 운영되고 있으며, 장기적으로는 2030년을 목표로 BIBFRAME 기반 체계로의 전환이 추진되고 있다. 디지털자원의 메타데이터는 MODS(Metadata Object Description Schema) 형식으로 통일하여 관리되고 있으며, KOLIS 등록 시에는 고유한 원부번호가 부여된다. 그러나 KOLIS 시스템 자체가 본질적으로 서비스 중심 구조로 설계되어 있어 장기보존 관점의 메타데이터 관리 및 구조적 지원에는 한계가 있다.

디지털자료와 관련해서는 KOLIS 시스템 내부에 PREMIS(Preservation Metadata: Implementation Strategies) 메타데이터 80개 입력 항목이 설계되어 있으나 해당 항목들은 아직 실제 운영 과정에서 활용되지 못하고 있는 상황이다.[7] 이 점을 고려할 때 향후 해당 기능의 실질적 활용 확대 및 강화의 필요성이 제기된다.

❑ 보존전략 및 스토리지 체계

KOLIS는 보존이 아닌 도서관 업무 중심의 구조로 운영되고 있어 장기보존 업무를 온전히 수행하기에는 어려움이 존재하고 있다. 보존용 파일과 활용용 파일을 명확히 구분하여 관리하는 체계가 부재한 상황으로 향후 보존전략 수립 시 해당 관리 체계의 도입의 고려가 필요하다.

스토리지 측면에서는 예산에 기반을 둔 지속적인 증설이 이루어지고 있으나 운영용 스토리지와 보존용 스토리지가 통합적으로 관리되고 있지 않아서 이에 대한 고려도 필요한 것으로 보여진다. 또한, 평창 등 외부 인프라와의 연계 가능성도 검토 중에 있으나 구체적인 이중화・재해복구(DR) 체계는 아직 정립 단계에 있는 것으로 조사되었다.

보존전략 측면에서는 장기보존 전용 저장소와 이에 대응하는 관리 기준이 부재한 상황이며, PREMIS(Preservation Metadata: Implementation Strategies)나 FF 패키지(보존 패키지) 수준의 기술 요건 역시 아직 미흡한 것으로 보여지며 이에 따라 마이그레이션, 에뮬레이션과 같은 장기보존 기술의 도입 필요성이 요구되어 진다.

또한 메타데이터와 디지털자료를 통합하여 장기적으로 보존할 수 있는 보존 패키지 체계 역시 부재한 상태로 이에 대한 추후 장기보존을 위하여 함께 고려되어야 할 부분이다.

---

7) 디지털자료 보존을 위한 보존패키지 분석결과 보고, 자료보존연구센터, 2020.11.

❑ 오아시스(OASIS)

국립중앙도서관은 소멸 위험이 높은 공개 인터넷 자원을 국가 차원의 디지털 지적·문화유산으로 수집·보존하기 위하여 오아시스(OASIS, Online Archiving & Searching Internet Sources) 사업을 추진하고 있다. OASIS(Online Archiving & Searching Internet Sources) 웹자원의 수집과 보존은 「도서관법」 제22조에 규정된 온라인자료 수집 근거를 바탕으로 수행되며, 이를 통해 국내에서 생산·유통되는 웹 기반 정보자원을 확보하고 장기적으로 관리하는 역할을 수행하고 있다. 저작권자의 이용 허락 조건에 따라 서비스 범위는 관내 이용과 관외 이용으로 구분되어 운영되고 있다. 이용 동의 대상은 웹사이트 전체뿐만 아니라 웹사이트 내에 포함된 문서, 동영상, 이미지 등 개별 웹자료까지 포괄하고 있다.

OASIS(Online Archiving & Searching Internet Sources)는 2011년부터 국제인터넷보존협의체(IIPC, International Internet Preservation Consortium)에서 권장하는 웹자원 수집 도구인 헤리트릭스(Heritrix)를 도입하고 국제 표준 웹사이트 보존 파일 형식인 WARC(Web ARChive)를 채택하여 현재까지 지속적으로 적용하고 있다. 이를 통해 웹자원의 수집·보존 형식에 있어 국제 표준에 부합하는 기술적 체계를 유지하고 있다.

웹자원 수집 방식은 크게 선택적 수집과 포괄적 수집의 두 가지 방식으로 운영된다. 선택적 수집은 정부·공공기관, 언론기관, 상업기관, 문화예술기관 등 주요 기관 중 신설 또는 개편된 웹사이트를 대상으로 수행되며, 국가적 재난, 대규모 행사, 특정 사회 이슈나 주제와 관련된 웹자원을 집중적으로 수집하는 방식으로 이루어진다. 반면 포괄적 수집은 한국인터넷진흥원(KISA)에 등록된 대한민국 국가 도메인(KR) 웹사이트 전체를 대상으로 연 1회 전수 수집하는 방식으로 약 110만 건에 이르는 웹사이트를 A부터 Z까지 누락 없이 수집하는 대규모 국가 단위 웹아카이빙 체계이다.

수집된 웹자료는 URL 단위로 WARC(Web ARChive format) 파일 형식으로 저장된 후 KOLIS와 연계되어 서비스되는 구조를 갖고 있다.

수집 주기는 웹사이트의 성격과 중요도에 따라 차등 적용된다. 일반적인 웹사이트는 1년 주기로 재수집되며, 주요 공공기관 사이트나 사회적 이슈와 연결되는 사이트의 경우에는 3개월, 6개월 단위 등 보다 짧은 주기로 반복 수집이 이루어진다. 특히 선거, 재난, 감염병 확산 등 사회적으로 중대한 이슈와 직접 연관된 사이트는 필요에 따라 상시 수집이 수행된다.

수집 대상은 우리나라에서 생산된 웹사이트 및 웹사이트 내에 공개된 웹자료

전반을 포괄한다. 구체적으로는 정부 부처, 지방자치단체, 공공기관, 교육・연구기관, 비영리 민간단체, 문화예술기관, 상업 및 언론기관의 웹사이트뿐만 아니라 대한민국 국가도메인 웹사이트와 한국 관련 해외 웹사이트도 포함된다. 또한 문자자료, 음성・음향자료, 영상자료, 이미지자료 등 웹을 통해 공유되는 다양한 형태의 콘텐츠가 폭넓게 수집되고 있다.

웹사이트 수집 및 보존 절차는 ‘대상자료 선택 → 자료선별 사이트 등록 → 자료수집 설정 → 수집 및 재수집 → 보존 프로세스로 체계화되어 운영된다. 이를 통해 수집 대상의 선정 단계부터 기술적 수집, 저장, 장기보존에 이르기까지 전 주기가 구조화된 절차에 따라 관리되고 있다.

웹수집 데이터는 그 자체로 대규모 비정형 데이터의 성격을 띠고 있어 AI 학습 자원으로의 활용 가능성이 높지만 현재는 이용 동의와 저작권 문제로 인해 실제 활용에는 많은 제약이 존재하는 것으로 보여진다. 다만 향후 제목자동생성, 메타데이터 자동생성, 자동 분류・색인, 포맷 분석・보존 포맷 변환・위험 탐지, AI 기반 복원(문서・이미지・영상・음원) 등의 영역에서는 AI 기술을 충분히 적용할 수 있을 것으로 인식되고 있으며, 이에 대한 검토가 필요할 것으로 보여진다.

2.2. 국내외 사례조사 및 분석

2.2.1. 해외 사례 분석

2002년 1월 OAIS(Open Archival Information System) 참조모델이 제안된 전후로 세계 각국은 자신의 기관에 맞게 목표를 정하고 장기보존 정책 및 전략을 추진하고 있다. 이에 본 연구에서는 전자기록물 장기보존 정책 동향을 파악하기 위해 문헌조사를 실시하였으며, 그 중 전자기록물 장기보존 정책을 명시한 국가의 대표 아카이브 기관을 선정하여 비교분석 하였다.

본 연구에서는 장기보존 정책을 선언한 미국, 영국, 호주, 캐나다, 프랑스의 기관을 선정하였으며, 5개 국가 6개 기관의 장기보존 현황을 분석하고 정책, 수집/이관 체계, 보존기술(전략), 메타데이터 및 표준, 무결성 검증, 거버넌스 등의 요소를 기준으로 각 국가 기관의 전자기록 장기보존 정책을 비교 분석하였다. 분석한 결과를 통해 우리나라 장기보존 정책 및 전략을 다양한 전자기록물의 유형에 맞게 재설계하기 위한 시사점을 도출하고자 하였다.

❑ 미국 의회도서관 (Library of Congress, LC)

○ 정책

미국 의회도서관(Library of Congress, LC)은 디지털자원 장기보존을 “미래 세대가 신뢰할 수 있는 형태로 접근할 수 있도록 디지털 정보를 유지・관리하는 공공 책임”으로 인식하고 있으며, 이를 장기 접근성개념을 중심으로 설명한다. LC의 디지털 보존 정책은 디지털 객체의 단순한 비트 유지가 아니라 기술 환경 변화 이후에도 정보의 의미와 기능이 유지될 수 있도록 관리하는 것을 핵심 목표로 삼는다.[8)]

이러한 정책 인식은 2000년 미 의회 승인으로 출범한 국가 디지털 정보 인프라 및 보존 프로그램(National Digital Information Infrastructure and Preservation Program, NDIIPP)을 통해 제도적으로 구체화되었다. NDIIPP는 디지털 정보가 특정 기술・플랫폼・상업 환경에 종속될 경우 장기적 접근이 위협받을 수 있다는 문제의식을 바탕으로 연방 차원의 협력・조정・가이드 제공을 LC의 핵심 역할로 설정하였다.[9)]

8) https://www.LC.gov/preservation/digital/
9) https://www.LC.gov/LC/lcib/0601/ndiipp2.html

○ 수집・이관

LC의 수집 및 이관 정책은 전통적인 기록관 중심의 강제 이관 체계와 달리, 수집 단계에서부터 장기보존 가능성을 통제하는 방식으로 설계되어 있다. LC는 단일한 전자기록 이관 규정보다는 디지털 콘텐츠가 생성・제출・수집되는 시점에서부터 장기보존에 적합한 기술 조건을 갖추도록 유도하는 정책 도구를 활용한다.

이를 대표하는 문서가 「Recommended Formats Statement(RFS)」이다. RFS는 디지털 콘텐츠 유형별로 장기 접근성 확보에 유리한 포맷을 권고하는 문서로, 수집 단계의 포맷 선택이 장기보존 위험을 결정짓는 핵심 요소라는 정책 인식을 반영한다. LC는 RFS를 통해 수집 정책과 보존 정책을 분리하지 않고 하나의 연속된 관리 체계로 결합한다.[10)]

○ 보존기술

보존기술 측면에서 LC는 특정 기술 전략(예: 마이그레이션 또는 에뮬레이션)을 단일한 표준 해법으로 규정하지 않는다. 대신 LC는 디지털 포맷 자체의 특성이 장기보존 위험에 미치는 영향을 평가하기 위해 「Sustainability of Digital Formats」 프레임워크를 제시한다.[11)]

이 프레임워크는 디지털 포맷의 장기보존 적합성을 판단하기 위한 기준으로 ① 공개성(Disclosure), ② 채택성(Adoption), ③ 투명성(Transparency), ④ 자기설명성(Self-documentation), ⑤ 외부 의존성(External dependencies), ⑥ 특허 영향(Impact of patents), ⑦ 기술적 보호조치(Technical protection mechanisms) 등 7가지 지속가능성 요인을 명시하고 있다. LC는 이 기준을 통해 보존기술을 사후 변환 행위가 아닌 사전적 위험 관리 전략으로 위치시킨다.

○ 메타데이터 및 표준

메타데이터 및 표준 차원에서 LC는 국제 디지털 보존 표준 형성 과정에 중요한 기여를 하고 있다. 특히 LC는 PREMIS(Preservation Metadata: Implementation Strategies) 보존 메타데이터 표준의 유지・개정 활동을 담당하는 PREMIS(Preservation Metadata: Implementation Strategies) Editorial Committee의 공식 웹 호스트 및 지원 기관으로 참여하고 있다.[12)]

PREMIS(Preservation Metadata: Implementation Strategies) 공식 문서에 따르면, PREMIS(Preservation Metadata: Implementation Strategies) Editorial Committee는

10) https://www.LC.gov/preservation/resources/rfs/introduction.htm
11) https://www.LC.gov/preservation/digital/formats/
12) https://www.LC.gov/standards/premis

PREMIS(Preservation Metadata: Implementation Strategies) Data Dictionary와 XML(Extensible Markup Language) 스키마의 유지・개정・발전 방향을 조정하며, LC는 이 활동을 기술적・제도적으로 지원한다. 다만 PREMIS(Preservation Metadata: Implementation Strategies)는 다기관・커뮤니티 기반 표준으로 운영되며, LC가 이를 단독으로 운영・주관하는 구조는 아니다.[13] 따라서 LC의 역할은 표준 유지・확산을 위한 핵심 지원 기관으로 규정하는 것이 정확하다.

○ 무결성

LC의 무결성 관리는 개별 체크섬 알고리즘이나 검증 주기를 공개적으로 고정하기보다는, 정책・표준 차원의 전제 조건으로 제시된다. LC는 디지털 객체가 미래에도 진정한 자원으로 기능하기 위해서는 변경 여부를 검증할 수 있어야 한다는 점을 반복적으로 강조한다.

무결성 검증 결과와 보존 행위는 보존 메타데이터로 기록되며, 이는 장기 접근성 확보를 위한 필수 요소로 간주된다.[14] 다만 세부 기술 매개변수는 LC의 공개 정책 문서에서 일괄적으로 규정되지 않으므로, 일괄적으로 규정하지 않고 있다.

○ 거버넌스

거버넌스 측면에서 LC는 연방 입법부 소속 기관으로서 안정적인 제도적 기반을 갖고 있으며, NDIIPP를 통해 연방기관・학계・문화기관・민간 부문이 참여하는 디지털 보존 협력 구조를 형성하였다. 이 과정에서 국가 디지털 스튜어드십 연합(National Digital Stewardship Alliance, NDSA)이 2010년 출범하였으며, LC는 초기 형성에 핵심적인 역할을 수행하였다.[15]

이후 NDSA의 제도적 운영 거점은 DLF(Digital Library Federation, CLIR 산하)로 이전되었으며, 이는 LC 중심의 프로그램형 거버넌스에서 커뮤니티 기반 연합 구조로의 전환을 의미한다. LC는 이후에도 가이드・정책・표준 제공 기관으로서 디지털 보존 생태계에서 핵심적인 조정 역할을 유지하고 있다.

❑ 미국 국립문서기록관리청 (National Archives and Records Administration, NARA)

○ 정책

미국 국립문서기록관리청(National Archives and Records Administration, NARA)은 연방정부가 생산・관리하는 기록의 장기적 보존과 접근을 책임지는 국가

13) https://www.LC.gov/standards/premis
14) https://www.LC.gov/preservation/digital/
15) https://www.LC.gov/LC/lcib/0601/ndiipp2.html

기록관리 기관으로 디지털 보존을 기관의 핵심 임무 중 하나로 명시하고 있다. NARA는 디지털 보존을 기록의 진정성(Authenticity), 정확성(Accuracy), 기능성(Functionality)을 시간의 경과에도 유지하는 것으로 정의하며, 단순한 데이터 저장이 아닌 장기적 신뢰성과 이용 가능성 확보를 정책 목표로 설정한다.[16]

이러한 정책 인식은 「Digital Preservation Strategy 2022-2026」에 구체적으로 반영되어 있다. 해당 전략에서 NARA는 디지털 보존을 데이터 무결성(Data integrity), 포맷 및 매체 지속가능성(Format and media sustainability), 정보보안(Information security)을 핵심 요소로 구성된 포괄적 접근(Comprehensive approach)으로 설명한다. 이는 디지털 보존을 기술적 조치의 집합이 아니라 정책・운영・보안・기술이 결합된 조직 차원의 전략 과제로 인식하고 있음을 보여준다.[17]

○ 수집・이관

NARA의 수집・이관 체계는 연방정부 기록관리 제도와 직접적으로 연계되어 있다. 연방기관은 기록을 생성・관리한 후, 보존 가치가 확인된 전자기록을 NARA로 이관해야 하며, 이 과정은 NARA가 제공하는 지침과 절차에 따라 수행된다. 이관은 단순한 전달 행위가 아니라 장기보존을 전제로 한 관리 체계에 기록을 편입시키는 공식 절차로 이해된다.[18]

이러한 이관・보존의 핵심 인프라는 ERA(Electronic Records Archives) 시스템이다. NARA는 ERA를 연방정부 전자기록을 수집・보존・제공하기 위해 구축된 아카이브 시스템으로 정의하며, ERA를 통해 다양한 유형의 전자기록을 장기적으로 관리하고 있다. 공식 설명에서 ERA는 전자기록의 수명주기 관리와 접근 제공을 지원하는 중앙 시스템으로 소개된다.[19]

○ 보존기술

보존기술과 관련하여 NARA는 특정 기술(예: 마이그레이션(Migration), 에뮬레이션(Emulation)을 단일 해법으로 명시하지 않는다. 대신 전략 문서에서는 디지털 보존을 위해 표준 기반 절차, 위험 평가, 포맷 및 매체 지속가능성 관리가 필요하다고 설명한다. 이는 기술 선택을 고정하기보다, 기술 변화에 대응할 수 있는 관리 프레임을 유지하려는 접근으로 해석된다.

NARA는 또한 보존전략의 실행을 위해 절차와 표준의 문서화(Documentation of standards and procedures)를 중요한 요소로 제시한다. 이는 디지털 보존이 개별

16) https://www.archives.gov/about
17) https://www.archives.gov/preservation/digital-preservation/strategy
18) https://www.archives.gov/preservation/digital-preservation/guidance
19) https://www.archives.gov/era

기술 도입의 문제가 아니라 반복 가능하고 감사 가능한 운영 체계로 구현되어야 함을 전제한다.[20]

○ 메타데이터 및 표준

메타데이터 및 표준 차원에서 NARA는 디지털 기록의 장기보존을 위해 국제적 개념과 표준을 준거로 삼는 방향성을 명시하고 있다. 전략 문서에서는 OAIS(Open Archival Information System)와 같은 국제 참조모형을 디지털 보존체계 이해의 기반으로 언급하며, 메타데이터가 디지털 기록의 진정성과 이용 가능성을 유지하는 핵심 요소임을 강조한다.

다만 NARA의 공식 전략·정책 웹페이지에서는 PREMIS(Preservation Metadata: Implementation Strategies), METS(Metadata Encoding and Transmission Standard) 등 개별 메타데이터 표준을 필수로 지정하지 않고 있다. 즉, 특정 표준의 강제 적용보다는 국제 표준 개념을 참조한 메타데이터 관리 원칙을 채택하고 있는 것이 특징이다.[21]

○ 무결성

무결성은 NARA 디지털 보존 정책의 핵심 구성 요소 중 하나로 명시된다. NARA는 디지털 기록이 시간이 지나도 신뢰할 수 있으려면 변경 여부를 검증할 수 있어야 한다는 점을 전제로 하며, 이를 데이터 무결성 개념으로 설명한다.

전략 문서에서 무결성은 정보보안 및 포맷 지속가능성과 함께 디지털 보존의 기본 요소로 제시되며, 이는 디지털 기록의 신뢰성과 법적·행정적 효력을 유지하기 위한 필수 조건으로 이해된다. 다만 체크섬 알고리즘이나 검증 주기와 같은 세부 기술 매개변수는 전략 문서에 규정되어 있지 않다.[22]

○ 거버넌스

거버넌스 측면에서 NARA는 연방정부 기록관리 체계의 중심 기관으로서 디지털 보존과 관련된 전략 수립, 지침 제공, 이관 규정 관리 역할을 수행한다. NARA는 연방기관을 대상으로 디지털 기록 관리 및 이관 지침을 제공함으로써 분산된 기록 생성 환경을 중앙 정책으로 조정한다.[23]

또한 NARA는 디지털 보존전략을 중장기 계획(2022-2026)으로 수립하고 이를

20) https://www.archives.gov/preservation/digital-preservation/strategy
21) https://www.archives.gov/preservation/digital-preservation/strategy
22) https://www.archives.gov/preservation/digital-preservation/strategy
23) https://www.archives.gov/preservation/digital-preservation/guidance

통해 기관 차원의 목표·우선순위·성과 지표를 관리한다. 이는 디지털 보존을 개별 프로젝트가 아닌 지속적이고 제도화된 행정 프로그램으로 운영하려는 거버넌스 구조를 보여준다.

❑ 영국 국가기록원(The National Archives, UK, TNA)

○ 정책

영국 국가기록원(The National Archives, TNA)은 영국 정부 기록의 장기보존과 접근을 책임지는 중앙 기록관리 기관으로 디지털 보존을 국가 기록관리의 핵심 기능으로 명시하고 있다. TNA는 디지털 보존을 디지털 기록이 시간의 경과에도 신뢰할 수 있고 이해 가능하며 이용 가능한 상태로 유지되도록 하는 활동으로 설명하며, 기술 변화에 대응하는 지속적 관리의 필요성을 강조한다.[24)]

이러한 정책 인식은 TNA가 공개한 「Digital Preservation Policies and Guidance」 전반에 반영되어 있다. TNA는 디지털 보존을 단일 시스템이나 기술의 문제가 아니라 기록 생애주기 전반에 걸친 관리 프레임워크로 위치시키며, 이를 통해 정부 전체의 디지털 기록관리 수준을 제고하는 것을 목표로 한다.[25)]

○ 수집·이관

TNA의 수집·이관 체계는 「Public Records Act」에 근거하여 운영되며, 중앙정부 기관이 생산한 디지털 기록은 일정 기간이 경과한 후 TNA로 이관된다. 디지털 기록은 생성 단계에서부터 기관 내 기록관리 체계에서 관리되며, 이관 시점에는 장기보존을 전제로 한 절차와 기준이 적용된다.

TNA는 디지털 기록의 이관과 보존을 위해 워크플로우 기반의 접근을 채택하고 있으며, 이를 구체화한 문서가 「Guidance for Digital Preservation Workflows」이다. 해당 지침은 디지털 객체가 수집·이관·보존되는 전 과정을 단계별로 설명하며, 이관이 단순한 파일 전달이 아니라 검증·기술·기록을 포함하는 절차적 과정임을 명확히 한다.[26)]

○ 보존기술

보존기술 측면에서 TNA의 가장 중요한 특징은 포맷 식별과 기술적 이해를 보존전략의 출발점으로 삼는다는 점이다. TNA는 디지털 보존에서 기술 선택 이

24) https://www.nationalarchives.gov.uk/about/
25) https://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-records/
26) https://cdn.nationalarchives.gov.uk/documents/information-management/guidance-for-digital-preservation-workflows.pdf

전에 대상이 무엇인지 정확히 아는 것이 필수적이라는 원칙을 명시한다.

이를 위해 TNA는 PRONOM이라는 파일 포맷 레지스트리를 운영한다. PRONOM은 파일 포맷의 구조, 버전, 관련 소프트웨어, 위험 요인 등을 기술한 기술 레지스트리로, 디지털 객체의 장기보존 위험 평가에 활용된다. PRONOM의 데이터는 TNA가 개발·배포하는 DROID(Digital Record Object Identification) 도구를 통해 실제 파일 식별 과정에 적용된다.[27)]

○ 메타데이터 및 표준

메타데이터 및 표준 차원에서 TNA는 특정 보존 메타데이터 표준을 단일 해법으로 강제하기보다는 보존에 필요한 기술 정보의 체계적 제공을 핵심 역할로 설정한다. PRONOM이 제공하는 포맷 서명(Signature)과 기술 정보는 OAIS(Open Archival Information System) 참조모형에서 정의하는 표현정보(Representation Information)와 유사한 기능을 수행한다.[28)]

TNA는 또한 디지털 보존 워크플로 지침에서 메타데이터가 보존 행위와 검증 결과를 기록하는 데 필수적임을 강조하며, 기술 메타데이터·관리 메타데이터·보존 메타데이터가 함께 관리되어야 함을 명시한다. 이는 메타데이터를 단순 기술 요소가 아닌 보존 행위의 증거 기록으로 인식하는 접근이다.[29)]

○ 무결성

무결성 검증은 TNA 디지털 보존 정책에서 명확히 확인되는 핵심 요소이다. 「Guidance for Digital Preservation Workflows」에서는 디지털 객체의 무결성을 확인하기 위한 체크섬(Checksum) 생성 및 검증 절차가 워크플로 단계에 포함되어야 함을 명시한다.

특히 해당 지침에서는 SHA-256과 같은 암호학적 해시 함수를 활용한 무결성 검증을 예시로 제시하며, 체크섬을 통해 파일의 변경·손상 여부를 탐지할 수 있음을 설명한다. 이는 무결성 검증이 권고 수준이 아니라 실제 운영 워크플로에 통합된 절차적 요소임을 보여준다.[30)]

○ 거버넌스

거버넌스 측면에서 TNA는 단일 기관 운영을 넘어, 국가 및 국제 디지털 보존

27) https://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-records/droid/
28) https://www.nationalarchives.gov.uk/PRONOM/
29) https://cdn.nationalarchives.gov.uk/documents/information-management/guidance-for-digital-preservation-workflows.pdf
30) https://cdn.nationalarchives.gov.uk/documents/information-management/guidance-for-digital-preservation-workflows.pdf

생태계의 핵심 조정자 역할을 수행한다. TNA는 PRONOM과 DROID를 무료로 공개·배포함으로써, 영국뿐 아니라 국제 디지털 보존 커뮤니티가 동일한 기술 기준을 활용할 수 있도록 지원한다.[31]

또한 TNA는 Digital Preservation Coalition(DPC)의 설립 및 활동에 깊이 관여하며, 정책·기술·교육·커뮤니티 차원의 협력 구조 형성을 지원한다. 이는 디지털 보존을 단일 기관의 기술 문제가 아닌 공공 부문 전체의 공동 책임 영역으로 확장하려는 거버넌스 전략으로 해석된다.[32]

❑ 호주 국가도서관(National Library of Australia, NLA)

❍ 정책

호주 국가도서관(National Library of Australia, NLA)은 호주의 출판유산과 문화기록을 수집·보존·제공하는 국가대표 기관으로서 디지털자원을 포함한 전체 컬렉션의 장기적 보존을 기관의 핵심 책무로 명시하고 있다. NLA는 보존을 현재와 미래 세대가 컬렉션에 지속적으로 접근할 수 있도록 보장하기 위한 활동으로 정의하며, 디지털자료 역시 물리적 컬렉션과 동일한 보존 정책의 적용 대상임을 분명히 한다.[33]

이러한 정책적 인식은 법정납본 제도의 디지털 확장을 통해 제도적으로 뒷받침된다. 호주는 「Copyright Act 1968」의 개정을 통해 2016년부터 전자출판물과 온라인 출판물을 법정납본 대상으로 포함시켰으며, 이에 따라 디지털 출판물의 수집과 보존은 국가적 의무로 규정되었다. NLA는 해당 법적 근거에 따라 디지털 출판물의 장기보존을 수행한다.[34]

❍ 수집·이관

NLA의 디지털자원 수집·이관 체계는 NED(National eDeposit) 시스템을 중심으로 운영된다. NED는 호주 전역의 출판사와 저자가 전자책, 전자저널 등 디지털 출판물을 온라인으로 납본할 수 있도록 구축된 국가 차원의 통합 시스템으로 NLA와 각 주·준주 주립도서관이 공동으로 운영한다. 제출된 디지털자료는 장기보존을 전제로 관리 체계에 편입되며, 법정납본의 디지털 구현 수단으로 기능한다.[35]

웹자원의 경우 NLA는 AWA(Australian Web Archive)를 통해 호주 웹 환경을

31) https://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-records/droid/
32) https://www.dpconline.org/about
33) https://www.library.gov.au/discover/about-collection
34) https://www.nla.gov.au
35) https://www.library.gov.au/services/publishers-and-self-publishers/national-edeposit

체계적으로 수집·보존한다. AWA는 정부 웹사이트, 호주 도메인(.au) 웹사이트의 정기적 대규모 수집과 함께, 사회·문화적으로 중요한 주제와 사건을 중심으로 한 선택적 웹 수집을 병행하는 구조로 운영된다. 이는 대량 수집과 큐레이션을 결합한 국가 웹아카이빙 모델이다.[36]

○ 보존기술

보존기술 측면에서 NLA는 특정 단일 기술을 장기보존의 해법으로 제시하지 않고 표준화된 포맷과 관리 가능한 구조를 채택하는 방향을 취한다. 웹 아카이브 자료는 국제 표준 포맷인 WARC(Web ARChive) 형식으로 저장되며, 이는 웹 콘텐츠의 구조와 맥락 정보를 함께 보존하기 위한 장기보존용 포맷이다.

또한 NLA는 디지털자원의 생성·관리·보존 과정에서 표준과 지침을 활용하여 품질과 지속 가능성을 확보하고자 하며, 디지털화 및 디지털자원 관리와 관련된 지침을 공개적으로 제공한다. 이러한 접근은 기술 변화에 대비한 유연한 보존전략으로 해석된다.[37]

○ 메타데이터 및 표준

메타데이터 및 표준 차원에서 NLA는 디지털 컬렉션의 기술·관리·접근을 위해 국제적으로 통용되는 도서관 및 디지털자원 관리 관행을 적용한다. 디지털 출판물과 기타 컬렉션은 서지 기술을 위해 MARC 기반 메타데이터를 활용하며, 웹 아카이브 자료는 WARC(Web ARChive) 메타데이터 구조에 따라 관리된다.

보존된 디지털자원은 Trove 플랫폼을 통해 검색·접근이 가능하도록 제공된다. Trove는 NLA가 운영하는 국가 통합 발견 서비스로, 보존과 접근을 연결하는 핵심 서비스 계층으로 기능한다. 이는 디지털 보존이 단순한 저장을 넘어 장기적 이용 가능성을 전제로 설계되어 있음을 보여준다.[38]

○ 무결성

무결성(Integrity) 관리는 NLA의 보존 정책에서 전제 조건으로 다루어진다. NLA는 보존 과정에서 디지털자원의 신뢰성과 변경 여부를 관리하는 것이 장기 접근성 확보의 핵심 요소임을 강조한다.

다만 NLA의 공개 정책 문서와 안내 페이지에서는 체크섬 알고리즘, 검증 주기 등 구체적인 기술적 운영 매개변수를 일괄적으로 명시하지 않는다. 따라서 확인 가능한 범위에서 NLA의 무결성 관리는 표준 포맷 사용, 반복 수집 구조, 안정적

36) https://www.library.gov.au/discover/what-we-collect/archived-websites
37) https://www.nla.gov.au/standards/digitisation-guidelines
38) https://trove.nla.gov.au/about

저장 환경을 통한 신뢰성 확보라는 정책적 수준에서 기술하는 것이 타당하다.

○ 거버넌스

거버넌스 측면에서 NLA의 디지털 보존은 단독 기관 운영이 아닌 국가 단위 협력 모델을 특징으로 한다. NED 시스템은 NLA와 주·준주 주립도서관이 공동으로 운영하는 구조이며, 디지털 출판물의 수집·보존 책임을 분산·공유한다.

Australian Web Archive 역시 NLA를 중심으로 정부 기관과 주립도서관이 협력하여 운영되며, 이는 대규모 디지털자원의 장기보존을 위한 분산 협력형 거버넌스 구조로 평가된다. 이러한 체계는 호주가 디지털 보존을 국가 문화유산 관리의 공동 책임으로 인식하고 있음을 보여준다.[39]

❑ 캐나다 국립도서관·기록관(Library and Archives Canada, LAC)

○ 정책

캐나다 국립도서관·기록관(Library and Archives Canada, LAC)은 캐나다 연방정부의 국가도서관과 국가기록관 기능을 통합한 기관으로 출판유산과 행정 기록을 포함한 국가 지식자원의 장기적 보존을 법적 책무로 수행한다.[40] LAC는 디지털자원 보존을 기관의 핵심 기능으로 명시하며, 디지털 기록과 출판물이 장기적으로 진정성(Authenticity), 신뢰성(Reliability), 접근 가능성(Accessibility)을 유지해야 함을 정책 차원에서 선언한다.

특히 LAC는 「Policy on Digital Preservation」을 통해 디지털 보존의 목적과 적용 범위를 명확히 규정한다. 해당 정책은 디지털 보존이 단순한 기술적 조치가 아니라 LAC의 통제 또는 관리(custody or control)하에 놓인 모든 디지털자원에 대해 지속적으로 적용되어야 할 조직 차원의 책임임을 분명히 한다.[41]

○ 수집·이관

수집·이관과 관련하여 LAC는 디지털자원의 보존 책임이 이관 완료 이후가 아니라 LAC가 해당 자원에 대한 통제 또는 관리 권한을 확보하는 시점부터 발생한다는 원칙을 채택한다. 이 원칙은 디지털 출판물과 정부 기록 모두에 동일하게 적용된다.

상위 정책 문서인 「Preservation Policy Framework」는 LAC의 모든 보존 활동이 이 프레임워크에 의해 지배되며, 출판물과 기록이 LAC의 관리하에 들어오는 즉시 보존 책임이 발생함을 명시한다. 이는 수집–이관–보존을 분절된 단계가 아닌 연속적인 관리 책임 구조로 이해하는 LAC의 특징을 보여준다.[42]

---

39) https://www.library.gov.au/services/publishers-and-self-publishers/national-edeposit
40) https://www.canada.ca/en/library-archives/corporate/about-us.html
41) https://www.canada.ca/en/library-archives/corporate/about-us/policies/digital-preservation.html

○ 보존기술

보존기술 측면에서 LAC는 특정 저장소 솔루션이나 단일 기술 전략을 공식 정책에서 고정하지 않는다. 대신 디지털 보존이 달성해야 할 핵심 속성—진정성, 무결성, 장기 접근성—을 먼저 규정하고 기술적 구현은 이러한 속성을 충족하도록 지속적으로 개선되어야 함을 강조한다.

이러한 접근은 「Strategy for a Digital Preservation Program」에서도 확인된다. 해당 전략 문서는 LAC의 디지털 보존 프로그램이 기술 변화에 대응할 수 있도록 유연하게 설계되어야 하며, 국제적 모범 관행과 표준(OAIS 참조모형 등)을 고려한 단계적 발전을 지향함을 설명한다. 이는 디지털 보존을 단기적 시스템 구축이 아닌 장기적 조직 역량으로 이해하는 관점에 기반한다.[43]

○ 메타데이터 및 표준

메타데이터 및 표준 차원에서 LAC는 디지털 기록의 진정성・신뢰성・이해 가능성 확보를 위해 메타데이터의 핵심적 역할을 명확히 한다. 「Policy on Digital Preservation」은 디지털자원이 장기적으로 보존되기 위해서는 생성・관리・이관 과정에서 축적된 메타데이터가 유지되어야 함을 강조한다.

또한 「Preservation Policy Framework」는 메타데이터를 보존 책임 이행의 필수 요소로 규정하며, 기록과 출판물이 LAC로 이관될 때 적절한 메타데이터가 동반되어야 함을 전제로 한다. 이와 같이 LAC는 특정 국제 표준(PREMIS, METS 등)을 정책 문서에서 일괄적으로 열거하기보다는 메타데이터의 기능적 필요성과 책임성을 제도적으로 확정하는 방식으로 접근한다.[44]

○ 무결성

무결성(Integrity)은 LAC 디지털 보존 정책의 핵심 원칙 중 하나로 명시된다. 「Policy on Digital Preservation」은 디지털자원이 보존 과정에서 변경・손상되지 않았음을 입증할 수 있어야 하며, 이는 진정성 유지의 전제 조건임을 분명히 한다.

다만 LAC의 공개 정책 문서와 전략 문헌에서는 체크섬 알고리즘, 검증 주기, 자동화 수준 등 세부 운영 매개변수를 구체적으로 명시하지 않는다. 따라서 확인 가능한 범위에서 LAC의 무결성 관리는 정책적 원칙과 책임 수준에서 명확히 규정되어 있으며, 구체적 기술 구현은 내부 운영 체계에 의해 관리되는 것으로 해석하는 것이 타당하다.[45]

---

42) https://www.canada.ca/en/library-archives/corporate/about-us/policies/preservation-framework.html
43) https://www.bac-lac.gc.ca/eng/about-us/publications/Documents/LAC-Strategy-Digital-Preservation-Program.pdf
44) https://www.canada.ca/en/library-archives/corporate/about-us/policies/preservation-framework.html
45) https://www.bac-lac.gc.ca/eng/about-us/publications/Documents/LAC-Strategy-Digital-Preservation-Program.pdf

○ 거버넌스

거버넌스 측면에서 LAC는 정책-프레임워크-전략 문서로 이어지는 다층 구조를 통해 디지털 보존을 관리한다. 「Preservation Policy Framework」는 LAC 전 조직에 적용되는 상위 준거로 기능하며, 개별 보존 활동과 기술적 구현은 이 프레임워크 하에서 수행된다.[46)]

국가도서관과 국가기록관 기능을 통합한 기관 구조는 출판물, 정부 기록, 디지털자원 등 다양한 객체군을 하나의 정책 체계 아래에서 포괄적으로 관리할 수 있게 한다. 이는 LAC가 캐나다 국가 차원의 디지털 유산 보존을 총괄하는 거버넌스 중심 기관으로 기능함을 보여준다.

❑ 프랑스 국립도서관(Bibliothèque nationale de France, BnF)

○ 정책

프랑스 국립도서관(Bibliothèque nationale de France, BnF)은 프랑스 국가의 출판 유산과 디지털 문화유산을 수집・보존・제공하는 대표 기관으로서 디지털 장기보존을 법제에 의해 명확히 부여된 국가적 책무로 수행한다.[47)] 프랑스의 법정납본(dépôt légal) 제도는 전통적인 인쇄자료를 넘어 전자출판물과 웹 기반 디지털자원까지 포괄하도록 확장되어 왔으며, BnF는 이 제도의 핵심 수행 기관으로 기능한다.[48)]

특히 프랑스는 2006년 제정된 「DADVSI 법(Loi n° 2006-961)」을 통해 웹 법정납본(dépôt légal de l’internet)을 도입하였고, 이후 관련 시행 규정을 통해 웹 콘텐츠의 수집・보존・접근 제공을 국가 책임으로 명문화하였다. 이에 따라 BnF는 프랑스 도메인 하에서 공개적으로 접근 가능한 웹사이트를 장기적으로 보존할 법적 권한과 의무를 가진다.[49)]

○ 수집・이관

BnF의 디지털자원 수집・이관 체계는 법정납본에 근거한 의무적 수집 구조를 특징으로 한다. 전자출판물과 온라인 출판물은 법에 따라 납본 대상이 되며, 웹 자원의 경우 BnF가 국가 차원에서 대규모 자동 수집을 수행한다.[50)]

웹아카이빙은 중앙 집중형 방식에 한정되지 않고 BnF와 지역 파트너 도서관 및 사서 네트워크가 협력하는 분산적 구조로 운영된다. BnF 공식 설명에 따르면,

46) https://www.canada.ca/en/library-archives/corporate/about-us/policies/preservation-framework.html
47) Digital preservation at the National Library of France, IFLA, 2008
48) A data-first approach to digital preservation: the SPAR project, Bermès, E., 2010
49) Semantic Web technologies for digital preservation: the SPAR projectBermès, E., et al., 2008
50) Digital preservation at the National Library of France, IFLA, 2008

사서들은 주제・사건・사회적 중요성에 따라 웹 콘텐츠를 선정하여 미래의 유산을 구축하는 데 참여하며, 이는 대량 수집과 선택적 큐레이션을 결합한 방식으로 이해된다.

❍ 보존기술

프랑스 국립도서관(BnF)의 디지털 장기보존 인프라는 SPAR(Scalable Preservation and Archiving Repository) 프로젝트를 중심으로 구축되어 왔다. SPAR는 BnF가 대규모 디지털자원을 장기적으로 보존하기 위해 설계한 저장소 아키텍처로, OAIS(Open Archival Information System, ISO 14721) 참조모형을 개념적 기반으로 삼아 모듈형・확장형 구조를 채택한 것이 특징이다.

IFLA(World Library and Information Congress)에서 발표된 Bermès(Bermès, Carbone, Ledoux, Lupovici) 등의 연구는 SPAR가 단일 기술 솔루션이 아니라 데이터 중심 접근을 취하는 장기보존 시스템임을 설명한다[2] . 해당 연구에 따르면 SPAR는 다양한 유형의 디지털 객체(전자출판물, 디지털화 자료, 웹 아카이브 등)를 수용하기 위해 저장 계층과 관리 계층을 분리하고 장기적 확장성과 기술 변화 대응을 고려한 구조로 설계되었다.

또한 SPAR는 초기 설계 단계부터 대량 데이터 처리와 장기적 신뢰성 확보를 주요 목표로 설정하였으며, 이는 국가대표 도서관이 직면한 대규모 디지털 수집 환경을 반영한 결과로 해석된다.[51)]

❍ 메타데이터 및 표준

SPAR 프로젝트 관련 학술 문헌들은 BnF의 디지털 보존체계가 OAIS(Open Archival Information System) 개념에 따라 정보 패키지와 보존 메타데이터를 관리하는 구조를 전제하고 있음을 공통적으로 지적한다. 특히 “A data-first approach to digital preservation” 논문은 SPAR가 METS(Metadata Encoding and Transmission Standard)와 PREMIS(Preservation Metadata: Implementation Strategies)와 같은 국제 표준 메타데이터 모델을 활용할 수 있도록 설계된 아키텍처임을 설명한다.

다만 BnF의 공식 웹페이지나 공개 정책 문서에서는 SPAR 운영에 적용되는 메타데이터 표준의 구체적 적용 범위(필수/선택 요소, 내부 프로파일 등)를 일괄적으로 규정하지는 않는다. 따라서 확인 가능한 사실은 SPAR가 OAIS(Open Archival Information System) 준거 저장소로서 국제 표준 메타데이터 체계를 수용할 수 있는 구조적 기반을 갖추고 있다는 점까지로 한정하는 것이 타당하다.

51) A data-first approach to digital preservation: the SPAR project, Bermès, E., 2010

○ 무결성

무결성(Integrity) 관리와 관련하여, SPAR 프로젝트 문헌들은 시스템이 신뢰 저장소(Trusted digital repository)를 지향하며, 장기보존 과정에서 디지털 객체의 신뢰성과 진정성을 유지하는 것을 핵심 목표로 삼고 있음을 명시한다. 이는 OAIS(Open Archival Information System) 준거 설계 자체가 장기보존 대상 객체의 무결성 유지와 변경 이력 관리의 필요성을 전제하기 때문이다.[52)]

그러나 체크섬 알고리즘(SHA-256 등), 검증 주기, 복제 정책과 같은 구체적 운영 매개변수는 공개 학술 문헌이나 BnF 공식 웹페이지에서 명시적으로 확인되지는 않는다. 이에 따라 본 사례에서는 무결성 관리가 SPAR의 설계 목표와 개념적 프레임에 포함되어 있음을 확인 가능한 수준으로 기술한다.

○ 거버넌스

SPAR는 BnF 단독의 기술 프로젝트가 아니라 프랑스 국가 차원의 디지털 유산 보존전략의 일부로 추진되었다. 관련 연구들은 SPAR가 BnF 내부 요구뿐 아니라 국가적 디지털 보존 역량을 장기적으로 강화하기 위한 인프라로 설계되었음을 강조한다.

이러한 접근은 BnF가 법정납본과 웹아카이빙을 통해 축적되는 방대한 디지털 자원을 안정적으로 관리하기 위해, 기술 아키텍처를 거버넌스 차원의 장기 전략과 결합하고 있음을 보여준다. 그러므로 프랑스의 디지털 장기보존은 법제 기반의 책무, 중앙기관(BnF)의 실행, 그리고 확장 가능한 기술 인프라(SPAR)가 결합된 구조로 이해할 수 있다.[53)]

52) Digital preservation at the National Library of France, IFLA, 2008
53) Semantic Web technologies for digital preservation: the SPAR projectBermès, E., et al., 2008

구분	미국 의회도서관 (LC)	미국 국립문서기록관리청 (NARA)	영국 국가기록원 (TNA)	호주 국가도서관 (NLA)	캐나다 국립도서관 · 기록관 (LAC)	프랑스 국립도서관 (BnF)
정책	• 장기 접근성(long-term access)을 핵심 목표로 설정하고, 포맷 지속가능성(sustainability factors)을 통해 수집 단계에서 장기보존 위험을 최소화하는 정책을 명시함	• 디지털 보존을 전략 문서로 명시하고, 데이터 무결성 · 포맷/매체 지속가능성 · 정보보안을 핵심 원칙으로 제시함	• 공공기록의 장기보존을 국가 책무로 규정하며, 디지털 보존을 기술 인프라와 지침 제공을 통한 국가적 기능으로 위치시킴	• 디지털을 포함한 전체 컬렉션의 장기보존을 기관 핵심 책무로 규정하고, 법정납본을 디지털 환경까지 확장함	• 디지털 보존을 조직 전체에 적용되는 정책적 책무로 규정하고, 통제/관리(custody or control) 개념을 핵심 기준으로 설정함	• 디지털 장기보존을 법정납본 제도에 의해 부여된 국가적 의무로 규정함
수집 /이관	• 권고 포맷 체계(RFS)를 통해 수집 단계에서 장기보존 적합성을 제도화하며, 수집-보존을 정책적으로 직접 연결함	• 연방기관 기록을 ERA 시스템을 통해 이관받으며, 이관을 보존체계 작동의 출발점으로 설정함	• 정부기관 기록을 SIP 형태로 이관받고, 웹 자원은 UK Web Archive를 통해 대규모 수집함	• National eDeposit(NED)을 통해 전자출판물 납본을 수행하고, Australian Web Archive를 통해 웹 자원을 수집함	• 출판물과 기록이 LAC의 통제/관리하에 놓이는 시점부터 보존 책임이 발생하는 구조를 채택함	• 전자출판물과 웹 자원을 법정납본에 따라 의무적으로 수집하며, 웹은 대규모 자동 수집과 선택적 큐레이션을 병행함
보존기술	• 특정 단일 기술을 고정하지 않고, 표준화된 포맷 선택을 통해 사전적 위험 저감을 강조함	• 마이그레이션과 운영 개선을 병행하며, 전략 문서를 통해 지속적 평가 · 개선을 강조함	• PRONOM 포맷 레지스트리와 DROID 식별 도구를 중심으로 보존전략을 기술적으로 지원함	• WARC 등 국제 표준 포맷을 사용하고, 디지털화 · 보존 지침을 통해 기술적 일관성을 유지함	• 특정 기술을 고정하지 않고, 국제 모범 관행(OAIS 등)을 고려한 단계적 발전 전략을 채택함	• OAIS 준거의 SPAR 저장소를 구축하여 대규모 디지털자원의 장기보존을 수행함
메타 데이터/표준	• PREMIS 유지 · 개정 활동을 지원하며, 커뮤니티 표준의 지속성을 통해 상호운용성을 확대함	• 특정 표준 열거보다는 ISO 16363, NDSA Levels 등 성숙도 평가 도구를 준거로 활용함	• PRONOM을 통해 포맷 기술 정보를 제공하고, 워크플로 지침에서 메타데이터 활용을 명시함	• 서지 · 웹 메타데이터를 통해 Trove 기반 발견 서비스를 제공하며, 보존-접근을 연계함	• 메타데이터를 진정성 · 신뢰성 확보의 필수 요소로 규정하고 정책 차원에서 요구함	• OAIS 정보패키지 개념에 기반한 메타데이터 관리 구조를 전제함
무결성	• 장기 접근성 확보의 전제조건으로 무결성을 강조하나, 운영 매개변수는 공개적으로 상세 규정하지 않음	• data 를 디지털 보존의 핵심 요소로 명시하고, 실무 공유 채널을 통해 절차를 다룸	• SHA-256 체크섬을 워크플로에 통합하여 무결성 점검을 절차적으로 명시함	• 표준 포맷 사용과 반복 수집 구조를 통해 무결성 확보를 전제함	• 무결성을 정책적 핵심 원칙으로 규정하되, 세부 기술 값은 내부 운영 영역으로 둠	• 신뢰 저장소를 지향하며 무결성 유지를 전제로 하나, 구체적 알고리즘은 공개 문헌에 명시되지 않음
거버넌스	• NDSA 출범 등 협력 네트워크를 통해 국가 · 커뮤니티 차원의 조정 역할을 수행함	• 전략-평가-운영 요소를 결합한 조직 내재형 거버넌스 구조를 형성함	• PRONOM · DROID · 지침을 공개하여 국가 및 국제 보존 생태계의 인프라 제공자로 기능함	• NLA-주립도서관 협력 모델을 통해 국가 단위 분산 거버넌스를 형성함	• 정책-프레임워크-전략 문서를 통해 통합형 국가 거버넌스를 구축함	• 법제-중앙기관-분산 협력 네트워크가 결합된 국가 주도형 거버넌스를 형성함

[표 8] 해외 사례 비교표

2.2.2. 국내 사례 분석

국내 사례 분석에서는 디지털자원 보존과 영상, 음원, 이미지, 3D 자료 등 다양한 형태의 자원을 관리하고 보존하는 국내의 기관들을 대상으로 보존 정책, 디지털 아카이빙 정책, 기록 관리 체계, 보존전략 수립 등 국립중앙도서관에서 필요한 부분에 대하여 비교 분석을 수행하고 이를 통해 실제 적용 가능성이 높은 시사점을 도출할 수 있다.

❑ 국가기록원

국가기록원의 디지털자원 보존은 공공기록물의 생산-관리-이관-보존-평가에 이르는 생애주기 관리체계에 기반하며, 기록물의 폐기는 기록물관리 전문요원 심사와 기록물평가심의회 심의를 거쳐야 한다는 법적 절차가 명시되어 있다.
수집(이관) 정책 측면에서 전자기록물은 장기보존을 위한 표준화된 장기보존 패키지 체계로 구조화・포맷화하여 관리하도록 공공표준・고시 체계가 운영되고 있으며, XML(Extensible Markup Language) 기반 장기보존패키지(NEO2) 기술규격이 행정규칙(국가기록원 고시) 형태로 제시되어 있다.

보존 기술・전략(포맷・패키징 포함)은 ① 문서유형 전자기록물의 보존포맷을 PDF/A-1b 기반으로 규격화하는 기술규격(고시)과 ② 전자기록물 장기보존패키지 기술규격(NEO2/NEO3)로 대표되는 보존포맷+패키징 표준의 조합으로 구현된다. 또한 보존포맷 범주에 웹기록물 WARC(Web ARChive) , 시청각 포맷(예: WAVE, MPEG-4, FFV1 등), 데이터세트 포맷(예: JSON, XML, CSV, SIARD 등)이 포함될 수 있음을 표 형태로 확인할 수 있다.

메타데이터・표준은 기록관리 메타데이터 표준(NAK 8)으로 공공표준 문서가 공개되어 있으며, 해당 표준이 NAK 8:2022(v2.3)로 제・개정 이력을 포함해 배포되는 형태임이 확인된다.

자료 폐기(제적) 제도는 「공공기록물 관리에 관한 법률」[54] 제27조에 따라 공공기관 기록물 폐기 시 전문요원 심사 및 평가심의회 심의를 요건으로 하는 절차 통제가 핵심이다.

54) 공공기록물 관리에 관한 법률(법률 제21066호, 시행 2025. 10. 1.)

❑ 한국영상자료원

정책·법적 기반과 수집 정책은 「영화 및 비디오물의 진흥에 관한 법률」[55] 제35조(영화필름 등의 제출)를 근거로 한 ‘개봉영화 제출제도’ 운영으로 확인되며, 제출대상(등급분류를 받은 작품), 제출기한(등급 부여일로부터 60일 이내) 등 운영 조건이 기관 안내문에 명시되어 있다.

보존 기술·전략(포맷·패키징 포함)에서 디지털시네마 의무제출 포맷을 DCP로 지정하고 DCP 제출이 불가한 경우 MOV(ProRes 422HQ 이상)로 제출하도록 한 구체 포맷 기준이 공지문으로 확인된다.

저작권 관리는 기관 ‘저작권정책’에서 공공누리 제1유형/제4유형을 통해 출처표기, 상업적 이용 가능 여부, 2차적 저작물 작성 가능 여부를 구분해 고지하고 있다.

자원 제공 방식은 제출·보존된 결과물이 기관 서비스(예: KMDb 등)로 제공되는 구조가 약관 및 기관 서비스 안내로 확인되지만, 이 응답에서는 제출제도 운영과 홈페이지 저작권 정책 고지처럼 문서로 직접 확인되는 범위만을 반영하였다.

❑ 국립국악원

정책·법적 기반으로는 국립국악원의 기능이 민족음악의 보존·전승(및 보급·발전)[56]에 있음을 직제 조문에서 확인할 수 있다.

수집 정책은 국악아카이브 ‘기증 안내’에서 국악(춤·연희·창작 포함) 관련 모든 형태 자료를 대상으로 하며, 희소성·사료적 가치·활용 가치 등을 ‘수증심의위원회’에서 심의해 수증 여부를 최종 결정한다고 명시되어 있다.

보존 기술·전략(포맷·패키징 포함)과 온라인자원(영상·음원·이미지·텍스트 등) 관리 흐름은 수집 자료의 세부정보 정리 후 디지털변환을 수행하고 원본과 복제본을 항온항습실에 보존한다는 운영 서술로 확인되며, 관리 대상 유형이 동영상·음향·이미지·텍스트 등임도 함께 확인된다.

메타데이터·표준 및 제공 방식(열람·복제)은 ‘이용안내’ 및 FAQ에서 ① 홈페이지/공간 내 PC를 통한 검색, ② 원본 열람 사전신청 및 내부 규정 검토에 따른 허가, ③ 열람·복제 횟수/점수/시간제한, ④ 공공누리 유형에 따른 다운로드/다운로드 신청 가능, ⑤ 공공누리 미지정 자료의 경우 별도 문의 필요 등의 조건이 명시되어 있다.

---

55) 영화 및 비디오물의 진흥에 관한 법률(법률 제21066호, 시행 2025. 10. 1.)
56) 국악진흥법(법률 제19567호, 시행 2024. 7. 26.)

저작권 관리는 저작권을 확보한 자료를 선별 가공하여 인터넷 열람 서비스를 제공한다는 원칙이 아카이브 소개에 명시되어 있어, 권리 확보 여부가 온라인 제공 가능성의 전제가 됨을 문서로 확인할 수 있다.

❑ 국립아시아문화전당

정책·법적 기반은 「아시아문화중심도시 조성에 관한 특별법」[57]에서 국립아시아문화전당 관련 조문 체계가 존재함을 확인할 수 있으며, 전당 내부 운영에 관해서는 ‘기본운영규정’, ‘아시아문화박물관 운영 규정’ 등 행정규칙 형태의 근거가 확인된다.

수집 정책(입수 경로)은 전당 아카이브가 수집을 통해 입수하는 자료(전문주제자료·기증자료)와 전당 부서로부터 이관받는 자료(ACC자료)를 함께 관리하는 구조로 서술되어 있다.

또한 열람 서비스는 온라인(아시아문화박물관 아카이브 누리집)과 오프라인(특별열람실 신청·승인 후 열람)의 병행 구조로 운영 조건이 안내되어 있다.

보존 기술·전략(포맷·패키징 포함)과 디지털화 기준은 (1) 영상 디지털화 세부기준으로 해상도/화면비/코덱/확장자(예: ProRes 422 MOV, H.264 MP4 등)를 제시한 표가 확인되며, (2) 박물류의 경우 2D는 보존본 TIFF·활용본 JPEG, 3D는 3D 스캐닝 원본(STL 등)-병합 데이터(PLY, STL 등)-후처리 데이터(PLY, OBJ 등)로 구분한 포맷 기준이 표로 확인된다.

이와 같이 이미지(예: TIFF/JPEG), 영상, 3D 등 다양한 형태의 온라인자원을 디지털화 기준(품질·포맷)에 따라 생산·관리하는 사실을 확인할 수 있다.

메타데이터·표준과 운영 체계는, 전당 아카이브 관리가 ‘사업정보 > 단위사업정보 > 폴더 > 아이템’ 계층 구조로 구성되어 생산 자료를 사업 단위로 그룹화해 관리한다는 구조가 명시되어 있으며, 저작권 관리를 위해 저작권 유형·기간·서류(첨부파일) 등 다수(예: 17개)의 저작권 메타데이터 요소를 설계하고 별도 메뉴로 권리 보유 현황을 점검하고 있다는 점을 확인할 수 있다.

기술 표준(검색·서비스/관리 기능)은 아카이브 관리시스템이 검색(DB/전문/분류/시청각 검색), 열람·대여 신청/승인/반납, 통계, 시소러스(대표어/동의어/상하위어 등) 관리, OpenAPI 관리, 감사이력 및 메타데이터 관리 등 기능 범위를 갖는 것으로 정리되어 있다.

57) 아시아문화중심도시 조성에 관한 특별법(법률 제21065호, 시행 2025. 10. 1.)

저작권 관리와 자원 제공 방식은 저작자를 확인하기 어렵거나 분쟁 소지가 있는 자료는 원칙적으로 온라인 서비스를 제공하지 않고 목록 게시 또는 신청 기반 방문열람으로 제공한다는 원칙, 모든 자료 공개를 원칙으로 하되 저작권이 해결되지 않은 자료는 비공개한다는 공개기준이 문서에 명시되어 있다.

또한 일부 컬렉션에서는 2차적 저작물 작성을 포함한 활용은 별도 문의 필요라는 조건이 개별 컬렉션 화면에서 확인된다. 3D 데이터는 열람에 별도 상용 소프트웨어가 필요하여 현재 열람에 제한이 있다는 기술적 제공 제약이 컬렉션 설명에 명시되어 있다.

구분	국가기록원	한국영상자료원	국립국악원	국립아시아문화전당
주요 보유 · 관리 자원 유형	공공기관이 생산한 전자기록물, 종이기록물, 시청각기록물, 행정정보 데이터세트	영화필름, 디지털 영화 파일, 영상 · 음향 자료, 관련 기록물	국악 공연 음원 · 영상, 악보, 공연 · 연구 기록	아카이브자료(전문주제자료 · 기증자료), 공연 · 전시 기록, 미디어 자료, 전문도서
자원 획득 · 이관 방식	법적 의무 이관에 따라 생산기관으로부터 단계적으로 이관	수집 · 기증 · 제작 지원 사업 등을 통해 확보	자체 제작 · 조사 · 공연 기록의 디지털화 및 축적	전당 내부 부서로부터 이관 및 외부 수집 · 기증
자원 관리 단위	기록물 단위(건 · 철 · 파일) 기반 관리	작품 · 필름 · 파일 단위 관리	공연 · 자료유형 단위 관리	프로그램 · 아이템 단위로 등록 · 관리
보존체계 및 관리 구조	생산–관리–이관–영구보존으로 이어지는 생애주기 관리 체계 운영	영상자료 보존 · 복원 중심 관리 체계 운영	디지털화 사업 중심의 자료 관리 체계 운영	아카이브통합관리시스템을 중심으로 통합 관리
관리 기준 · 지침	기록관리 기준표, 전자기록물 관리 지침, 보존포맷 선정기준 등 국가 기준 적용	내부 보존 · 복원 기준 및 사업 지침 적용	사업 단위 디지털화 기준 및 내부 관리 기준 적용	「아카이브 관리 규정」, 「아카이브자료 등록 · 기술지침」 적용
메타데이터 구성	국가기록관리 메타데이터 표준 및 NEO 패키지 구조 적용	영상자료 기술 메타데이터 활용	자료유형별 기술 · 관리 정보 메타데이터 활용	등록 메타데이터 영역별 요소 정의(유형, 보존상태, 저작권 등)
디지털화 · 보존 포맷 관리	문서 · 시청각 · 데이터 유형별 보존 · 수용 포맷을 구분하여 관리	필름 디지털 변환 및 영상 포맷 관리 기준 적용	음원 · 영상 · 악보에 대한 디지털화 기준 적용	이미지 · 도서 · 영상 등 유형별 디지털화 세부 기준 명시
무결성 · 품질 관리	전자기록물의 위 · 변조 방지 및 무결성 관리 절차를 공식 지침에 따라 수행	디지털 파일 품질 점검 및 복원 과정 중심 관리	디지털화 결과물의 품질 관리 수행	보존상태 구분 및 관리 이력 기록 수행
관리시스템	기록관리시스템(RMS) 및 중앙 보존 인프라	내부 영상자료 관리 · 복원 시스템	내부 디지털 아카이브 시스템	아카이브통합관리시스템 + 아카이브 누리집
대국민 · 연구자 서비스 연계	공개 가능한 기록물에 대해 검색 · 열람 서비스 제공	영상 자료 상영 · 열람 및 연구 지원	국악 자료 열람 · 교육 · 연구 활용	프로그램 정보와 연계한 아카이브 열람 · 콘텐츠 제공

[표 9] 국내 사례 비교표

2.3. 시사점

앞서 국립중앙도서관의 디지털자원 현황과 관련 시스템 구조, 업무 운영 체계를 분석하고, 국내외 주요 기관의 디지털 장기보존 정책과 기술 적용 사례를 조사・비교・검토하였다. 현황 분석을 통해서는 수집・정리・보존・서비스로 이어지는 디지털자원 관리 체계의 구조적 특징과 한계를 확인하였으며, 사례 분석을 통해서는 장기보존 체계 구축을 위한 국제적 기준과 기술적 적용 방향을 검토하였다.

국립중앙도서관은 납본 제도, OASIS(Online Archiving & Searching Internet Sources) 웹아카이빙, KOLIS 기반 통합관리체계 등 디지털자원 수집과 서비스 측면에서는 국가대표 기관으로서 일정 수준의 기반을 구축하고 있으나, 디지털 장기보존 관점에서 볼 때 체계적 한계와 구조적 제약이 여전히 존재하는 것으로 분석된다. 이러한 한계는 기술적・제도적・조직적 측면에서 복합적으로 나타나며, 이를 해결하기 위한 중장기적 개선 과제가 요구된다.

첫째, 보존 중심이 아닌 서비스 중심으로 설계된 시스템의 구조적 특징이다. 현재 KOLIS는 본질적으로 수집과 대국민 서비스 제공을 목적에 두고 설계된 시스템으로 장기보존을 전제로 한 AIP(Archival Information Package) 관리, 보존 이력 추적, 무결성 자동 검증, 마이그레이션(Migration) 이력 관리와 같은 핵심 보존 기능과는 상반된 목적의 시스템이다. 이에 따라 향후에는 KOLIS와 분리된 장기보존 전용 리포지터리(TDR) 구축 또는 OAIS(Open Archival Information System) 기반 보존 시스템의 별도 도입이 필수적인 과제로 도출된다.

둘째, 보존용 데이터와 서비스용 데이터의 구분을 포함한 다양한 목적에서 생산되는 파생 데이터의 관리 체계 미비 또한 중요한 한계로 나타난다. 현재는 이들 전체가 각 개체별로 명확히 분리, 식별되는 구조로 관리되지 못하고 있으며, 이는 장기적 관점에서 데이터의 진본성 유지와 보존 안정성 확보에 취약 요인으로 작용한다. 이에 따라 향후에는 다양한 파생 데이터들을 명확하게 구분할 수 있는 다면적 데이터 구조 및 메타데이터 관리체계가 요구된다.

셋째, 장기보존 전용 스토리지와 통합적인 저장 정책의 부재 역시 한계로 지적된다. 현재 스토리지는 예산 상황에 따라 지속적으로 증설되고 있으나, 운영용 스토리지와 보존용 스토리지가 논리적으로 분리되지 않은 상태이며, 장기보존을 전제로 한 이중화・지리적 분산・재해복구(DR) 정책도 아직 초기 검토 단계에 머물러 있는 것으로 보여진다. 이는 지속적으로 자원이 증가하는 상황에서 현행 저장 구조는 장기적 지속가능성을 담보하기에 한계가 있다. 따라서 향후에는 다양한 포맷의 파일, 디지털 원본 자료 등을 안정적으로 수용할 수 있는 장기보존

전용 스토리지 인프라 구축과 함께, 무결성 자동 검증과 재해복구 체계가 연계된 통합 저장 정책 수립이 시급한 과제로 제시된다.

넷째, 업무 분절에 따른 전주기 통합 관리의 한계가 구조적 문제로 작용하고 있다. 현재 국립중앙도서관의 디지털자원 관련 업무는 수집, 제작, 메타데이터 구축, 보존, 서비스 기능이 부서별로 분산되어 운영되고 있으며, 이로 인해 입수-정제-등록-보존-서비스로 이어지는 전주기 데이터 흐름이 단절적으로 관리되고 있다. 이러한 분절 구조는 장기보존을 위한 책임 주체의 불명확성, 데이터 이력 관리의 불연속성, 표준 적용의 비일관성, 연관 데이터들 간의 상관관계 단절 등으로 이어질 위험이 크다. 이에 따라 향후에는 자료보존연구센터 또는 보존 전담 조직을 중심으로 한 통합 거버넌스 체계 구축, 그리고 부서 간 역할과 책임을 명확히 규정한 전주기 관리 체계 수립이 필요하다.

다섯째, 시스템적인 측면에서 여전히 서비스 중심의 시스템 구조 안에서 보존 체계가 관리되고 있으며, 장기보존을 전제로 한 보존 전담 아키텍처는 충분히 분리·독립되어 있지 않은 것으로 나타난다. KOLIS는 수집-등록-서비스 업무를 통합적으로 지원하는 데에는 효과적이나, 장기보존 패키지(AIP) 단위로 독립적인 보존 이력·무결성·이벤트를 관리하는 구조에는 한계가 있다. 이는 디지털자원의 수명 주기를 전주기적으로 관리하는 체계로 확장하는 데 구조적 제약 요인으로 작용한다.

여섯째, 보존 기술 전략과 표준 적용의 미흡 또한 중요한 개선 과제로 도출된다. 현재 PREMIS(Preservation Metadata: Implementation Strategies), METS(Metadata Encoding and Transmission Standard) 등 표준이 고려되고 있으나, 실제 운영에서는 보존 행위, 무결성 검증(Fixity), 이벤트 관리, 포맷 변환, 마이그레이션 이력 관리, 포맷 위험도 평가와 같은 핵심 보존 기능이 충분히 구현되지 못하고 있다. 또한 마이그레이션과 에뮬레이션과 같은 적극적 장기보존 기술에 대한 필요성 인식은 존재하지만, 이를 제도와 시스템 차원에서 구현할 구체적인 실행 체계는 아직 부재한 상황이다. 따라서 향후에는 OAIS(Open Archival Information System)-PREMIS(Preservation Metadata: Implementation Strategies)-ISO 16363 기반의 보존 기술 표준을 실질적으로 적용하고, 자동화된 보존 워크플로를 구축하는 것이 핵심 과제로 설정되어야 한다.

일곱째, 저작권 제약으로 인한 보존 및 활용의 구조적 한계도 지속적으로 제기된다. 현재 많은 디지털자료와 웹 수집 자료는 저작권 문제로 인해 관내 열람 중심의 제한적 이용 구조를 유지하고 있으며, 대국민 서비스와 AI 학습 데이터 활용에도 상당한 제약이 존재한다. 이는 보존 자체뿐 아니라 보존 자원의 활용이라는 측면에서도 국립중앙도서관의 공공적 기능을 제한하는 요인으로 작용한다.

따라서 향후에는 디지털 납본 자료와 웹자원에 대한 저작권 처리 기준의 명확화, 장기보존 목적 이용에 대한 법적 근거 강화, 디지털자원 활용 범위의 제도적 정비가 병행되어야 할 필요가 있다.

마지막으로 AI 및 신기술 활용을 위한 데이터 준비 수준의 한계도 중요한 과제로 부각된다. 국립중앙도서관이 보유한 웹자원과 디지털 콘텐츠는 AI 학습 자원으로서 잠재적 가치가 매우 높지만, 현재는 저작권 문제, 메타데이터 품질의 비정형성, 표준화 부족 등의 이유로 실질적인 활용이 제한되고 있다. 이에 따라 향후에는 AI 기반 메타데이터 자동생성, 자동 분류・색인 (KDC/LCC/DDC・시소러스), 포맷 분석・보존 포맷 변환・위험 탐지, AI 기반 복원 기술을 장기보존 체계와 연계하여 단계적으로 도입하는 전략이 요구된다.

종합하면, 국립중앙도서관의 디지털 장기보존체계는 수집과 서비스 중심 구조에서 출발하여 점진적으로 보존 기능을 확대해 온 단계적 발전 과정에 있으나, 현재 단계에서는 서비스 중심 시스템과 보존 중심 시스템 간의 협업과 연계 구조에 대한 정의 부재, 저장소 구조의 한계, 업무 분절, 표준 적용 미흡, 저작권 제약, 신기술 연계의 한계라는 복합적인 구조적 문제를 내포하고 있다. 따라서 향후 국립중앙도서관이 국가 차원의 실질적 디지털 장기보존 기관으로 기능하기 위해서는, 기술 인프라 구축과 더불어 거버넌스 개편, 법・제도 정비, 표준 기반 운영 체계 확립, 보존 전문 인력 확충이 통합적으로 추진되어야 할 핵심적인 중장기 개선 과제로 설정될 필요가 있다.

제3장 보존 포맷, 인프라 및 AI 기술 적용방안

3.1. 포맷 레지스트리 기술

3.1.1. 장기보존전략 유형

디지털자원의 장기보존을 위해서는 기술 환경의 변화에도 불구하고 자료의 접근성과 진본성을 유지할 수 있는 전략적 접근이 필요하다. 현재 디지털 장기보존 분야에서는 매체전환(Refreshment), 마이그레이션(Migration), 포맷정규화(Normalization), 캡슐화(Encapsulation), 에뮬레이션(Emulation)과 같은 대표적인 보존전략이 활용되고 있으며, 각 전략은 적용 대상 자료의 특성과 보존 목적에 따라 상이한 장단점을 지닌다.

❑ 매체전환(Refreshment) 전략

매체전환(Refreshment)은 디지털자원의 장기보존 과정에서 저장매체의 물리적 노후화나 기술적 구식화에 대응하기 위해, 디지털 정보를 동일한 형식의 새로운 저장매체로 옮겨 담는 전략이다. 이 전략의 핵심은 데이터의 비트스트림(bitstream) 자체를 변경하지 않고 그대로 유지한 채, 물리적 저장 환경만을 교체하는 데 있다. 그러므로 파일 형식이나 논리 구조를 변환하지 않으며, 디지털 객체의 내용은 원형 그대로 보존된다.

매체전환(Refreshment)은 데이터를 한 유형의 저장매체에서 동일한 유형의 다른 저장매체로 이전하는 과정으로 정의된다. 예를 들어, 장기간 사용되어 열화 위험이 있는 CD-ROM에 저장된 데이터를 새로운 CD-ROM으로 복사하는 행위는 대표적인 매체전환(Refreshment) 사례에 해당한다. 이러한 점에서 매체전환(Refreshment)은 데이터 형식 자체를 새로운 기술 환경에 맞게 변환하는 마이그레이션(Migration) 전략과 명확히 구별되며, 오직 물리적 저장매체의 문제 해결에 중점을 둔다.

매체전환(Refreshment)의 주요 기능과 목적은 첫째, 저장매체의 노후화와 열화에 따른 데이터 손실 위험을 예방하는 데 있다. 자기 테이프나 광디스크와 같은 물리적 매체는 시간이 경과함에 따라 성능 저하나 손상이 발생할 수 있으며, 주기적인 매체전환(Refreshment)은 이러한 물리적 실패(media failure)로 인한 데이터 손실 가능성을 낮추는 역할을 수행한다. 둘째, 데이터를 읽는 데 필요한 특정 저장 장치가 더 이상 생산되거나 지원되지 않아 접근이 불가능해지는 저장

장치 차원의 기술적 구식화에 대비하는 기능을 가진다.

또한 매체전환(Refreshment) 과정에서는 데이터를 새로운 매체로 옮기는 동시에 체크섬(Checksum) 검사와 같은 정합성 검사가 수행된다. 이를 통해 파일이 손상되지 않았음을 확인할 수 있으며, 만약 손상이 발견될 경우 백업본을 활용하여 복원할 수 있는 기회를 제공한다. 이러한 절차는 디지털자원의 무결성을 지속적으로 확인하고 유지하는 데 중요한 역할을 한다.

매체전환(Refreshment)는 장기적인 디지털 보존 정책에서 지속적인 접근성 유지를 위한 기본 전략으로 간주된다. 이는 데이터 자체의 유용성을 유지하기 위한 마이그레이션 전략과 병행되어 사용되며, 디지털 객체의 논리적·의미적 보존을 담당하는 다른 전략들의 전제 조건으로 기능한다. 다시 말해, 매체전환(Refreshment)는 저장매체 수준에서 데이터가 안정적으로 존재할 수 있도록 하는 기초적인 유지보수 활동으로서 보다 복잡한 장기보존전략이 효과적으로 수행될 수 있는 기반을 제공한다.

종합적으로 볼 때, 매체전환(Refreshment)는 디지털 아카이브에서 데이터 비트 수준 보존(Bit-Level Preservation)을 담당하는 핵심 전략이다. 이는 디지털 정보를 기술 환경 변화 속에서도 살아있는 상태로 유지하는 데 기여하며, 디지털 장기보존체계에서 필수적인 유지관리 전략이다.

❑ 마이그레이션(Migration) 전략

마이그레이션(Migration)은 디지털 콘텐츠를 장기간 열람 가능하게 유지하기 위해, 파일 형식·저장매체·플랫폼을 새로운 환경으로 주기적으로 변환하는 보존전략이다. 이 전략은 기술 환경의 변화에 따라 기존 포맷이나 매체가 더 이상 지원되지 않을 경우에도 콘텐츠 접근성을 유지하기 위한 방법으로 활용된다. 예를 들어, 문서 파일을 Word 형식에서 PDF/A로 변환하거나, VHS 영상 자료를 FFV1 코덱과 Matroska 컨테이너(MKV) 형식으로 변환하는 방식, 데이터베이스를 SIARD 형식으로 변환하는 사례 등이 이에 해당한다.

마이그레이션(Migration) 전략의 특징은 포맷, 저장매체 또는 소프트웨어 버전의 변경에 대응하기 위해 주기적으로 수행되어야 한다는 점이다. 이 과정에서는 변환 전후의 무결성을 점검하여, 콘텐츠의 내용이 손실되거나 변형되지 않았는지를 검증하는 절차가 필수적으로 수반된다. 또한 자동화 시스템과의 결합이 용이하다는 특성으로 인해, 대규모 디지털자료를 관리하는 기관에서 비교적 널리 활용되고 있다.

마이그레이션(Migration)의 장점은 변환 도구가 다양한 공급자에 의해 지속적으로

개발·제공되고 있어, 실행을 위한 개발 및 시행 비용이 비교적 낮다는 점이다. 특히 텍스트 기반 디지털자료의 경우, 내용 보존의 신뢰성이 높은 전략으로 평가되며, 장기보존 관리 측면에서 효율적으로 적용될 수 있다. 반면 단점으로는, 원본의 논리적 구조나 렌더링 방식이 변환 과정에서 달라질 수 있어, 반복적인 변환이 누적될 경우 데이터 또는 속성의 일부가 변경될 가능성이 존재한다는 점이 지적된다. 특히 복잡한 레이아웃을 가진 자료의 경우 외형이 손상될 위험이 있으며, 디지털 환경이 변화할 때마다 변환을 반복해야 하므로 장기적으로는 보존비용이 증가할 수 있다.

❑ 포맷정규화(Normalization) 전략

포맷정규화(Normalization)은 디지털자료를 생성 또는 입수 단계에서부터 장기보존에 적합한 표준 형식으로 변환하거나 통일하는 장기보존전략이다. 이는 사후 대응적 보존전략이 아니라 생산 단계에서부터 보존을 고려하는 사전 예방적 접근 방식이라는 점에서 특징적이다. 문서의 경우 PDF/A나 ODF, 이미지의 경우 TIFF나 JPEG, 음원의 경우 WAV나 FLAC, 영상의 경우 FFV1, MJ2, MKV 등과 같은 표준 포맷이 예시로 제시된다.

포맷정규화(Normalization) 전략은 생산 단계(Born-Digital)에서부터 표준 포맷을 적용함으로써, 이후 보존 관리 과정에서 발생할 수 있는 복잡성을 줄이고자 한다. 국가기관의 경우 권장 포맷을 정의하고 이를 활용하고 있다.

이 전략의 장점은 장기보존 관리의 효율성을 높이고, 향후 마이그레이션에 소요되는 비용을 크게 절감할 수 있다는 점이다. 동일 유형의 파일을 표준 포맷으로 유지함으로써 보존과 접근이 용이해지며, 기관 간·시스템 간 교환과 통합이 수월해져 상호운용성이 강화된다. 그러나 변환 과정에서 일부 서식, 구조, 스크립트 정보가 손실될 가능성이 있으며, 3D 모델이나 인터랙티브 미디어와 같이 표준화가 어렵거나 불완전한 자료유형에는 적용이 제한될 수 있다. 또한 모든 자료를 표준 포맷으로 강제할 경우, 디지털 표현의 다양성이 제한될 수 있다는 점 등이 단점으로 지적된다.

자료 유형	보존 포맷	설명	적용 사례	품질 기준
문서	PDF/A-1b	기본 가시성 보존 중심	한국 국가기록원 NEO3	• 폰트 내장 (embedded) • 무손실 압축 • DRM, 암호화, 멀티미디어 객체 삽입 금지
	PDF/A-2u	유니코드 문자 포함 (검색 가능 문서용)	EU Publications Office	
	PDF/A-4	최신 XML 기반 표준	Library of Congress (PDF/A 및 XML 기반 문서 병행)	
	ODF (OpenDocument Format)	오픈오피스 및 리브레오피스 기반의 개방형 문서표준.	ISO/IEC 26300	
텍스트	TXT / XML (UTF-8)	구조화되지 않은 텍스트나 메타데이터 중심 기록물은 단순 텍스트 포맷으로 장기보존	미국 의회도서관(LC)	• UTF-8 인코딩
이미지	TIFF 6.0	장기보존의 국제표준 포맷	NARA(National Archives, U.S.)	• 문서형 이미지 300 ~600 dpi 이상 / 사진류 3000 픽셀 이상 • 흑백 8bit, 컬러 24bit (RGB) 또는 48bit (16bit×3) • 무손실 압축
	JPEG 2000	무손실 또는 고품질 압축 이미지	NARA(National Archives, U.S.) Europeana	
영상	FFV1	FFmpeg Video Codec v1, lossless 오픈소스 무손실 코덱	한국영상자료원 Europeana	• 해상도 : 1920×1080 (HD) 이상 / 보존용은 원본 해상도 유지 • 프레임 : 24 fps (필름) ~ 30 fps (방송) • 비트깊이 : 10-12 bit YUV 4:2:2 이상 • 오디오 : 48 kHz / 24 bit PCM
	Motion JPEG 2000	각 프레임이 독립적으로 압축되어 편집 및 복원 용이	NARA & LC	
	MKV	Matroska 오픈형 멀티미디어 컨테이너	Europeana	
	MXF	Material Exchange Format, SMPTE 377M 방송 · 영화용 프로페셔널 규격	NARA & LC	
음원	WAV	Linear PCM 비압축 PCM 16bit 또는 24bit	한국 국가기록원(24bit) LC(기본)	• 샘플링레이트: 48 kHz (표준), 96 kHz (고음질) • 비트깊이: 24 bit • 채널: 모노(1ch) 또는 스테레오(2ch)
	FLAC	무손실 압축포맷	LC(보조)	

[표 10] 포맷 표준 형식

❑ 캡슐화(Encapsulation) 전략

캡슐화(Encapsulation)은 디지털 콘텐츠 자체뿐 아니라 해당 콘텐츠와 관련된 메타데이터, 실행 및 문맥 정보, 기술 문서 등을 하나의 정보 패키지로 묶어 저장하는 장기보존전략이다. 이는 단순히 파일을 보관하는 수준을 넘어, 파일이 언제, 어떤 조건에서 누가, 어떤 방식으로 생성·보존되었는지를 함께 기록하는 데

목적이 있다. OAIS(Open Archival Information System) 참조모델의 AIP(Archival Information Package), BagIt과 METS(Metadata Encoding and Transmission Standard)/PREMIS(Preservation Metadata: Implementation Strategies)의 결합 구조, 호주 빅토리아주의 VEO, 한국 국가기록원의 NEO 등이 대표적인 사례로 제시된다.

캡슐화(Encapsulation) 전략의 특징은 PREMIS(Preservation Metadata: Implementation Strategies)의 object, event, agent, rights 요소와 METS(Metadata Encoding and Transmission Standard)의 구조적 요소(fileSec, amdSec 등)를 결합하여, 완전한 정보 패키지 형태의 보존 단위를 구성한다는 점이다. 이를 통해 콘텐츠와 그에 대한 관리·이력 정보가 통합적으로 관리된다.

이 전략의 장점은 콘텐츠, 메타데이터, 관리 정보가 하나의 패키지로 통합되어 있어 이동·복제·검증이 용이하다는 점이다. 또한 디지털자료의 의미, 출처, 저작권, 변환 이력 등을 함께 보존할 수 있어, 신뢰할 수 있는 보존 증거로 활용 가능하다. METS(Metadata Encoding and Transmission Standard), PREMIS(Preservation Metadata: Implementation Strategies), 더블린 코어(Dublin Core), MODS(Metadata Object Description Schema) 등 국제 표준 기반 구조를 활용하기 때문에 시스템 간 이식성과 상호운용성이 높으며, 패키지 단위로 체크섬을 관리하여 변조나 손실을 방지할 수 있다는 점도 중요한 장점이다. 반면, 초기 구축 과정이 복잡하고 파일 크기가 증가할 수 있으며, 표준 스키마에 대한 충분한 이해와 기술적 지원이 요구된다. 또한 실행환경 보존이 필요한 경우에는 에뮬레이션 전략과 병행해야 완전한 복원이 가능하다는 한계가 있다.

❑ 에뮬레이션(Emulation) 전략

에뮬레이션(Emulation)은 더 이상 사용되지 않는 운영체제, 소프트웨어, 하드웨어 환경을 가상으로 재현하여 원본 파일을 그대로 실행하는 방식의 장기보존전략이다. 이 전략은 과거의 실행환경을 소프트웨어적으로 모사함으로써, 원본 디지털자료를 당시의 환경과 동일한 조건에서 열람할 수 있도록 하는 데 목적이 있다. 고전 게임이나 디지털 아트, Flash 기반 콘텐츠 웹에서 구동 및 이미 단종된 Word Perfect 편집프로그램을 실행하기 위한 MS-DOS 환경을 재구성 등의 사례가 이에 해당한다.

에뮬레이션(Emulation) 전략은 응용프로그램 실행 파일, 인터랙티브 미디어, 디지털 예술작품, 구형 게임 등 특정 실행환경에 강하게 의존하는 자료에 특히 적합하다. 하드웨어, 운영체제, 응용프로그램을 가상화함으로써 원본 환경을 재현한다는 점이 핵심적인 특징이다.

이 전략의 가장 큰 장점은 원본 데이터의 ‘외형과 느낌(look and feel)’을 그대로 재현할 수 있다는 점이다. 원본 디지털자료를 변환하지 않고 그대로 보존할 수 있어, 자료의 진본성을 가장 정확하게 유지할 수 있으며, 마이그레이션과 달리 시간의 흐름에 따라 반복적인 변환을 수행할 필요가 없다. 또한 자료유형별로 일관된 보존전략을 수립할 수 있다는 장점도 있다. 반면, 에뮬레이터 소프트웨어를 개발·유지하는 데 고도의 기술력과 상당한 비용이 요구되며, 지적재산권과 저작권 문제가 발생할 가능성이 있어 장기보존전략의 효율성과 안정성을 저해할 수 있다. 더불어, 다양한 시스템과 소프트웨어 환경에 대응하기 위해 지속적인 에뮬레이터 개발이 필요하다는 점도 한계로 지적된다. 이와 함께, 에뮬레이션을 통해 기록을 재생산하는 데 필요한 절차를 기술하는 메타데이터가 추가적으로 생산·관리되어야 한다.

❑ 디지털 장기보존전략 간 비교 분석

디지털 장기보존전략은 단일한 기술적 해법이 아니라 보존 대상의 성격과 보존 목적, 기술 환경 변화의 양상에 따라 상호 보완적으로 적용되는 다층적 접근 방식으로 이해되어야 한다. 포맷정규화(Normalization), 매체전환(Refreshment), 마이그레이션(Migration), 에뮬레이션(Emulation), 캡슐화(Encapsulation)은 모두 장기보존을 목표로 하지만, 개입 수준과 보존 대상, 그리고 진본성 유지 방식에서 본질적인 차이를 지닌다.

○ 보존 개입 수준과 대상의 차이

매체전환(Refreshment)는 다섯 가지 전략 가운데 가장 낮은 수준의 개입 전략으로 디지털 객체의 비트스트림을 변경하지 않고 저장매체만을 교체한다는 점에서 보존 개입이 물리적 수준에 한정된다. 이 전략은 데이터의 논리적 구조나 표현 방식에는 전혀 영향을 미치지 않으며, 저장매체의 노후화와 장치 차원의 기술적 구식화에 대응하는 데 초점을 둔다.

반면 마이그레이션(Migration)과 포맷정규화(Normalization)는 파일 형식 수준에서의 개입을 수반한다. 마이그레이션(Migration)은 시간의 흐름에 따라 포맷·플랫폼·소프트웨어 환경 변화에 대응하여 주기적으로 파일을 변환하는 전략인 반면, 포맷정규화(Normalization)는 생성 또는 입수 단계에서부터 장기보존에 적합한 표준 포맷으로 통일하는 사전 예방적 전략이라는 점에서 차이를 보인다. 두 전략 모두 데이터 형식의 변경을 전제로 하며, 물리적 저장매체보다는 논리적

접근성과 관리 효율성에 초점을 둔다.

캡슐화(Encapsulation)와 에뮬레이션(Emulation)은 보다 상위 수준의 보존 개입 전략으로 볼 수 있다. 캡슐화(Encapsulation)는 디지털 콘텐츠와 함께 메타데이터, 관리 정보, 기술 문서를 하나의 정보 패키지로 통합하여 보존함으로써, 파일 자체뿐 아니라 그 맥락과 관리 이력을 함께 유지하는 데 중점을 둔다. 에뮬레이션(Emulation)은 실행환경 자체를 재현하여 원본 파일을 변형 없이 실행하는 전략으로 개별 파일이 아니라 환경 전체를 보존 대상으로 확장한다는 점에서 가장 포괄적인 개입 방식을 취한다.

○ 진본성 및 원본성 유지 방식의 비교

진본성 유지 측면에서 매체전환(Refreshment)은 비트 수준 보존(bit-level preservation)을 통해 데이터의 동일성을 유지하지만, 논리적 접근성이나 표현 방식의 재현까지는 보장하지 않는다. 마이그레이션(Migration)과 포맷정규화(Normalization)는 접근성과 관리 효율성을 확보하는 데 효과적이지만, 반복적인 변환 과정에서 원본의 논리 구조나 외형이 일부 변경될 가능성을 내포하고 있어, 진본성 손상 위험이 존재한다.

이에 비해 에뮬레이션(Emulation)은 원본 데이터와 실행환경을 그대로 재현함으로써, 디지털자료의 외형과 느낌을 가장 충실하게 유지할 수 있는 전략으로 평가된다. 캡슐화(Encapsulation) 역시 콘텐츠와 함께 출처, 저작권, 변환 이력, 관리 정보를 통합적으로 보존함으로써, 보존 과정 전반의 신뢰성과 증거성을 강화하는 역할을 수행한다. 다만 캡슐화(Encapsulation)는 실행환경 자체를 보존하지 않기 때문에, 원본 실행 경험의 완전한 재현을 위해서는 에뮬레이션(Emulation)과의 병행이 요구된다.

○ 운영비용과 기술적 난이도의 차이

운영비용과 기술적 난이도 측면에서 매체전환(Refreshment)는 비교적 단순한 절차와 낮은 기술 요구 수준을 가지며, 장기보존체계에서 기본적인 유지보수 전략으로 기능한다. 마이그레이션(Migration) 역시 자동화 시스템과 결합하기 용이하고 변환 도구가 지속적으로 제공되기 때문에, 대규모 디지털자료 관리에 널리 활용될 수 있다. 그러나 장기적으로는 반복적인 변환이 요구되므로 누적 비용이 증가할 가능성이 있다.

포맷정규화(Normalization)은 초기 단계에서 표준 포맷 정책을 수립하고 이를 적용해야 하므로 제도적·정책적 준비가 필요하지만, 일단 정착될 경우 이후 관리 비용을 크게 절감할 수 있다는 장점을 가진다. 반면 캡슐화(Encapsulation)은 METS(Metadata Encoding and Transmission Standard), PREMIS(Preservation Metadata: Implementation Strategies) 등 표준 스키마에 대한 이해와 시스템 구축이 필요하여 초기 도입 비용과

기술적 복잡성이 상대적으로 크다. 에뮬레이션(Emulation)은 에뮬레이터 개발과 유지에 고도의 기술력과 상당한 비용이 요구되며, 다양한 하드웨어·소프트웨어 환경에 대응해야 한다는 점에서 가장 높은 기술적 부담을 수반한다.

○ 전략 간 상호 보완적 관계

이들 전략은 상호 배타적인 대안이라기보다, 서로 다른 보존 수준을 담당하는 상호 보완적 전략으로 이해될 필요가 있다. 매체전환(Refreshment)는 모든 장기보존전략의 전제 조건으로서 데이터의 물리적 생존을 보장하며, 마이그레이션(Migration)과 포맷정규화(Normalization)은 논리적 접근성과 관리 효율성을 확보하는 역할을 수행한다. 캡슐화(Encapsulation)은 보존 행위와 맥락을 구조적으로 기록·유지함으로써 장기보존의 신뢰성을 강화하고, 에뮬레이션(Emulation)은 특정 자료유형에 대해 원본 실행 경험을 유지하는 최종적인 보존 수단으로 기능한다.

구분	매체전환(Refreshment)	마이그레이션(Migration)	포맷정규화(Normalization)	캡슐화(Encapsulation)	에뮬레이션(Emulation)
보존 개입 수준	물리적 저장매체 수준	파일 형식 · 플랫폼 수준	파일 형식 수준 (사전 예방)	정보 패키지 수준	실행환경 수준
핵심 목적	저장매체 노후화 · 물리적 실패 대응	기술 환경 변화에 따른 접근성 유지	장기보존에 적합한 표준 형식 유지	콘텐츠와 맥락 · 이력의 통합 보존	원본 실행환경과 경험 재현
데이터 변경 여부	변경 없음 (비트스트림 유지)	있음 (포맷 · 구조 변환)	있음 (표준 포맷으로 변환)	없음(파일 자체는 유지)	없음(원본 파일 유지)
주요 적용 시점	주기적 유지관리 단계	기술 환경 변화 시점마다 반복	생성 · 입수 단계	보존 패키지 생성 단계	접근 · 활용 단계
대표 예시	CD-ROM → 새 CD-ROM	Word → PDF/A, VHS → FFV1	PDF/A, TIFF, WAV 등	OAIS AIP, METS+PREMIS, NEO	구형 OS · 소프트웨어 가상화
접근성 유지 방식	매체 교체를 통한 물리적 접근 유지	최신 환경에서 직접 열람 가능	표준 포맷 기반 직접 열람	패키지 기반 관리 · 이식	원본 환경에서 실행
진본성 (원본성) 영향	매우 낮음	변환 과정에서 손상 가능	일부 정보 손실 가능	출처 · 이력 보존으로 신뢰성 강화	외형 · 느낌까지 가장 정확히 유지
무결성 관리	체크섬 검증 필수	변환 전 · 후 무결성 검증 필수	변환 시 무결성 검증 필요	패키지 단위 체크섬 관리	실행환경 재현 중심
자동화 용이성	높음	높음	중간	중간	낮음
기술적 난이도	낮음	낮음~중간	중간	중간~높음	매우 높음
운영비용 특성	낮음(지속적 유지관리)	반복 수행 시 누적 증가	초기 비용 후 관리 효율 높음	초기 구축비용 높음	개발 · 유지비용 매우 높음
적합한 자료유형	모든 디지털자료	텍스트 · 영상 · DB 등	정형 자료(문서·이미지 등)	모든 유형	게임 · 디지털아트 · 인터랙티브
전략적 위치	비트 수준 보존의 기초 전략	접근성 유지 핵심 전략	사전 예방적 보존전략	신뢰성 · 증거성 강화 전략	최종적 · 특수 목적 전략

[표 11] 디지털 장기보존전략 간 비교 분석

3.1.2. 장기보존전략(에뮬레이션) 적용 사례

디지털자료가 지닌 본래의 기능적 속성 및 원래의 0과 1의 비트스트림을 유지하며 원래의 모습을 그대로 유지할 수 있는 방안은 디지털자료 장기보존을 위한 핵심이며, 이를 실현하기 위해서는 에뮬레이션 전략을 적극적으로 적용할 필요가 있다.

에뮬레이션(Emulation)에 대하여 JISC(Joint Information Systems Committee)에서는 '구형 컴퓨터 환경을 재현하고, 오래되고 쓸모없게 된 파일들을 실행시키기 위해 현재의 하드웨어와 소프트웨어를 사용하는 것'이라고 정의하고 있으며, The LIFE(Lifecycle Information for E-Literature)에서는 구형의 운영체제 또는 소프트 또는 소프트웨어의 조각을 최신의 운영체제 또는 소프트웨어 조각에 재생산하여 구형 시스템과 관련된 디지털 객체에 대한 접근을 허용하는 과정', DPT(Digital Preservation Testbed)에서는 '원본을 이용하는 것과 동일한 효과를 내기 위해 다른 컴퓨터에 어떤 장치나 프로그램을 이용하는 것과 관련되는 다양한 모든 기법을 말하며, 보존을 위해 미래의 컴퓨터상에서 구형의 노화된 컴퓨터를 재생산하고, 원본의 노화된 소프트웨어가 미래에 작동되게 하여, 미래 컴퓨터가 노화된 컴퓨터를 구현하도록 하는 것'이라고 정의하고 있다.

디지털자료의 장기보존전략 중 에뮬레이션(Emulation)을 적용한 대표적인 사례는 미국의 Internet Archive(archive.org), 호주의 AusEaaSI(auseaasi.org), MAME 프로젝트(mamedev.org), KVM + QEMU 기반 EaaS, 영국 TNA 디지털 기록 재현(Replay) 프로젝트, 독일 DNB EaaS 프로젝트 등을 통하여 확인할 수 있다.

❑ 미국의 Internet Archive(archive.org) : 소프트웨어 라이브러리 & 에뮬레이션 프레임워크

미국의 Internet Archive(이하 IA)는 텍스트, 영상, 소프트웨어, 음원, 웹사이트 등 수억 건에 이르는 디지털 객체를 수집·보존하는 비영리 디지털 도서관으로 디지털 문화유산의 장기적 보존과 접근 제공을 주요 목표로 하는 대표적인 국제 사례이다. IA는 단순한 디지털 파일 보존을 넘어, 소프트웨어 라이브러리(Software Library)와 에뮬레이션 프레임워크(Emulation Framework)를 통해 과거 디지털 환경을 재현하고, 이용자가 웹 브라우저 환경에서 직접 실행 가능한 접근 경험을 제공하고 있다.

IA의 소프트웨어 라이브러리는 초창기 개인용 컴퓨터 환경에서 사용되던 다양한 소프트웨어와 게임, 애플리케이션을 포함하며, 플로피 디스크나 카세트 테이프 기반의 디지털 객체까지 폭넓게 수집·관리한다. 이러한 자료는 DOSBox, MAME 등과 같은 브라우저 기반 에뮬레이터를 통해 제공되며, 이용자는 별도의 프로그램 설치 없이 웹 브라우저만으로 해당 소프트웨어를 실행할 수 있다. 이와 같은 접근 방식은 초창기 PC, 맥(Mac) 환경을 포함한 다양한 플랫폼과 운영체제를 포괄적으로 지원한다는 점에서 특징적이다.

IA의 에뮬레이션 프레임워크는 개별 에뮬레이터를 단순히 나열하는 방식이 아니라 에뮬레이션 환경을 각 디지털 객체의 컬렉션 아이템 페이지와 통합하여 제공한다. 이를 통해 하나의 화면 안에서 메타데이터, 원본 파일, 그리고 에뮬레이터가 결합된 형태의 접근 경험이 구현되며, 이용자는 디지털 객체의 설명과 맥락을 확인하는 동시에 실제 실행 결과를 체험할 수 있다. 이러한 통합적 제공 방식은 에뮬레이션을 단순한 기술적 실험이 아닌 디지털 컬렉션 서비스의 핵심 구성 요소로 자리매김하게 한다.

메타데이터 구조 측면에서 IA는 DC(Dublin Core)를 기본 구조로 유지하면서도, 대규모·이질적 디지털 객체를 효율적으로 관리하기 위해 확장 가능한 Key-Value 기반의 유연한 메타데이터 설계를 채택하고 있다. IA의 메타데이터는 크게 세 가지 유형으로 구분된다. 첫째, 시스템 예약(System-Reserved) 필드는 아카이브 시스템이 내부적으로 인식하거나 아이템 관리에 필수적으로 요구하는 주요 메타데이터 항목으로 구성된다.

IA 시스템 예약 필드	설명	DC 매핑	PREMIS 구분	PREMIS 상세
identifier	자료의 디렉토리를 참조하기 위해 사용되는 식별자	Identifier	Object	objectIdentifier
title	상세 페이지에서 표시되는 아이템의 설명적인 제목	Title	Object	objectCharacteristicsExtension
creator	아이템의 내용에 대해 목록화된 저자 또는 제작자	Creator	Agent	Producer / Creator
collection	아이템이 파일로 묶이는 위치	Relation (isPartOf)	Object	relationshipType
search collection	“가상” 컬렉션을 생성하는 검색 질의어		Rights	rightsGranted / restriction
mediatype	movies, texts, audio, software, data, web 등이 있으며, 각각 세부 정보 페이지에서 자동으로 표시되는 방식이 다름	Type	Object	objectCategory

IA 시스템 예약 필드	설명	DC 매핑	PREMIS 구분	PREMIS 상세
date	아이템의 공식적인 날짜로 영화의 개봉일, 책의 발행일, 소프트웨어의 제작일 등이 해당함	Date	Event	eventDateTime (create)
subject	아이템에 대한 키워드 역할을 수행하는 용어(들)로 아이템에 포함된 콘텐츠 유형, 아이템에 등장하는 인물 등이 해당함	Subject	Object	objectCharacteristics Extension
licenseurl	아이템에 적용된 라이선스에 대한 설명 URL		Rights	rightsBasis
pick	아이템이 포함된 컬렉션에서 pick인지 여부(대부분 더 이상 사용되지 않음)			
noindex	아이템이 검색 엔진/검색 결과에 표시되는지 여부		Rights	rightsGranted / restriction
hidden	noindex 중 더 이상 사용되지 않는 버전으로 향후 병합/제거될 가능성이 있음		Rights	rightsGranted / restriction
publicdate	아이템이 발행된 날짜(읽기 전용) – 아카이브 내부 스캐닝 작업의 일부인 경우에는 addeddate와 다름		Event	eventDateTime (publish)
publisher	아이템 또는 아이템이 생성된 원본 아티팩트의 알려진 발행자	Publisher	Agent	organization
addeddate	아이템 식별자가 IA에 추가되고 예약된 날짜와 시간(읽기 전용)		Event	eventDateTime (ingest)
adder	IA의 내부 스캐닝 작업에서 항목 식별자를 DB에 추가한 직원 또는 계약자		Agent	individual / contractor
updater	IA의 내부 스캐닝 작업에서 아이템의 정보를 마지막으로 갱신한 직원 또는 계약자		Agent	individual / contractor
updatedate	IA의 내부 스캐닝 작업에서 아이템이 마지막으로 갱신된 날짜		Event	eventDateTime (modify)
uploader	아카이브에 아이템을 처음 업로드한 계정		Agent	individual / contractor
notes	아이템의 품질, 추가된 날짜, 아이템 자체와 관련된 요소 또는 고려 사항에 대한 설명을 위한 추가 정보 요소		Object	objectCharacteristicsNote
rights	licenseurl이 언급된 아이템에 대한 추가 권한 및 고려사항을 가진 사람 또는 조직에 대해 알려진 정보	Rights	Rights	rightsBasis
contributor	도서관, 외부 아카이브 또는 아이템을 제공한 기금/개인과 같이 아이템을 후원하는 기관 또는 출처	Contributor	Agent	organization
language	아이템에 대한 공식적인 언어	Language	Object	objectCharacteristics/formatLanguage
coverage	해당 아이템이 지리적 또는 국가적 측면에서 포괄하는 범위	Coverage	Object	objectCharacteristics/coverageNote

[표 12] Internet Archive 시스템 예약 필드 (System-Reserved)

둘째, 매직 필드(magic fields)는 일반적인 메타데이터 항목과 달리, 특정 키가 설정될 경우 아이템의 처리 방식에 영향을 미치거나 자동화 스크립트와 연동되는 기능적 성격의 메타데이터이다.

매직 필드	설명
mobygames	ModyGames(mobygames.com)의 URL을 제공하면 내부 스크립트가 실행될 때마다 해당 URL에서 다른 메타데이터가 자동으로 채워짐
isbn	ISBN 번호를 입력하면 MARC 및 기타 도서 메타데이터 소스에서 수집한 정보가 항목에 채워짐
ppi-direct	ppi-direct 메타데이터 쌍을 설정하면 소프트웨어가 자료를 보고 결정하는 값 대신 dpi 재정의가 제공됨
reviews-allowed	항목에 새로운 리뷰가 추가되지 않도록 고정으로 설정함

[표 13] Internet Archive 매직 필드 (Magic)

셋째, 기타(Everything Else) 필드는 시스템 예약 필드나 매직 필드에 해당하지 않는 모든 메타데이터를 포함하며, 개별 아이템의 특성에 따라 다양한 메타데이터 요소를 자유롭게 추가할 수 있도록 설계되어 있다.

매체 유형	공통 항목	고유 항목
Text	Title, Publisher, Description, Publication date, Item Size, Addeddate, Identifier, Scanner	Access-restricted-item, Firstfiledate, Lastfiledate, Pages, Scanneddate
Movie	Title, Publisher, Description, Topics, Language, Item Size, Contact Information, Addeddate, Collection_added, Idenfifier, Scanner	Color, Sound
Audio	Title, Publisher, Description, Publication date, Usage(이용허락), Topics, Language, Item Size, Description, Addeddate, Call number, Collection_added, Identifier	Run time
Software	Title, Publisher, Publication date, Topics, Item Size, Description, Addeddate, Collection_added, Identifier, Scanner	Emulator, Emulator_ext, Pce_drive_fd1, Pce_drive_fd2, Year
Images	Title, Publisher, Usage(이용허락), Topics, Item Size, Description, Addeddate, Collection_added, Identifier, Scanner	

[표 14] Internet Archive 매체유형별 메타데이터 항목

또한 IA는 매체 유형별 특성을 고려하여 메타데이터 항목을 공통 항목과 매체 유형별 고유 항목으로 구분하여 관리한다. 이를 통해 텍스트, 소프트웨어, 게임, 음원 등 서로 다른 디지털 객체 유형을 하나의 통합 아카이브 환경에서 유연하게 수용하면서도, 각 매체의 특수성을 반영한 기술적·내용적 설명을 가능하게 한다.

이와 같은 IA의 소프트웨어 라이브러리와 에뮬레이션 프레임워크 사례는, 실행형 디지털 객체를 단순히 파일 단위로 보존하는 것을 넘어, 웹 기반 에뮬레

이션을 통해 실제 이용 경험까지 포함한 장기 접근을 실현하고 있다는 점에서 중요한 시사점을 제공한다. 이는 디지털 장기보존에서 에뮬레이션 전략이 기술적 선택을 넘어, 이용자 접근성과 컬렉션 서비스 설계 전반과 밀접하게 결합될 수 있음을 보여주는 대표적인 사례라 할 수 있다.

IA의 소프트웨어 라이브러리 및 에뮬레이션 프레임워크 운영 사례는, 실행형 디지털 객체를 포함한 대규모 이종 디지털자원을 장기적으로 관리하기 위해 유연하면서도 구조적인 메타데이터 설계 전략이 필요함을 보여준다. IA는 기본적으로 DC(Dublin Core)를 기반으로 메타데이터 요소를 설계하되, 모든 디지털 객체를 동일한 고정 스키마로 제한하지 않고 필요에 따라 확장 가능한 유연한 구조를 적용하는 방식을 채택하고 있다.

이러한 접근은 DC(Dublin Core)가 지닌 범용성과 단순성을 유지하면서도, 소프트웨어·게임과 같은 복잡한 디지털 객체의 기술적·운영적 특성을 충분히 기술할 수 있도록 하는 데 목적이 있다. 그러므로 최소한의 공통 메타데이터 집합을 유지하면서, 각 객체의 성격에 따라 추가적인 메타데이터 요소를 자유롭게 확장할 수 있는 구조를 통해 대규모 컬렉션의 이질성을 수용하고 있다.

특히 IA는 이른바 ‘매직 필드(magic fields)’와 같은 메타데이터 개념을 도입하여, 단순한 기술·설명용 메타데이터를 넘어 시스템 동작과 자동화 프로세스에 직접적인 영향을 미치는 메타데이터를 활용하고 있다. 이러한 필드는 특정 키 값이 설정될 경우 아이템의 처리 방식이 달라지거나, 자동화 스크립트 및 시스템 로직과 연동되어 에뮬레이터 실행, 접근방식, 파일 처리 흐름 등에 영향을 미치는 기능적 메타데이터로 작동한다. 이는 메타데이터를 단순한 ‘설명 정보’가 아니라 디지털 객체 관리와 제공을 제어하는 운영 메커니즘의 일부로 활용하고 있음을 보여준다.

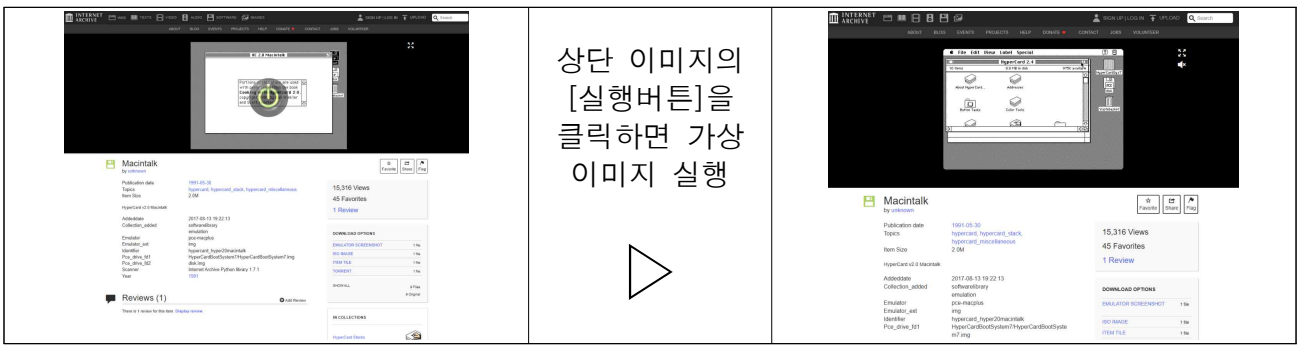
또한 IA의 메타데이터 설계는 텍스트, 이미지, 영상, 음원, 소프트웨어 등 매체 유형별 특성을 반영한 계층적 구조를 전제로 한다. 공통적으로 적용되는 기본 메타데이터 항목과 함께, 각 매체 유형에 특화된 고유 메타데이터 항목을 구분하여 설계함으로써, 이질적인 디지털 객체를 하나의 아카이브 환경에서 통합적으로 관리하면서도 매체별 기술적·내용적 차이를 충분히 표현할 수 있도록 한다. 이러한 방식은 실행형 디지털 객체와 정형 문헌 자료가 혼재하는 환경에서 메타데이터의 과도한 단순화나 불필요한 복잡화를 동시에 방지하는 역할을 수행한다.

이와 같은 IA의 메타데이터 설계 전략은 디지털 객체의 장기보존 관점에서 보존 메타데이터 표준과의 연계 필요성을 시사한다. 특히 디지털 객체의 보존대상, 보존행위, 행위 주체, 권리관계를 체계적으로 기록하기 위해서는 PREMIS(PREservation

Metadata Implementation Strategies) 표준의 핵심 요소인 object, event, agent, rights와의 매핑을 고려할 필요가 있다. DC 기반 기술·서지 메타데이터와 PREMIS(Preservation Metadata: Implementation Strategies) 기반 보존 메타데이터를 연계함으로써, 개별 디지털 객체에 대해 무엇이 보존 대상이며(object), 어떤 보존 행위가 수행되었고(event), 누가 이를 수행했으며(agent), 어떠한 권리 조건 하에서 관리·제공되는지(rights)를 구조적으로 설명할 수 있다.

이러한 매핑은 메타데이터 간의 데이터 연계성을 확보하는 동시에, 디지털 객체 보존관리 체계의 구조적 완결성을 검증하는 역할을 수행한다. 그러므로 실행형 디지털 객체에 대한 에뮬레이션 제공 여부, 접근 통제 방식, 변환·이관 이력 등이 개별 시스템의 내부 규칙에 의존하지 않고 표준화된 보존 메타데이터 구조 안에서 설명·재현 가능해진다.

종합적으로 볼 때, IA 사례는 실행형 디지털 객체를 포함한 대규모 디지털 컬렉션을 대상으로 할 경우, DC 기반의 범용 메타데이터 구조, 기능적 메타데이터(매직 필드), 매체 유형별 확장 구조, 그리고 PREMIS(Preservation Metadata: Implementation Strategies)와의 체계적 연계가 결합된 다층적 메타데이터 설계가 효과적으로 작동할 수 있음을 보여준다. 이는 에뮬레이션 기반 접근을 장기보존전략으로 고려하는 기관에 있어, 메타데이터가 단순한 기술 요소를 넘어 보존전략 실행의 핵심 인프라로 기능해야 함을 시사한다.



[그림 3] 미국의 Internet Archive(archive.org)

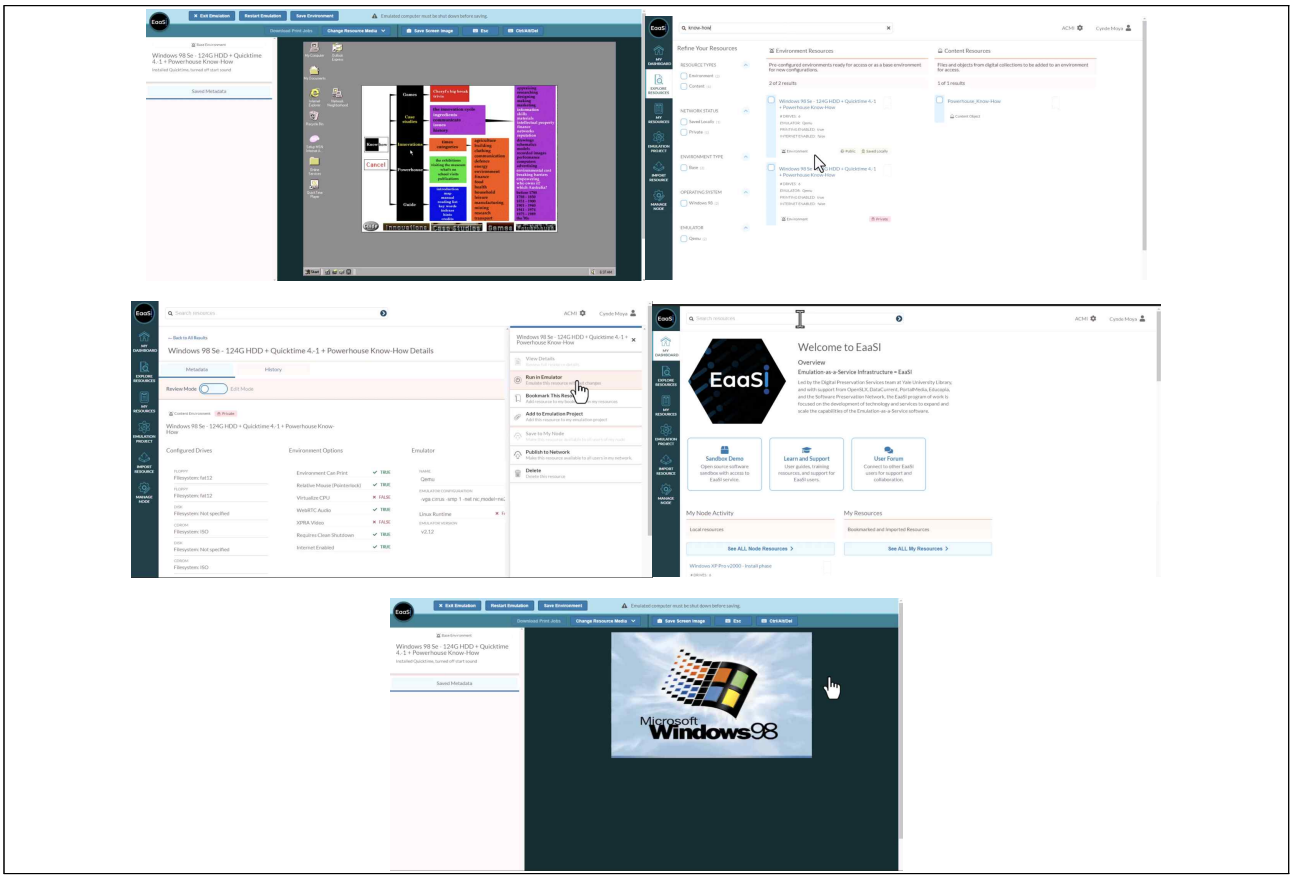
❑ 호주 AusEaaSI(auseaasi.org) : 분산 네트워크 기반 EaaS 인프라의 공공 적용

AusEaaSI(Australian Emulation as a Service Infrastructure)는 호주에서 진행 중인 「The Australian Emulation Network: Born Digital Cultural Collections Access」 프로젝트의 일환으로 구축된 에뮬레이션 기반 접근 인프라이다. 이 프로젝트는 EaaSI(Emulation as a Service Infrastructure) 개념을 토대로, 호주 내 디지털 문화유산 컬렉션에 대한 장기 접근과 활용을 가능하게 하는 것을 목표로 한다.

AusEaaSI는 OpenSLX 및 소프트웨어 보존 네트워크의 지원을 받으며, 예일대학교 도서관 디지털 보존 서비스 팀이 주도하는 연구 프로그램과 연계되어 운영되고 있다. 기술 인프라는 AARNet에서 제공하는 네트워크 및 호스팅 환경을 기반으로 하며, 이를 통해 사용자는 웹 브라우저를 통해 EaaSI 플랫폼에 접근하고, 네트워크를 통해 공유되는 사전 구성된 에뮬레이션 환경에서 디지털 미디어와 상호작용할 수 있다.

이 프로젝트의 중요한 특징은 단일 기관 중심의 서비스가 아니라 호주 내 갤러리, 도서관, 기록관, 박물관(GLAM) 조직들이 네트워크로 참여하는 확장형 협력 구조를 채택하고 있다는 점이다. AusEaaSI는 문화적으로 중요한 디지털 유물을 보유한 여러 기관이 공동의 목표 아래 사람과 기술을 연결하는 그랜드 컨소시엄 형태로 확장되고 있으며, 분산된 컬렉션을 하나의 에뮬레이션 접근 플랫폼에서 안정적으로 제공하는 것을 지향한다.

AusEaaSI는 단순히 디지털 객체를 저장하는 데 그치지 않고 원형 경험(preserving look and feel)의 유지를 핵심 가치로 설정한다. 이에 따라 미디어 아트, 건축 및 산업 디자인, 게임과 애플리케이션, AR/VR, 웹 등장 이전의 네트워킹 환경에 이르기까지 다양한 분야의 디지털자료에 대한 연구와 접근을 가능하게 한다. 예시로는 1996년 파워하우스 박물관(Powerhouse Museum)이 제작한 멀티미디어 CD-ROM 「Know-How: The Guide to Innovation in Australia」와 같은 자료가 제시된다.



[그림 4] 호주의 AusEaaSI(auseaasi.org)

❑ MAME(Multiple Arcade Machine Emulator) 프로젝트 (www.mamedev.org)

MAME(Multiple Arcade Machine Emulator)는 1997년 2월 5일 이탈리아 프로그래머 Nicola Salmoria와 초기 개발진에 의해 공개된, 아케이드 게임 하드웨어의 소프트웨어 기반 복원(emulation)을 목표로 한 오픈소스 프로젝트이다. MAME는 초기에는 아케이드 게임 하드웨어를 에뮬레이션하기 위해 설계되었으나, 이후 비디오 게임 콘솔, 오래된 개인용 컴퓨터, 기타 다양한 시스템을 포함하도록 범위를 확장하였다.

초기 MAME는 MS-DOS 전용으로 개발되었으나, 이후 Unix 계열 시스템(X/MAME), Macintosh(Mac MAME 및 이후 MAME OS X), Windows(MAME32) 등 다양한 운영체제로 이식되었다. 현재 MAME의 소스 코드는 공개 GitHub 저장소를 통해 관리되며, 필요한 전문 지식과 도구를 갖춘 개발자들이 최신 버전을 빌드하고 개선 사항을 풀 리퀘스트 형태로 기여할 수 있는 개방형 개발 구조를 유지하고 있다.

MAME의 설계 철학에서 핵심적인 점은 실행 속도나 성능 최적화보다 하드웨어의 논리적 동작을 정확히 재현하는 것을 우선한다는 점이다. 이를 위해 수천 개의

아케이드 PCB(board)에 대한 구조, ROM 맵, 입출력 포트, 그래픽 및 사운드 레이아웃을 코드 수준에서 상세히 기록하고 있으며, CPU, I/O, 버스 인터페이스를 가능한 한 정확하게 모사하도록 설계되어 있다. 이러한 특성으로 인해 MAME는 단순한 게임 실행 도구를 넘어, 과거 디지털 시스템의 동작 원리를 기록·재현하는 기술 문서적 성격을 동시에 지닌다.

2015년 5월 27일에는 게임 콘솔 및 컴퓨터 시스템 에뮬레이터인 MESS(Multi Emulator Super System)가 MAME와 통합되면서, 단일 프레임워크 안에서 보다 광범위한 시스템 에뮬레이션이 가능해졌다. MAME의 대중적 확산은 개인 매니아들이 자체 아케이드 캐비닛을 제작하는 문화로 이어졌으며, 동시에 일부 기업에서 불법적인 MAME 파생 제품을 생산하는 현상도 나타났다.

❍ MAME의 구조와 디지털 아카이브 적용

MAME는 계층형 구조를 기반으로 설계되어 있다. 최상위에는 호스트 환경(Host Layer)이 위치하며, 이는 일반적인 PC, 서버, 또는 웹 브라우저 환경을 의미한다. 이 계층은 Windows, Linux, macOS와 같은 운영체제뿐 아니라 Internet Archive에서 활용되는 WebAssembly 기반 브라우저 환경도 포함한다. 호스트 계층은 CPU 자원, 메모리, 화면 출력, 사운드 출력, 입력 장치 등 MAME 코어가 가상 하드웨어를 실행하는 데 필요한 기본 자원을 제공한다.

그 위에는 MAME 코어(Core Framework)가 위치하며, 이는 에뮬레이션 엔진의 핵심을 구성한다. 코어에는 여러 CPU와 디바이스의 타이밍을 조정하는 스케줄러, Z80·6502·68000·x86 등 다양한 CPU를 소프트웨어적으로 구현한 CPU 에뮬레이션 모듈, 메모리 맵 관리, 사운드·그래픽·입력 장치 등을 모듈화한 디바이스 프레임워크가 포함된다. 또한 비디오 및 오디오 서브시스템, 입력 시스템, 설정 및 NVRAM 관리 기능을 통해 하드웨어 동작과 사용자 설정을 재현·보존한다.

머신 드라이버(Machine Drivers)는 각 아케이드 기판, 콘솔, 레트로 컴퓨터별 하드웨어 구성을 코드로 정의하는 계층으로 CPU 종류와 클럭, 주소 공간 구성, 그래픽 및 사운드 장치 연결, 입력 매핑 등의 정보를 포함한다. 최하위에는 실제 게임이나 응용프로그램 바이너리로 구성된 ROM 또는 디스크 이미지가 위치하며, MAME는 이들 ROM 자체를 배포하지 않고 외부에서 제공된 덤프(dump)를 로드하여 실행하는 방식을 취한다.

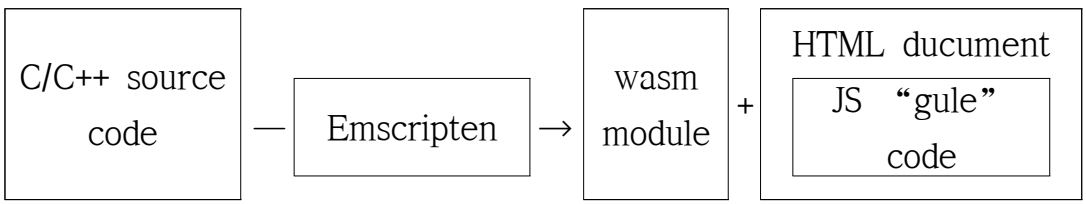
○ Internet Archive에서의 MAME 활용과 브라우저 기반 에뮬레이션

앞서 Internet Archive의 디지털 장기보존 전략과 에뮬레이션 적용 방향을 검토하였다. 특히 실행 환경 재현을 통해 디지털 객체의 기능적 진본성과 접근 가능성을 유지하는 전략적 의의를 확인하였다. 이러한 전략이 실제 시스템 환경에서 어떠한 기술 구조로 구현되고 있는지를 구체적으로 조사·분석할 필요가 있으며, Internet Archive에서 MAME를 활용한 브라우저 기반 에뮬레이션 구조를 중심으로 실행 환경 재현 모델의 기술적 특징을 검토하고자 한다.

Internet Archive는 MAME 에뮬레이터와 Emscripten 기술을 결합하여, Internet Arcade를 포함한 브라우저 기반 에뮬레이션 서비스를 제공하고 있다. Emscripten은 C/C++ 등으로 작성된 소스 코드를 웹 브라우저 또는 WebAssembly(WASM) 런타임에서 실행 가능하도록 변환하는 컴파일러로, 고성능이 요구되는 게임이나 에뮬레이션 코드를 브라우저 환경에서 실행할 수 있게 한다.

IA의 보안 정책상 로컬 파일 접근은 차단되어 있으므로, 사용자가 개인 PC에 있는 파일을 직접 업로드하거나 실행하는 기능은 제공되지 않는다. 대신, 특정 응용프로그램을 실행할 수 있는 구형 운영체제 이미지와 해당 프로그램, 그리고 열람 대상 파일을 하나의 이미지 파일로 통합하여 등록하는 방식을 사용한다. 예를 들어, MS-DOS 6.22 운영체제, Word 2.0 응용프로그램, 그리고 example.doc 파일을 하나의 디스크 이미지(*.img)로 구성하면, MAME 에뮬레이션 환경 내에서 해당 파일을 실행하고 열람할 수 있다.

이와 같은 방식은 실행형 디지털 객체의 보존과 접근을 파일 단위가 아닌 ‘실행 가능한 환경 단위’로 관리한다는 점에서 디지털 아카이브에서 에뮬레이션 전략을 적용하는 하나의 구체적 운영 모델을 제시한다.

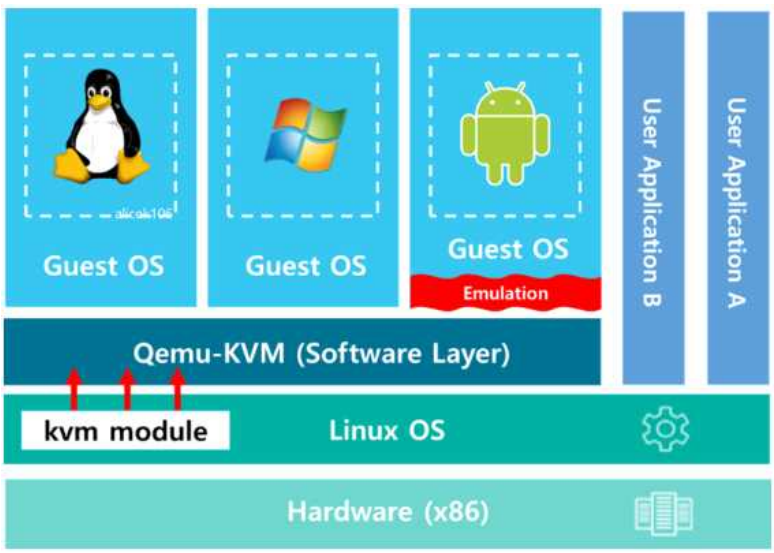


[그림 5] MAME 웹어셈블러

❑ KVM+OpenStack / KVM+QEMU(클라우드/가상화 기반 에뮬레이션 서비스)

디지털 장기보존 분야에서 에뮬레이션은 더 이상 개별 연구자의 실험적 기술이 아니라 공공기관 차원의 보존·접근 전략으로 점차 제도화되고 있다. 이러한 흐름의 기술적 기반에는 QEMU(Quick EMUlator)와 KVM(Kernel-based Virtual Machine)이 핵심 구성요소로 자리하고 있다. QEMU는 서로 다른 CPU 아키텍처와 운영체제를 소프트웨어적으로 재현할 수 있는 범용 에뮬레이터로, 구형 운영체제와 응용프로그램을 실행하는 데 필수적인 도구로 활용되어 왔다. KVM은 리눅스 커널에 통합된 하이퍼바이저로서, QEMU와 결합될 경우 가상머신 수준의 안정성과 성능을 확보할 수 있다. 이러한 조합은 디지털 객체의 실행환경을 재현 가능한 단위로 관리하는 데 적합한 구조를 제공한다.

이러한 에뮬레이션·가상화 기술은 OpenStack과 같은 IaaS(Infrastructure as a Service, 서비스형 인프라) 플랫폼과 결합될 경우, 단일 시스템 수준을 넘어 EaaS(Emulation as a Service) 형태로 확장될 수 있다. OpenStack은 가상머신의 생성·배포·자원 할당·네트워크·스토리지 관리를 통합적으로 제어할 수 있는 오픈소스 클라우드 플랫폼으로 기본 하이퍼바이저로 KVM을 지원한다. 디지털 보존 분야의 연구와 실무 논의에서는 이러한 구조가 공공기관 환경에서 에뮬레이션을 개별 PC에서의 실행이 아닌 중앙 관리형 서비스로 제공할 수 있는 현실적인 기술 토대임을 지적하고 있다. 그러므로 에뮬레이션은 더 이상 보존만을 위한 내부 기술이 아니라 이용자 접근을 전제로 한 서비스 인프라로 전환되고 있다.



[그림 6] KVM-QEMU 모듈

❑ 영국 TNA 디지털 기록 재현(Replay) 프로젝트

영국 국가기록원(The National Archives, UK, TNA)은 에뮬레이션 및 재현 개념을 기관의 공식 디지털 전략에 명시적으로 포함시킨 대표적 사례이다. TNA는 『Digital Strategy 2017-2019』에서 디지털 기록을 현재의 소프트웨어로 더 이상 읽거나 사용할 수 없는 상태가 될 위험이 있는 대상으로 규정하고, 이러한 위험에 대응하기 위한 우선적 접근방식으로 에뮬레이션(Emulation)을 채택한다고 명시하였다. 특히 TNA는 마이그레이션(Migration)을 보조적 수단으로 위치시키고, 에뮬레이션을 통해 기록이 생성·활용되던 방식 그대로 재현(replay)하는 것이 장기적 접근성 확보에 더 적합하다고 서술한다.

이러한 전략은 TNA의 웹 아카이브 및 디지털 기록 제공 정책과도 직접적으로 연결된다. TNA는 웹 기록과 디지털 기록이 단순한 정적 파일이 아니라 코드·스크립트·데이터·시각적 렌더링이 결합된 복합 객체임을 전제로, ‘capture and replay’라는 개념을 통해 수집 이후의 재현 품질을 보존전략의 핵심 요소로 다루고 있다. 이는 전통적인 의미의 파일 보존을 넘어, 기록의 행위성(behavior)과 경험적 특성을 보존대상으로 포함시키는 접근으로 평가된다. TNA 사례는 공공기록관리 기관이 에뮬레이션(Emulation)과 재현을 단순한 기술 실험이 아니라 제도적으로 승인된 장기보존전략으로 채택할 수 있음을 보여준다.

❑ 독일 DNB EaaS 프로젝트 (KVM/QEM 기반 웹 스트리밍)

독일 국립도서관(Deutsche Nationalbibliothek, DNB)은 유럽 차원의 협력 프로젝트를 통해 에뮬레이션(Emulation) 기반 접근 제공을 실질적으로 구현한 대표적 공공기관이다. DNB가 참여·주도한 EMiL(Emulation of Multimedia Objects in Libraries) 프로젝트는, CD-ROM과 같은 멀티미디어 객체 및 구형 소프트웨어 의존 자료를 장기적으로 이용 가능하게 만드는 것을 목표로 수행되었다. EMiL은 보존 대상 객체의 기술적 특성을 자동으로 분석하고, 해당 객체에 적합한 운영체제 및 응용프로그램 환경을 에뮬레이션을 통해 자동 매칭·실행하는 프레임워크를 제안하였다.

EMiL의 중요한 특징은 에뮬레이션(Emulation)을 보존 이후의 부가적 서비스가 아니라 도서관 접근 서비스의 핵심 구성요소로 설계했다는 점이다. 프로젝트 결과 문헌에서는 사서나 관리자가 개별 객체마다 수작업으로 환경을 구성하는 방식이 아니라 대규모 컬렉션에 적용 가능한 자동화된 에뮬레이션 접근 제공이 필수적임을 강조한다. 이는 에뮬레이션이 연구실 수준을 넘어, 실제 공공 도서관 서비스 환경에서도 운영 가능함을 실증한 사례로 평가된다.

또한 DNB 및 관련 유럽 프로젝트들은 에뮬레이션 환경을 중앙 서버에서 실행하고, 이용자는 웹 인터페이스를 통해 해당 환경에 접근하는 방식이 저작권 관리와 운영 통제 측면에서 효과적임을 지적한다. 이러한 구조는 결과적으로 EaaS(Emulation as a Service) 모델로 수렴되며, 이 과정에서 QEMU 기반 에뮬레이션과 가상화 기술이 핵심 실행 계층으로 활용된다.

3.1.3. 국립중앙도서관 디지털자원 보존전략(안)

앞서 디지털 장기보존전략의 주요 유형과 그 기술적 특성을 비교·분석함으로써, 각 전략이 지니는 적용 범위와 한계를 검토하였다. 그러나 장기보존전략은 기술적 선택의 문제가 아니라, 기관의 역할과 법적 책무, 자원 구성 특성, 서비스 정책을 종합적으로 반영하여 결정되어야 할 정책적 사안이다. 특히 국립중앙도서관과 같이 국가 지식자원 보존을 수행하는 기관의 경우, 보존전략은 단순한 기술 도입을 넘어 제도화·표준화·거버넌스 체계와 연계되어야 한다는 점에서 정책적 판단의 중요성이 더욱 크다.

❑ 보존전략 선택 기준

국립중앙도서관의 디지털자원은 납본·수집 제도에 의해 출판물 중심의 정형 자료부터 웹자원, 멀티미디어, 데이터세트에 이르는 매우 다양한 유형으로 구성되어 있다. 따라서 단일 보존전략을 모든 자료에 일괄 적용하는 방식은 현실적이지 않으며, 자료유형별 특성과 보존 목적에 따라 복수의 전략을 조합하여 적용하는 선택 기준이 필요하다.

첫째, 자료의 논리적 구조와 표현 복잡성은 핵심 기준이다. 텍스트 중심의 정형 자료는 포맷 변환이나 표준화에 따른 정보 손실 위험이 낮은 반면, 복잡한 레이아웃이나 인터랙션을 포함한 자료는 마이그레이션 시 진본성 훼손 가능성이 높다. 이 경우 실행환경 보존이나 패키징 전략의 비중이 높아질 수밖에 없다.

둘째, 접근성과 이용 지속성의 요구 수준이다. 국립중앙도서관의 주요 디지털자원은 장기적으로 읽을 수 있는 것이 핵심 가치이므로, 다수 이용자가 범용 환경에서 접근할 수 있는 포맷 유지가 중요하다. 이에 따라 텍스트·이미지·일반 영상 자료는 마이그레이션(Migration) 및 포맷정규화(Normalization) 전략이 우선 적용 대상이 된다.

셋째, 비트 수준 보존의 안정성은 모든 자료유형에 공통적으로 요구되는 전제 조건이다.

저장매체의 노후화와 기술적 구식화는 자료유형과 무관하게 발생하므로, 마이그레이션(Migration)은 국립중앙도서관의 모든 디지털자원에 기본 전략으로 적용되어야 한다.

넷째, 보존 책임의 증거성·신뢰성 요구이다. 국립중앙도서관은 국가대표 도서관으로서 보존 행위의 정당성과 추적 가능성을 확보해야 하므로, 대부분의 디지털자원은 캡슐화(Encapsulation)(OAIS AIP, METS/PREMIS)을 통해 관리되는 것이 바람직하다.

다섯째, 저작권·법적 제약과 기술적 실현 가능성이다. 에뮬레이션은 진본성 측면에서 가장 강력한 전략이지만, 저작권 및 기술적 부담이 크기 때문에 모든 자료에 일반화하기보다는, 대체 수단이 없는 특정 유형에 제한적으로 적용하는 것이 합리적이다.

❑ 자료유형별 보존전략 현황 및 적용 방향

○ 전자책·전자저널·학위논문(PDF, ePub 등)

이들 자료는 국립중앙도서관 디지털 장서의 가장 핵심적이고 대량적인 유형으로 구조가 비교적 안정적이고 텍스트 중심이라는 특성을 지닌다. 따라서 장기보존 전략으로는 포맷정규화(Normalization)와 마이그레이션(Migration)을 중심 전략으로 삼는 것이 적절하다. PDF/A와 같은 장기보존 적합 포맷을 우선 적용하고, 기술환경 변화에 따라 주기적 마이그레이션(Migration)을 수행함으로써 접근성을 유지할 수 있다. 동시에 OAIS(Open Archival Information System) 기반 캡슐화(Encapsulation)을 통해 출처, 권리, 변환 이력을 패키지 단위로 관리하고, 매체전환(Refreshment)를 통해 비트 수준 보존을 지속하는 것이 기본 전략의 조합이 된다.

○ 디지털화된 인쇄자료 이미지(TIFF, JPEG 등)

스캔 이미지 형태의 디지털화 자료는 원본 인쇄물의 대체물로서 내용적 정확성과 이미지 품질 유지가 중요하다. 이 경우 무손실 또는 최소 손실 포맷으로의 포맷 정규화(Normalization)이 우선되며, 장기 접근성을 위해 마이그레이션(Migration)이 병행된다. 캡슐화(Encapsulation)을 통해 디지털화 과정, 해상도, 생성 이력 등을 함께 관리함으로써 보존의 신뢰성을 강화할 수 있다. 다만 에뮬레이션(Emulation)은 상대적으로 적용하기에 적합성이 낮은 편이다.

○ 웹자원(OASIS 수집 웹사이트, WARC(Web ARChive) 등)

웹자원은 구조적·동적 요소를 포함하는 복합 객체로, 단순 포맷 변환만으로는

원형 유지가 어렵다. 따라서 캡슐화(Encapsulation)는 WARC(Web ARChive) 기반 패키징이 핵심 전략이 되며, 접근성을 유지하기 위해 제한적인 마이그레이션(Migration)이 병행된다. 매체전환(Refreshment)은 저장 안정성 확보 차원에서 필수적으로 적용된다. 특정 웹 기술이나 스크립트에 의존하는 경우에는, 장기적으로 에뮬레이션(Emulation)은 전략을 활용하여 보완적으로 고려될 수 있다.

○ 시청각 자료(영상・음원)

영상과 음원 자료는 용량이 크고 기술 변화 속도가 빠르기 때문에, 포맷정규화(Normalization)와 마이그레이션(Migration)의 병행 적용이 핵심 전략이 된다. FFV1, WAV 등과 같은 장기보존 친화적 포맷을 기준으로 관리하되, 접근 환경 변화에 따라 주기적 마이그레이션이 요구된다. 캡슐화(Encapsulation)을 통해 기술 메타데이터와 변환 이력을 관리하고, 매체전환(Refreshment)를 통해 저장매체의 안정성을 유지한다. 에뮬레이션(Emulation)은 일반적인 영상・음원에는 적용 적합성이 낮다.

○ 데이터세트・연구데이터(CSV, XML, DB 덤프 등)

데이터세트는 구조적 정보와 의미적 맥락이 중요하므로, 단순 파일 보존을 넘어 구조와 메타데이터를 함께 보존하는 캡슐화(Encapsulation) 전략이 중요하다. SIARD와 같은 표준 포맷으로의 포맷정규화(Normalization)과 마이그레이션(Migration)이 병행되며, 매체전환(Refreshment)은 기본 전제 조건으로 적용된다. 실행환경 재현보다는 구조적 재현이 핵심이므로 에뮬레이션(Emulation)의 필요성은 제한적으로 적용될 수 있다.

○ 특수 디지털 객체(인터랙티브 콘텐츠, 구형 소프트웨어 등)

이 유형은 국립중앙도서관 내에서 상대적으로 비중은 낮지만, 장기보존 난이도가 가장 높은 자료군이다. 마이그레이션(Migration)이나 포맷정규화(Normalization)만으로는 원형 보존이 어려운 경우가 많아, 에뮬레이션(Emulation) 전략이 핵심적인 선택지가 된다. 다만 에뮬레이션 환경을 설명・재현하기 위한 메타데이터 관리가 필요하므로 캡슐화(Encapsulation)와의 결합이 필수적이며, 매체전환(Refreshment)은 저장 안정성 확보를 위한 기본 전략으로 적용된다.

자료유형	주요 특성	핵심 보존전략	보조 전략	비고
전자책 · 전자저널 · 학위논문	정형 · 텍스트 중심	Normalization, Migration	Encapsulation, Refreshment	PDF/A 중심
디지털화 인쇄자료 이미지	고해상도 이미지	Normalization	Migration, Encapsulation, Refreshment	무손실 이미지
웹자원(WARC)	동적 · 복합 구조	Encapsulation	Migration, Refreshment	OASIS 수집
시청각 자료	대용량 · 기술 변화 빠름	Migration, Normalization	Encapsulation, Refreshment	FFV1, WAV 등
데이터세트 · DB	구조 · 맥락 중요	Encapsulation	Normalization, Migration, Refreshment	SIARD 등
특수 디지털 객체	환경 의존성 높음	Emulation	Encapsulation, Refreshment	제한적 적용

[표 15] 자료유형별 보존전략 선택 기준 및 적용

❑ 보존전략 적용 시 고려사항

국립중앙도서관에서 디지털 장기보존전략을 적용할 때 가장 우선적으로 유의해야 할 점은 보존전략이 개별 기술의 선택 문제가 아니라 기관의 책임 범위와 운영 논리를 구체화하는 과정이라는 점이다. 국립중앙도서관은 국가대표 도서관으로서 방대한 규모의 디지털자원을 지속적으로 수집·관리해야 하며, 모든 자료에 최고 수준의 보존전략을 동일하게 적용하는 것은 현실적으로 불가능하다. 따라서 전략 적용은 반드시 자료유형·보존 목적·법적 책임 수준에 따른 차등적 접근을 전제로 해야 한다.

첫째, 보존전략의 선택이 곧 '책임 선언'으로 해석될 수 있음을 인식할 필요가 있다. 예컨대 에뮬레이션(Emulation) 전략을 적용한다는 것은 단순히 기술을 도입하는 차원이 아니라 해당 자료에 대해 원본 실행 경험까지 장기적으로 재현할 책임을 진다는 의미를 내포한다. 이는 향후 기술 유지, 인력 전문성, 저작권 문제까지 포함하는 장기적 부담으로 이어질 수 있으므로, 국립중앙도서관은 모든 자료에 대해 최대 수준의 전략을 적용하기보다는, 전략별 적용 범위를 명확히 설정하고 이를 내부 정책으로 문서화할 필요가 있다.

둘째, 매체전환(Refreshment)·마이그레이션(Migration)·포맷정규화(Normalization)과 같은 기본 전략과 캡슐화(Encapsulation)· 에뮬레이션(Emulation)과 같은 고급 전략을 명확히 구분하여 운영해야 한다. 매체전환(Refreshment)는 모든 디지털자원에 공통적으로 적용되어야 할 최소한의 유지관리 전략인 반면, 에뮬레이션(Emulation)은 특정

자료유형에 한정하여 선택적으로 적용되어야 하는 고비용·고위험 전략이다. 이 구분이 명확하지 않을 경우, 보존전략이 기술 중심으로 과잉 확장되거나, 반대로 필요한 영역에서 적용되지 못하는 문제가 발생할 수 있다.

셋째, 보존전략은 서비스 중심 시스템(KOLIS)과 분리된 논리로 설계되어야 한다. 국립중앙도서관의 기존 시스템은 대국민 서비스 제공을 중심으로 구축되어 있어, 접근성과 편의성은 높지만 장기보존을 위한 AIP(Archival Information Package) 중심 관리, 보존 이력 추적, 무결성 자동 검증 등에는 구조적 한계가 있다. 보존전략을 적용할 때에는 서비스 요구에 의해 보존 정책이 왜곡되지 않도록, 보존 영역을 독립적인 책임 단위로 설정하고 서비스는 보존 객체를 참조하는 구조를 유지해야 한다.

넷째, 보존전략 적용 시 자동화 가능성과 지속 운영 가능성을 함께 고려해야 한다. 마이그레이션(Migration)이나 Fixity 검증은 자동화와 결합될 때 대규모 장서 환경에서 실효성을 갖는다. 반면, 수작업에 의존하는 보존 절차는 단기적으로는 가능하더라도 장기적으로는 누적 비용과 오류 위험을 급격히 증가시킨다. 따라서 국립중앙도서관은 전략 선택 단계에서부터 해당 전략이 향후 자동화·시스템화될 수 있는지 여부를 중요한 판단 기준으로 삼아야 한다.

다섯째, 보존전략 적용은 메타데이터 체계와 분리하여 논의할 수 없다. 마이그레이션(Migration), 포맷정규화(Normalization), 에뮬레이션(Emulation)등 모든 보존 행위는 그 자체로 보존 이벤트이며, 이는 PREMIS(Preservation Metadata: Implementation Strategies)와 같은 보존 메타데이터로 기록되지 않을 경우 장기적으로 증명 가능성을 상실하게 된다. 국립중앙도서관은 보존전략을 적용할수록 보존 메타데이터의 중요성이 기하급수적으로 증가한다는 점을 인식하고 전략 적용과 메타데이터 생산·관리 체계를 항상 연동하여 설계해야 한다.

여섯째, 저작권·라이선스·법적 제약은 전략 적용의 기술적 가능성과 별개로 검토되어야 한다. 특히 SW·게임·웹 콘텐츠와 같이 실행환경을 재현하거나 복제·변환이 수반되는 자료유형의 경우, 기술적으로는 가능하더라도 법적으로 허용되지 않는 경우가 존재할 수 있다. 국립중앙도서관은 보존전략을 기술 중심으로만 판단하지 말고, 전략 적용 가능 범위를 법적·윤리적 관점에서 사전에 명확히 설정해야 한다.

마지막으로 장기보존전략은 ‘완결된 상태’를 목표로 해서는 안 되며, 점진적·누적적 체계로 인식되어야 한다. 디지털 환경은 지속적으로 변화하므로, 특정 시점에 완성된 보존전략은 존재하지 않는다. 국립중앙도서관은 전략을 일회성 사업이나 단기 과제로 인식하기보다는, 장기적으로 축적·보완·조정되는 운영 프레임워크로 관리해야 하며, 이를 전제로 한 조직·예산·평가 체계를 함께 설계할 필요가 있다.

❑ 사례를 통한 종합 시사점

이상의 사례를 종합하면, 에뮬레이션과 재현 기반 보존전략은 단순히 특수한 디지털 객체를 위한 보조적 방법이 아니라 소프트웨어・게임・인터랙티브 콘텐츠・웹 기록 등 실행환경 의존 자료 전반을 포괄하는 핵심 전략으로 자리매김하고 있음을 확인할 수 있다. 영국 TNA는 에뮬레이션을 공식 디지털 전략에 포함시켜 정책적 정당성을 부여하였고, 독일 DNB는 EMiL 프로젝트를 통해 이를 실제 도서관 서비스 환경에 적용하였다. 이들 사례는 공공기관이 에뮬레이션 기반 접근을 기술・정책・운영 차원에서 통합적으로 설계할 수 있음을 보여준다.

특히 KVM・QEMU와 같은 오픈소스 가상화・에뮬레이션 기술은 OpenStack과 결합될 경우 공공기관 내부에서 통제 가능하고 확장 가능한 EaaS 인프라를 구축할 수 있는 현실적인 선택지로 평가된다. 이는 에뮬레이션을 개별 연구자나 단일 PC 환경에 국한하지 않고 국가 차원의 디지털 보존 서비스로 확장할 수 있는 기술적 근거를 제공한다.

이러한 국제적 사례는 국립중앙도서관을 포함한 국가대표 도서관이 향후 디지털 보존전략을 수립함에 있어, 마이그레이션・정규화 중심의 기존 접근을 보완하는 방식으로 에뮬레이션 기반 보존(EaaS, replay)을 중장기 전략에 포함시킬 필요성을 시사한다. 이는 단지 기술 도입의 문제가 아니라 디지털 문화유산의 진본성, 이용 가능성, 그리고 미래 세대의 연구・활용 가능성을 동시에 보장하기 위한 필수적인 전략적 선택으로 이해될 수 있다.

3.2. 스토리지 및 백업 인프라 구성 방안

3.2.1. 파일저장 스토리지 구성 방안

3.2.1.1. 매체 유형에 따른 스토리지 분류

❑ HDD (Hard Disk Drive) 자기식 저장장치

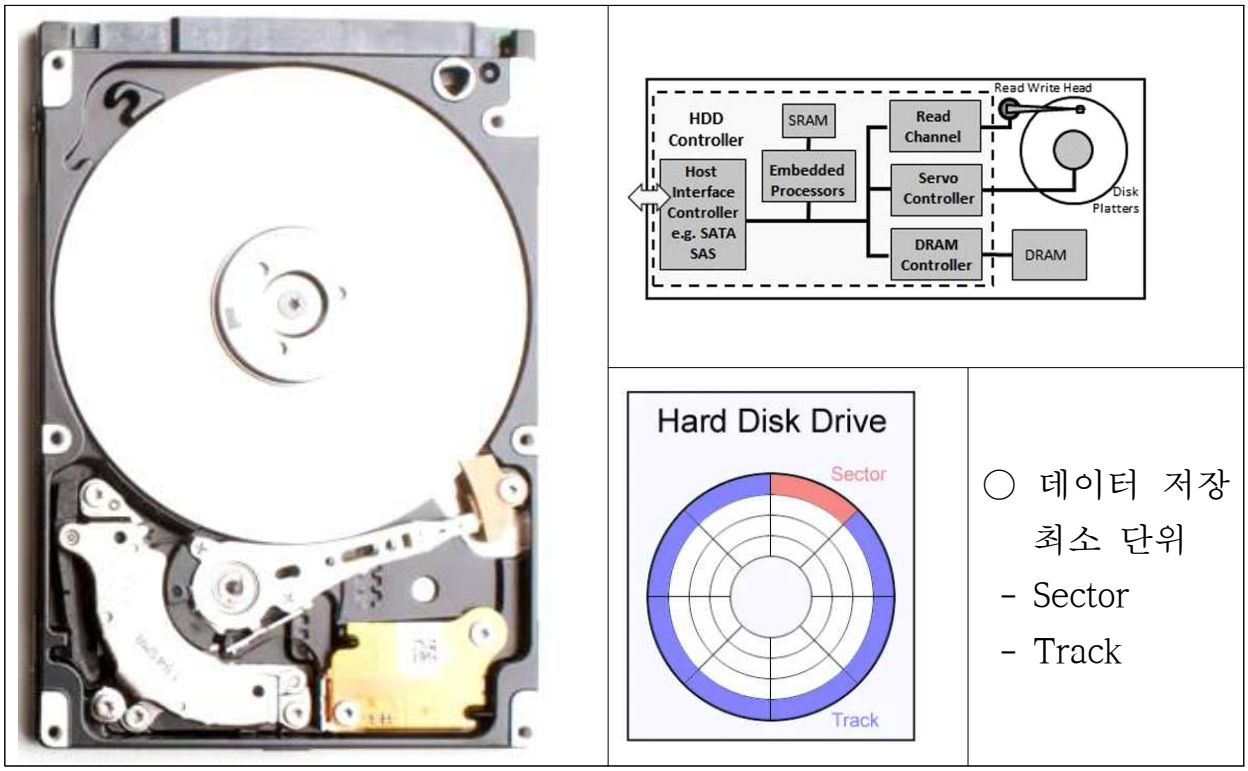
하드디스크 드라이브(Hard Disk Drive, HDD)는 기계적으로 회전하는 플래터와 자기 기록 헤드를 이용하여 데이터를 저장하고 읽어들이는 저장매체이다. 데이터는 플래터 표면에 형성된 자성 패턴의 형태로 기록되며, 기록된 데이터에 접근하기 위해서는 자기 헤드가 물리적으로 해당 위치로 이동해야 한다. 이러한 구조적 특성으로 인해 하드디스크 드라이브(Hard Disk Drive, HDD)는 고정된 메모리 셀에 직접 접근하는 방식의 저장장치에 비해 접근 시간이 상대적으로 길다.

하드디스크 드라이브(Hard Disk Drive, HDD)는 여러 기계적·전자적 구성요소로 이루어져 있다. 플래터는 데이터를 저장하는 원형의 자성 디스크로, 분당 약 5,400회에서 15,000회까지 회전하며, 기록 밀도에 따라 전체 저장 용량이 결정된다. 스핀들 모터는 플래터를 고속으로 회전시키는 역할을 수행하며, 하드디스크 드라이브(Hard Disk Drive, HDD)의 전력 소비와 발열에 큰 영향을 미치는 요소이다. 읽기/쓰기 헤드는 플래터 표면의 자성을 변화시키거나 이를 감지하여 데이터를 처리하는 정밀 장치이며, 액추에이터 암은 해당 헤드를 원하는 트랙 위치로 이동시키는 기계적 구조를 담당한다. 이와 함께 컨트롤러 PCB는 명령 처리, 버퍼 관리, 오류 정정 코드(ECC) 수행, 헤드 제어, 그리고 SATA 또는 SAS 통신을 담당하며, 캐시 메모리는 읽기 및 쓰기 성능을 보완하기 위한 버퍼 역할을 수행한다.

하드디스크 드라이브(Hard Disk Drive, HDD)는 사용 환경과 목적에 따라 서로 다른 인터페이스를 통해 시스템과 연결된다. SATA(Serial ATA)는 일반 개인용 컴퓨터에서 널리 사용되는 표준 인터페이스로, 비교적 저렴한 비용과 플러그 앤 플레이 방식을 지원하며, SATA 3 기준으로 최대 6Gb/s의 전송 속도를 제공한다. SAS(Serial Attached SCSI)는 서버 및 엔터프라이즈 환경을 위한 인터페이스로, SATA보다 높은 최대 12Gb/s 이상의 속도를 제공하며, 높은 신뢰성과 확장성, 양방향 통신 기능을 지원한다. 두 인터페이스 모두 하드디스크 드라이브(Hard Disk Drive, HDD)와 솔리드 스테이트 드라이브(Solid State Drive, SSD)에서 공통적으로 사용될 수 있다.

하드디스크 드라이브(Hard Disk Drive, HDD)의 주요 장점은 대용량 저장을 매우

낮은 비용으로 제공할 수 있다는 점이다. 저장 용량 대비 비용이 최소화되어 대규모 데이터 저장에 유리하며, 제조 기술이 성숙 단계에 도달하여 대량 생산과 유지보수가 비교적 용이하다. 또한 순차적인 읽기 및 쓰기 작업에서는 여전히 우수한 성능을 제공한다. 반면, 기계적 회전과 이동 구조로 인해 충격과 진동에 취약하며, 탐색 시간과 회전 지연으로 인해 랜덤 접근 속도가 다른 저장매체에 비해 현저히 느리다. 더불어 지속적인 전력 소비로 인한 발열과 소음이 발생하고 기계적 마모에 따라 저장장치의 수명이 영향을 받는 한계를 가진다.



[그림 7] HDD 스토리지 구조

❑ SSD (Solid State Drive) 전자식 저장장치

솔리드 스테이트 드라이브(Solid State Drive, SSD)는 기계적인 구동부를 전혀 포함하지 않고 반도체 기반의 플래시 메모리를 전기 신호로 제어하여 데이터를 저장하는 저장매체이다. 회전 디스크나 기계적 헤드 없이 낸드 플래시 메모리와 컨트롤러로 구성되기 때문에, 물리적 이동에 따른 지연이 발생하지 않으며, 이로 인해 매우 빠른 처리 속도와 높은 내구성을 제공한다.

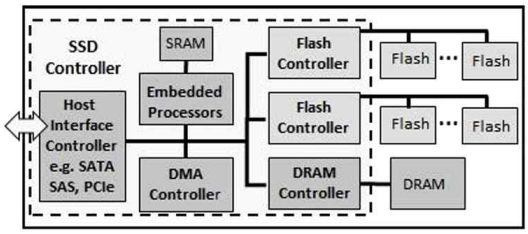
솔리드 스테이트 드라이브(Solid State Drive, SSD)의 핵심 구성요소는 컨트롤러와

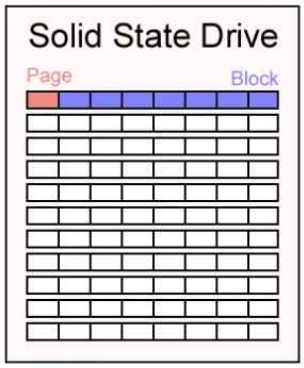
낸드 플래시 메모리이다. 컨트롤러는 낸드 플래시 메모리에 대한 데이터 저장과 관리, 읽기・쓰기 동작을 제어하는 중심 부품으로서 솔리드 스테이트 드라이브(Solid State Drive, SSD)의 안정성과 신뢰성, 전반적인 성능에 직접적인 영향을 미친다. 낸드 플래시는 전원이 차단되더라도 데이터가 유지되는 비휘발성 메모리로, SLC, MLC, TLC, QLC 등 다양한 셀 구조를 통해 구현된다. 낸드 플래시 메모리는 데이터 처리 단위에 따라 페이지와 블록으로 구성되는데, 페이지는 실제 데이터를 저장하는 최소 기본 단위이며, 블록은 다수의 페이지를 묶은 상위 단위로서 데이터 삭제와 관리의 기본 단위가 된다. 이와 함께 DRAM 또는 캐시 버퍼는 메타데이터 매핑 테이블 관리와 쓰기 캐싱 등에 활용되어 입출력 성능을 향상시키는 역할을 수행한다.

솔리드 스테이트 드라이브(Solid State Drive, SSD)는 다양한 인터페이스와 폼팩터를 통해 시스템과 연결된다. SATA 인터페이스는 하드디스크 드라이브(Hard Disk Drive, HDD) 시대부터 사용된 범용 규격으로 최대 6Gbps 속도를 제공하며 보급형 솔리드 스테이트 드라이브(Solid State Drive, SSD)에 적용된다. SAS 인터페이스는 서버 및 스토리지 환경을 위한 고신뢰성 규격으로 듀얼 포트와 12~24Gbps의 대역폭을 지원한다. PCIe 인터페이스는 CPU와 직접 연결되는 고속 버스로, 솔리드 스테이트 드라이브(Solid State Drive, SSD)가 최고 수준의 성능을 발휘할 수 있는 물리적 기반을 제공한다. NVMe는 PCIe 기반 솔리드 스테이트 드라이브(Solid State Drive, SSD) 전용 프로토콜로서 초저지연과 높은 IOPS 특성을 제공한다. 이 외에도 SATA 또는 NVMe 방식을 모두 지원하는 M.2 카드형 폼팩터, 서버 환경에서 핫스왑을 지원하는 U.2 인터페이스, SATA・SAS・NVMe를 단일 백플레인에서 지원하는 U.3 인터페이스, 그리고 데이터센터 및 AI 서버를 위한 고성능・고밀도 폼팩터인 EDSFF(E1/E3) 규격이 사용된다.

솔리드 스테이트 드라이브(Solid State Drive, SSD)의 장점은 기계적 부품이 없다는 구조적 특성에서 비롯된다. 이로 인해 랜덤 입출력 성능이 매우 우수하며, 전력 소비가 낮고 발열, 소음, 진동이 거의 발생하지 않는다. 또한 외부 충격에 강하고 데이터 위치와 무관하게 접근 지연이 거의 동일한 임의 접근 특성을 제공한다. 반면 솔리드 스테이트 드라이브(Solid State Drive, SSD)는 하드디스크 드라이브(Hard Disk Drive, HDD)에 비해 저장 용량 대비 비용이 높아 대용량 저장 시 비용 부담이 발생하며, 낸드 플래시의 물리적 특성으로 인해 쓰기 수명에 한계가 존재한다. 더불어 일부 컨트롤러와 플래시 메모리 조합에서는 성능 저하나 데이터 유지와 관련된 문제가 발생할 가능성도 지적된다.







○ 데이터 저장 최소 단위
- Page
- Block

[그림 8] SSD 스토리지 구조

※ 블록은 HDD의 트랙에 대응되고, 페이지는 HDD의 섹터에 대응되므로, 데이터 저장의 최소 단위는 HDD에서는 섹터이고, SSD에서는 페이지임

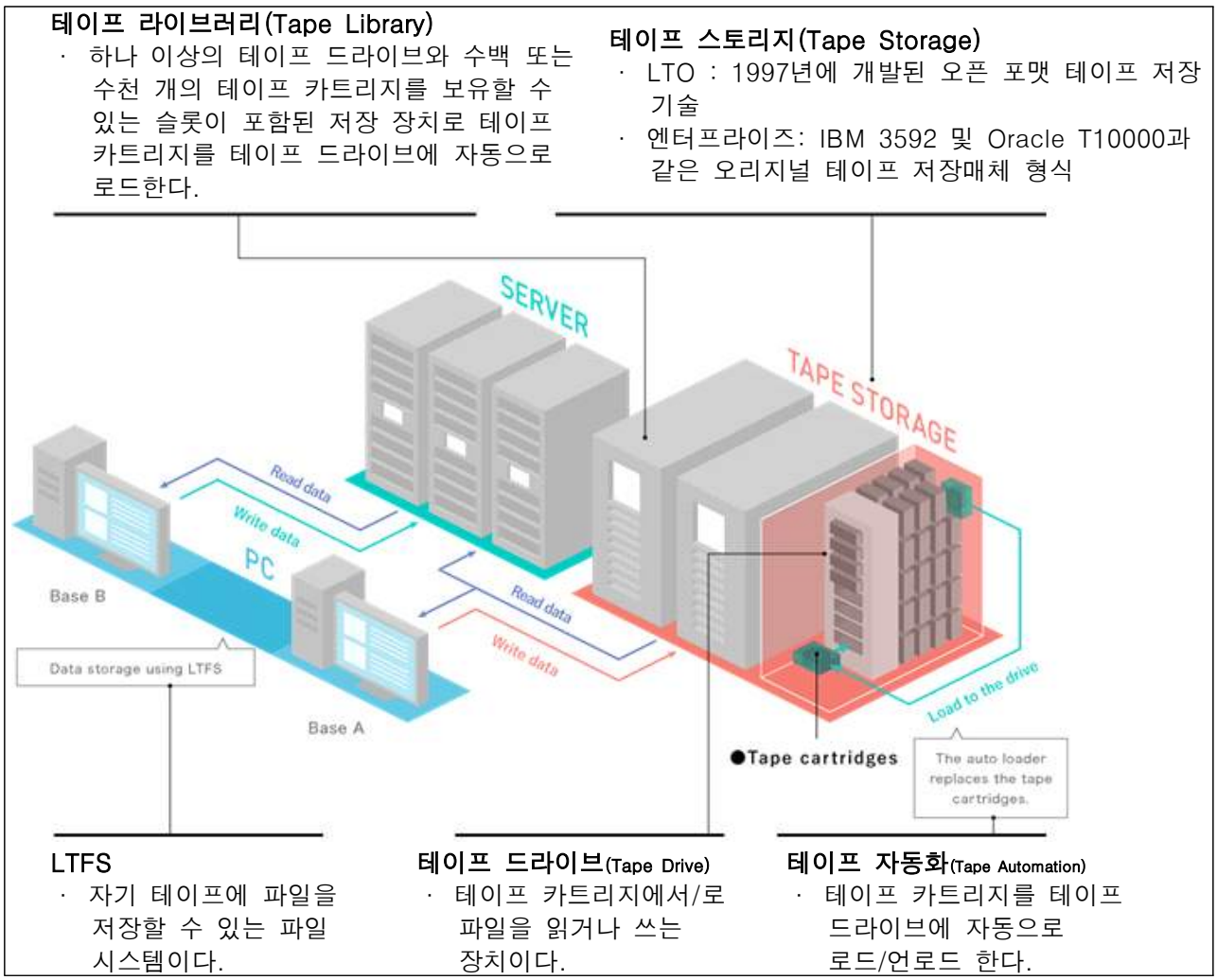
❑ 테이프 저장장치

LTO(Linear Tape-Open) 테이프는 길게 이어진 자기 테이프에 데이터를 순차적으로 기록하는 저장매체로서, 대용량 데이터의 장기보존과 백업을 목적으로 설계된 기술이다. 테이프 카트리지는 보관 중 전력을 전혀 사용하지 않는 구조를 가지며, 이로 인해 저전력·저비용·장기 안정성 측면에서 높은 효율성을 제공하는 저장 기술로 평가된다.

LTO(Linear Tape-Open) 테이프 시스템은 여러 구성요소로 이루어진다. 테이프 카트리지는 실제 데이터를 저장하는 자기 테이프가 감겨 있는 물리적 매체로, 미디어 자체는 데이터 보관 상태에서 전력을 소비하지 않는다. 테이프 드라이브는 테이프를 이동시키고 데이터를 기록하거나 읽는 장치로서, 내부에 헤드, 모터, 서보 시스템을 포함한다. 서보 트랙은 테이프 헤드가 정확한 위치를 추적하고 정렬할 수 있도록 안내하는 정렬 정보이며, 테이프 헤드는 자기 신호를 기록하거나 판독하는 핵심 부품이다. 또한 LTFS(Linear Tape File System)는 테이프를 파일시스템처럼 인식하도록 하여, 사용자가 일반 파일과 유사한 방식으로 데이터에 접근할 수 있도록 지원하는 기술이다.

LTO(Linear Tape-Open) 테이프의 장점은 대용량 장기보존 환경에 적합한 특성에서 나타난다. 저장 용량 대비 비용이 가장 저렴하여 페타바이트(PB)에서 엑사바이트(EB) 규모의 데이터 보존에 유리하며, 미디어 자체가 전력을 거의 사용하지 않아 에너지 효율이 매우 높다. 또한 30년 이상 보존 가능한 장기 안정성을 제공하고 외부 충격에 강하며 오프라인 저장이 가능하다는 점에서 재해나 랜섬웨어 대응 측면에서도 효과적인 저장매체로 평가된다. 아울러 대용량 데이터를 순차적으로 처리하는 입출력 성능이 뛰어난 특징을 가진다.

반면 LTO(Linear Tape-Open) 테이프는 구조적 특성으로 인해 랜덤 접근이 매우 불리하며, 순차 접근만 가능하다는 한계를 가진다. 이로 인해 데이터 검색이나 복원에 상대적으로 긴 시간이 소요될 수 있으며, 테이프 라이브러리와 테이프 드라이브 구축을 위한 초기 시스템 비용이 발생한다. 또한 데이터의 지속적인 수정이나 빈번한 갱신보다는 백업 및 장기보존 목적에 보다 적합한 저장매체로 활용되는 경향이 있다.



[그림 9] LTO 테이프 스토리지 구조

❑ 매체 유형별 스토리지 비교

솔리드 스테이트 드라이브(Solid State Drive, SSD)는 속도 및 지연시간이 중요한 환경에 적용하고 하드디스크 드라이브(Hard Disk Drive, HDD)는 대용량 및 경제성이 중요한 환경에 적용하고 그리고 테이프 드라이브는 장기보존, 보안(Air-gap), 최저 비용, 저전력 등이 중요한 환경에 적용한다. 가장 이상적인 스토리지 설계는 솔리드 스테이트 드라이브(Solid State Drive, SSD), 하드디스크 드라이브(Hard Disk Drive, HDD), 테이프 드라이브를 데이터 저장 계층에 따라 목적별로 적용해서 설계하는 것이 바람직할 것이다.

비교기준	SSD	HDD	LTO Tape
구동 방식	전자식(반도체)	기계식(회전 플래터)	자기 테이프 순차 기록
핵심 부품	NAND Flash, 컨트롤러, 버퍼	플래터, 헤드, 스핀들, 액추에이터, 컨트롤러	카트리지, 드라이브, 헤드
접근 방식	임의 접근(Random)	임의 접근(기계적)	순차 접근 (Sequential Only)
데이터 접근 속도	매우 빠름	보통	느림(검색 · 로드 필요)
구입 비용	매우 높음 200 ~ 330 원/GB (3.84TB NVMe 가격 약 80만~100만원 전후 → 약 210~260원/GB)	낮음 40 ~ 70 원/GB (14TB HDD 가격 약 6만~8만원 → 약 43~57원/GB)	매우 낮음 5 ~ 7 원/GB (LTO-9 18TB 매체 약 20~30만원 전후 압축 기준 2.5:1 기대 시 45TB 수용 → 5~7원/GB 수준)
내구성	중간(쓰기수명 제한) 5~10년	낮음(기계 부품 마모) 5~7년	매우 높음 (30년 이상 보존)
비트 오류율(BER)	10-17 ~ 10-18	10-15	10-19 (LTO-9 기준)
충격 민감성	매우 낮음	매우 높음	테이브 낮음 드라이브 높음
소음/발열	없음	있음	테이프 없음 드라이브 있음(약간)
전력 소비	낮음 상시 전력 필요 Idle : 0.6~2W Active : 3~10W (NVMe 최대 20W) 5년 TCO 330 ~ 540 원/Gb	보통 상시 전력 필요 Idle : 4~7W Active : 6~12W 5년 TCO 80 ~ 135 원	테이프 0kW 드라이브 매우 낮음 저장 후 0W Tape Library 사용 시 Idle : 5~10W Active : 20~40W 5년 TCO 27 ~ 68 원
Gb 당 비용	비쌈	저렴	매우 저렴

[표 16] 매체 유형별 스토리지 비교

※ 비트 오류율(BER) : 스토리지 장치가 저장하거나 읽어들인 비트(bit) 중에서 오류가 발생하는 비율을 나타내는 지표로 10^{-n} 의 형태(10^{-15} : 1조(10^{12})비트 당 1개의 오류 발생 가능)로 표현되는 확률 값임

❑ 매체 유형별 스토리지 적용 분야

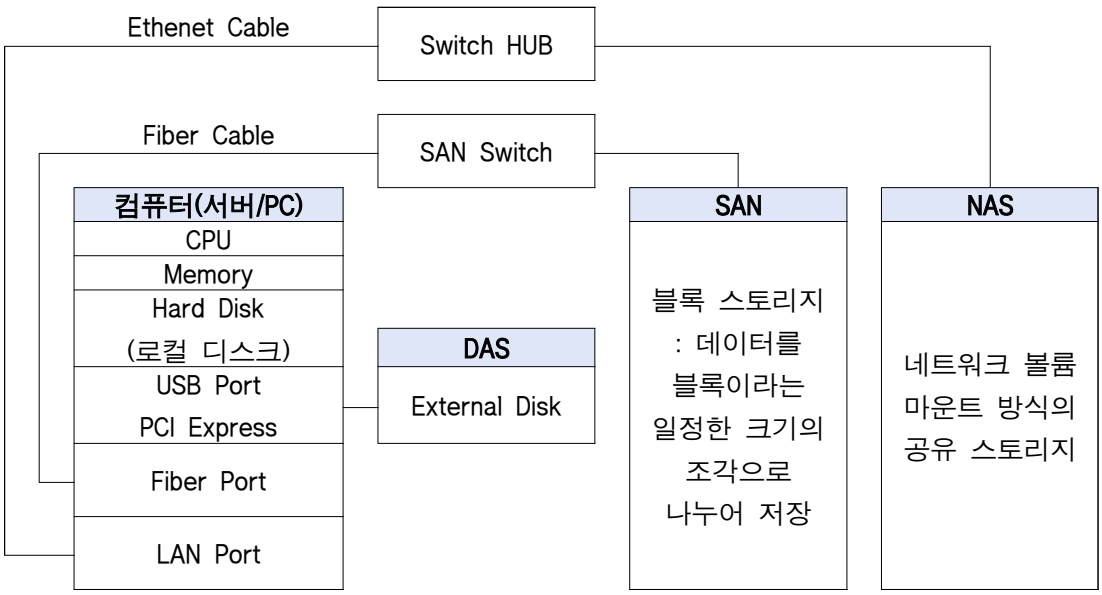
솔리드 스테이트 드라이브(Solid State Drive, SSD)는 높은 처리 성능과 낮은 지연시간이 요구되는 다양한 정보시스템 환경에서 활용된다. 운영계 데이터베이스 관리시스템(DBMS)과 트랜잭션 처리 시스템에서는 OLTP, 인덱스 조회, 초고속 트랜잭션 처리가 필수적인 환경에 솔리드 스테이트 드라이브(Solid State Drive, SSD)가 적용된다. 또한 AI 및 머신러닝 학습과 추론 환경에서는 데이터 셋을 캐시하고 NVMe 기반으로 GPU 서버의 데이터 병목을 해소하기 위한 용도로 활용된다. 가상화 및 클라우드 네이티브 인프라 영역에서는 Ceph RBD, VMware 등의 환경에서 고성능 가상머신과 컨테이너 운영을 지원하기 위해 SSD(Solid State Drive)가 사용되며, 빠른 VM 부팅, 스냅샷 처리, 고속 블록 I/O 제공에 기여한다. 이와 함께 Elastic, Splunk와 같은 검색엔진 및 로그 분석 시스템에서 초당 수천 건의 쓰기와 읽기가 요구되는 경우에도 솔리드 스테이트 드라이브(Solid State Drive, SSD)가 적용된다. 미디어 편집용 워크스테이션과 같이 4K~8K 영상 처리 환경에서 대용량 파일을 반복적으로 실시간 처리해야 하는 경우에도 솔리드 스테이트 드라이브(Solid State Drive, SSD)의 활용도가 높다. 더불어 최신 데이터나 자주 접근되는 데이터를 처리하는 핫 데이터 계층(Storage Tier 0 또는 Tier 1)에서 솔리드 스테이트 드라이브(Solid State Drive, SSD)는 최적화된 저장매체로 활용된다.

하드디스크 드라이브(Hard Disk Drive, HDD)는 비용 효율성과 대용량 저장 특성을 기반으로 한 저장 환경에서 주로 활용된다. 대용량 파일 스토리지 환경에서는 문서, 이미지, 영상, 음원 등 대규모 파일을 저장하기 위한 NAS나 파일 서버에 하드디스크 드라이브(Hard Disk Drive, HDD)가 적용된다. 또한 백업 및 스냅샷의 중간 저장 계층으로서 솔리드 스테이트 드라이브(Solid State Drive, SSD)보다 비용이 저렴하면서 테이프보다 빠른 복원 속도가 요구되는 환경에서도 활용된다. 데이터레이크 환경에서는 분석용 원본 파일, 로그, 중간 결과 등을 저장하는 웜(Warm) 티어에 하드디스크 드라이브(Hard Disk Drive, HDD)가 사용되며, 미디어 서버 환경에서는 VOD 및 스트리밍과 같이 순차 읽기 중심의 접근 패턴에서 높은 비용 효율성을 제공한다. 이외에도 가상머신이나 컨테이너 이미지 저장을 위한 세컨더리 티어로서, 자주 사용되지는 않으나 필요 시 빠른 접근이 요구되는 데이터를 저장하는 용도로 활용된다. 대규모 오브젝트 스토리지 환경에서는 Ceph나 MinIO의 하드디스크 드라이브(Hard Disk Drive, HDD) 풀을 기반으로 Erasure Coding을 적용하여 대규모 저장 확장이 가능한 구조를 구성하는 데에도 하드디스크 드라이브(Hard Disk Drive, HDD)가 적용된다.

테이프 드라이브는 디지털자원의 장기보존을 목적으로 하는 환경에서 주로 활용된다. 문헌, 원본 영상 및 음원, 스캔본 등 디지털자원의 장기보존이 필요한 환경에서 테이프 드라이브는 핵심 저장매체로 적용된다. 또한 주요 시스템의 오프라인 백업(Air-gapped Backup)을 통해 랜섬웨어 대응이 요구되는 경우에도 활용되며, 재해복구(DR) 환경에서는 원격지 보관과 소산(Offsite Storage)을 위한 백업 매체로 사용된다. 아울러 접근 빈도가 낮은 오래된 원본 파일 데이터를 저장하기 위한 대규모 콜드 스토리지 환경에서 페타바이트에서 엑사바이트 규모의 저장소를 구성하는 데 적합한 매체로 활용된다.

3.2.1.2. 연결방식에 따른 스토리지 유형

디지털자원의 저장과 관리를 위한 스토리지 구조는 성능, 확장성, 가용성, 운영 복잡도 측면에서 서로 다른 특성을 지니며, 기관의 업무 성격과 데이터 규모에 따라 적합한 유형을 선택할 필요가 있다. 일반적으로 로컬 스토리지, DAS(Direct Attached Storage), NAS(Network Attached Storage), SAN(Storage Area Network)은 디지털자원 관리 환경에서 대표적으로 활용되는 스토리지 유형으로 구분된다.



[그림 10] 연결 방식에 따른 스토리지 유형

❑ 로컬 스토리지(Local Storage)

로컬 스토리지는 개별 컴퓨터나 서버 내부에 직접 장착되어 메인보드에 연결되는 저장 장치로, SSD(Solid State Drive)나 HDD(Hard Disk Drive) 형태로 운영체제(OS)와 애플리케이션이 직접 접근하는 가장 기본적인 일대일 스토리지 구조이다. 이 방식은 SATA, NVMe, PCIe와 같은 내부 인터페이스를 통해 장비 내부에 연결되며, 네트워크를 거치지 않고 직접 접근이 이루어진다.

로컬 스토리지는 지연시간이 매우 낮고 성능이 뛰어나다는 특징을 지니며, 특정 서버나 PC의 단독 사용 목적에 적합하다. 별도의 네트워크 구성없이 사용할 수 있고 구조가 단순하여 초기 구축비용이 비교적 저렴하다는 장점이 있다. 반면, 저장 장치가 특정 서버에 고정되어 있어 확장성이 부족하고 여러 서버 간 데이터 공유가 어렵다는 한계를 가진다. 또한 서버 장애가 발생할 경우 해당 장비에 저장된 데이터가 함께 손실될 위험이 상대적으로 높다.

❑ DAS(Direct Attached Storage)

DAS(Direct Attached Storage)는 컴퓨터나 서버에 USB, SAS, Thunderbolt 등의 케이블을 통해 직접 연결되는 외장형 스토리지로, 부족한 저장 공간을 가장 간단하게 확장할 수 있는 방식이다. DAS(Direct Attached Storage)는 네트워크를 경유하지 않고 특정 서버 또는 소수의 서버에 직접 연결되어 운영되며, 물리적으로 가까운 위치에서 데이터 접근이 가능하다는 특성을 지닌다.

이 방식은 로컬 스토리지에 비해 상대적으로 확장성이 높고, NAS(Network Attached Storage)에 비해 구조가 단순하며 비용이 저렴하다는 장점을 가진다. 그러나 네트워크 기반 공유 기능을 제공하지 않기 때문에 다수 서버 간 데이터 공유에는 적합하지 않으며, 대규모 데이터 증가에 따라 외부 저장 장치를 지속적으로 추가할 경우 서버 운영 효율이 저하될 수 있다. 또한 특정 서버가 중단될 경우 해당 서버에 연결된 DAS 역시 함께 사용할 수 없다는 구조적 제약이 존재한다.

❑ NAS(Network Attached Storage)

NAS(Network Attached Storage)는 네트워크(LAN)에 연결되어 파일 수준(File-level)으로 데이터를 공유하는 저장 장치로, 서로 다른 네트워크 클라이언트에 대해 데이터 접근 권한을 제공한다. NAS(Network Attached Storage)는 일반적으로

SMB/CIFS나 NFS와 같은 파일 공유 프로토콜을 사용하며, 파일 시스템을 NAS(Network Attached Storage) 장비 자체가 관리한다.

NAS(Network Attached Storage)의 주요 특징은 여러 사용자나 컴퓨터가 네트워크를 통해 동시에 접근할 수 있다는 점이며, 랜 기반 공유 스토리지로서 협업 환경에 적합하다. 웹 기반 관리 인터페이스, 접근 제어 목록(ACL), 스냅샷 기능 등을 제공하여 관리가 비교적 용이하고 구축비용이 낮으며 확장성과 클러스터링 구성이 가능하다는 장점을 가진다. 반면, 파일 수준 접근 방식으로 인해 블록 기반 스토리지에 비해 지연시간이 증가할 수 있으며, 네트워크 속도와 품질에 따라 파일 전송 성능이 영향을 받는다. 또한 데이터베이스와 같이 대규모 I/O가 요구되는 환경에서는 효율성이 떨어질 수 있다.

❑ SAN(Storage Area Network)

SAN(Storage Area Network)은 파일이 아닌 블록 단위(BLCk-level) 입출력을 기본으로 하는 스토리지 네트워크로, 각 블록이 저장된 위치에 대한 주소를 기반으로 서버의 요청에 따라 데이터를 재구성하여 전달하는 구조를 가진다. SAN은 별도의 고속 데이터 전달 통로를 통해 스토리지 시스템에 접근하므로, 일반 네트워크 트래픽의 영향을 받지 않고 신속한 데이터 액세스를 제공한다.

SAN(Storage Area Network) 환경에서는 여러 서버가 하나의 스토리지 풀을 공유할 수 있으며, 서버 입장에서는 SAN(Storage Area Network) 스토리지가 로컬 디스크처럼 인식된다. 이를 위해 Fibre Channel(FC)과 같은 고속 전용 네트워크 기술과 SAN(Storage Area Network) 스위치를 사용하여 별도의 스토리지 네트워크를 구성한다. 이러한 특성으로 인해 SAN(Storage Area Network)은 고성능, 고대역폭, 매우 낮은 지연시간을 요구하는 대규모 데이터베이스, 가상화 환경, 금융 시스템 등에 적합하다.

다만 SAN(Storage Area Network)은 구축비용이 높고, FC 스위치 구성, SAN(Storage Area Network) 존닝(zoning) 등 관리가 복잡하며, 운영과 관리를 위해 전문 인력이 필요하다는 단점을 가진다. 그럼에도 불구하고 고가용성(HA) 구조를 적용할 경우 이중화와 확장성이 뛰어나 대규모 기업 시스템에서 핵심 인프라로 활용된다.

3.2.1.3. 접근 방식에 따른 스토리지 유형

컴퓨터 시스템에서 디지털 파일에 접근하는 방식에 따라 스토리지는 일반적으로 파일 스토리지, 블록 스토리지, 오브젝트 스토리지로 구분된다. 이들 스토리지 유형은 데이터의 저장 구조와 접근 방식, 확장성, 성능 특성에서 차이를 보이며, 디지털자원의 성격과 관리 목적에 따라 적합성이 달라진다.

❑ 파일 스토리지(File Storage)

파일 스토리지는 데이터를 폴더와 파일로 구성된 계층적 구조 안에 저장하는 방식으로 각 파일에는 파일명, 파일 유형, 생성 시점 및 최종 수정 시점과 같은 기본적인 메타데이터가 부여된다. 데이터는 디렉토리와 하위 디렉토리 구조에 따라 조직되며, 시스템은 파일 경로를 통해 원하는 파일에 접근한다.

이러한 구조는 물리적 파일 캐비닛에 종이 문서를 보관하는 방식에 비유할 수 있다. 서랍은 디렉토리에, 폴더는 하위 디렉토리에 해당하며, 이용자는 경로를 따라가며 필요한 파일을 찾는다. 파일 스토리지는 비교적 소량의 데이터를 정리하고 관리하는 데 용이하며, 연구자나 업무 담당자가 이해하고 운영하기 쉬운 구조를 가진다. NAS(Network Attached Storage)가 대표적인 파일 스토리지 방식이다.

그러나 디지털자원의 규모가 증가함에 따라 파일과 디렉토리 수가 기하급수적으로 늘어나면, 메타데이터 조회와 탐색 속도가 점차 저하되고 관리 부담이 증가한다. 백업, 스냅샷, 마이그레이션과 같은 운영 작업 역시 데이터 규모 증가에 따라 복잡성과 비용이 함께 상승하게 된다. 이러한 특성으로 인해 파일 스토리지는 대규모 장기보존보다는 디지털자원의 생산, 수집, 가공 과정에서 사용하는 작업용 또는 서비스용 스토리지로서의 적합성이 더 높다. 마스터 파일의 최종 보존 계층으로 활용하기에는 확장성과 관리 비용 측면에서 한계를 가진다.

❑ 블록 스토리지(Block Storage)

블록 스토리지는 파일 스토리지의 한계를 보완하기 위해 등장한 방식으로 데이터를 동일한 크기의 데이터 블록으로 분할하여 개별 주소에 저장한다. 파일 시스템의 계층적 구조에 의존하지 않으며, 운영체제가 고유 주소를 기반으로 데이터 블록을 재조합하여 파일 형태로 제공한다. 이 과정에서 디렉토리 탐색이 필요하지 않기 때문에 접근 효율성과 성능이 향상된다.

블록 스토리지는 짧은 지연 시간과 높은 입출력 성능(IOPS)을 요구하는 환경에 적합하며, 트랜잭션 데이터베이스, 가상 머신 디스크, 핵심 업무 시스템과 같은 애플리케이션 중심의 스토리지로 활용된다. SAN(Storage Area Network)과 DAS(Direct Attached Storage)가 대표적인 블록 스토리지 유형에 해당한다.

다만 블록 스토리지는 구축과 운영이 상대적으로 복잡하고 비용이 높으며, 대규모 비정형 데이터(원문 파일, 이미지, 영상 등)를 장기간 대량 보존하는 용도로는 효율성이 낮다. 이러한 특성으로 인해 블록 스토리지는 디지털자원 그 자체를 저장하기보다는, 전자도서관 서비스 데이터베이스, 서지·메타데이터 관리 시스템, 검색 인덱스, 로그 분석 시스템 등 디지털자원을 관리·제공하는 애플리케이션을 지원하는 스토리지로서 필수적인 역할을 수행한다.

❑ 오브젝트 스토리지(Object Storage)

오브젝트 스토리지는 데이터를 파일이나 블록 단위가 아닌 주소 지정이 가능한 독립적인 오브젝트 단위로 저장하는 방식이다. 각 오브젝트는 데이터 자체, 해당 데이터에 대한 메타데이터, 그리고 고유 식별자를 포함하는 완결된 단위로 구성된다. 전통적인 파일 시스템과 달리 복잡한 디렉토리 계층 구조를 사용하지 않으며, 구조적으로 평면적인 환경에서 대규모 확장이 가능하도록 설계된다.

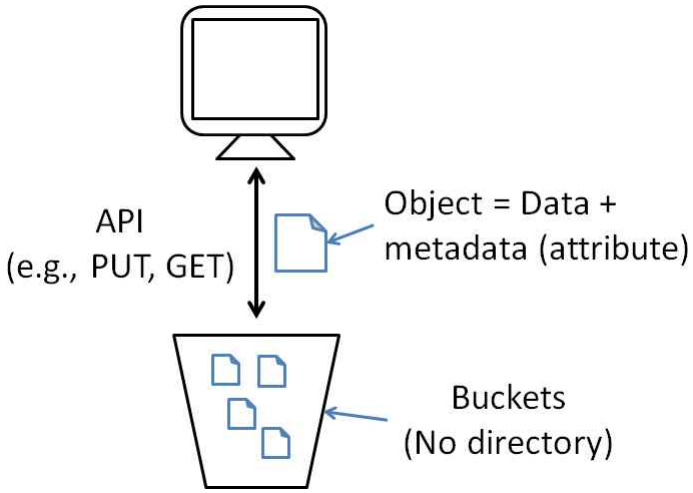
오브젝트 스토리지는 대규모 비정형 데이터에 대해 비용 효율적이며, 무제한에 가까운 수평적 확장성과 높은 내구성을 제공한다. 복제, 삭제 코딩, 다중 데이터센터 분산과 같은 기술을 통해 디스크 장애, 노드 장애, 심지어 사이트 단위 장애까지 견딜 수 있는 복원력을 갖춘 설계가 가능하다. 또한 오브젝트 단위에 다양한 메타데이터를 직접 부착할 수 있어, 생산 기관, 수집 유형, 저작권 상태, 보존 등급, 포맷 정보 등 보존·관리 메타데이터를 체계적으로 관리할 수 있다.

오브젝트 스토리지는 구조적으로 오브젝트의 부분 수정이 불가능하므로, 데이터 변경이 필요한 경우 기존 오브젝트를 삭제하고 새로운 오브젝트를 생성하는 방식으로 처리된다. 이러한 특성으로 인해 자주 수정되는 데이터보다는, 수집 이후 거의 변경되지 않는 정적 데이터에 적합하다. 그러므로 WORM(Write Once Read Many) 특성을 지닌 디지털 장기보존 자료에 특히 적합한 구조라 할 수 있다.

국립중앙도서관의 디지털자원은 문서(PDF/A), 이미지(TIFF), 음원(WAV,MP3), 영상(MXF, MP4), 웹 자료(WARC, Web ARChive) 등 다양한 비정형 데이터로 구성되어 있으며, 보존 단계로 이관된 이후에는 삭제나 수정이 거의 발생하지

않는다. 또한 비압축 포맷을 포함한 대규모 데이터 축적에 따라 저장 용량이 TB, PB, 나아가 EB 단위로 급격히 증가하는 특성을 보인다. 이러한 환경에서는 OAIS(Open Archival Information System) 기반 장기보존전략을 적용하기 위한 보존 메타데이터(PREMIS, METS), 무결성 검증(fixity), 포맷 식별과 같은 기능이 필수적으로 요구된다.

이와 같은 요구 사항은 오브젝트 스토리지가 지닌 확장성, 비용 효율성, 내구성, 메타데이터 중심 구조와 높은 정합성을 가지며, 국가 단위 디지털자원을 장기간 또는 영구적으로 보존해야 하는 환경에서 핵심적인 스토리지 계층으로 기능할 수 있음을 시사한다.



[그림 11] 오브젝트 스토리지의 구조

3.2.1.4. 국립중앙도서관 디지털자원 보존 스토리지 적용 방안

국립중앙도서관 디지털자원 관리에서 스토리지 유형은 성능, 확장성, 공유 방식, 운영 복잡도 측면에서 서로 상이한 특성을 지닌다. 따라서 로컬 스토리지, DAS, NAS, SAN은 디지털자원의 성격과 기관의 운영 목적에 따라 선택적으로 활용될 필요가 있으며, 특히 대규모 디지털 자산을 장기적으로 관리해야 하는 기관 환경에서는 단일 스토리지 유형에 의존하기보다 용도와 계층에 따라 복합적으로 구성하는 전략이 요구된다. 이러한 관점에서 접근 방식에 따른 스토리지 유형을 비교하면, 파일 스토리지는 업무 및 서비스 중심의 운영 환경에, 블록 스토리지는 고성능 애플리케이션 지원에, 오브젝트 스토리지는 대규모 비정형 데이터의 장기보존에 각각 적합한 특성을 지닌다. 따라서 디지털 장기보존을 수행하는 기관은 업무·서비스·보존 목적에 따른 계층적 스토리지 전략을 수립하는 것이 바람직하다.

국립중앙도서관의 보존 스토리지 전략은 단일 기술 선택의 문제가 아니라 보존용 스토리지(Storage for Preservation)와 서비스/업무용 스토리지(Storage for Access & Operation)를 분리하여 설계하는 계층형·역할 분리형 접근으로 구체화될 필요가 있다. 그러므로 추가 중심·장기보존·대규모 비정형 데이터는 오브젝트 스토리지를 중심으로 설계하고 고성능 트랜잭션 처리와 검색 인덱스, 업무 시스템 데이터베이스(DB)와 같은 워크로드는 블록 스토리지(SAN/DAS) 중심으로 설계하며, 업무용 파일 공유·중간 산출물·임시 작업 데이터는 파일 스토리지(NAS)를 활용하는 방식으로 기능별 적합성을 확보해야 한다.

❑ 계층형 스토리지 구조

국립중앙도서관의 디지털자원 장기보존을 위한 스토리지 체계는 보존 목적과 활용 목적을 구분한 계층형 스토리지 구조로 설계될 필요가 있다. 이 구조는 보존 안정성, 서비스 성능, 운영 효율성을 동시에 확보하기 위해 각 계층별로 역할과 기술적 특성을 명확히 분리하는 것을 전제로 한다.

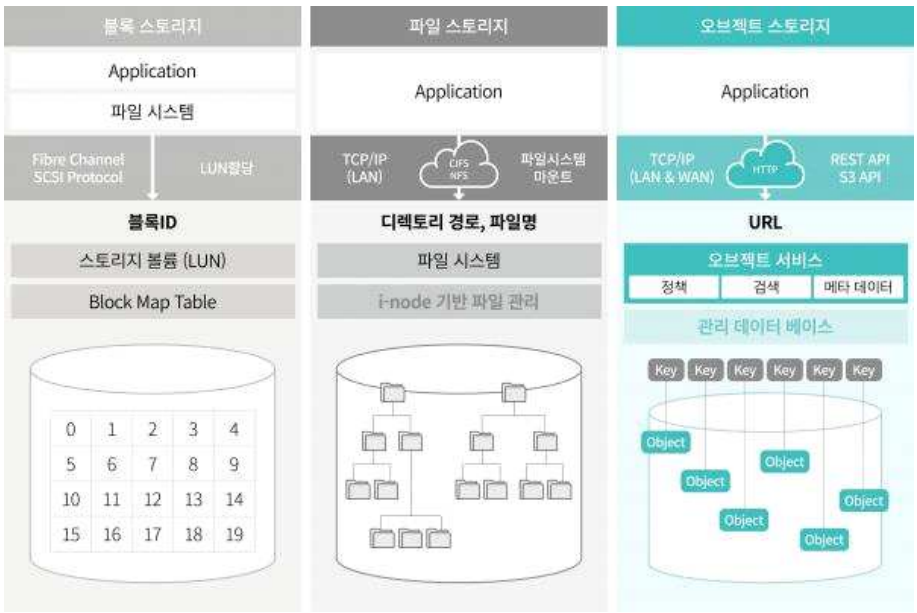
우선 보존 마스터 계층(Preservation Master Tier)은 디지털자원의 장기보존을 책임지는 핵심 계층으로 오브젝트 스토리지를 기반으로 구성된다. 이 계층에는 디지털 납본 자료(전자책, 전자저널, 웹 아카이브 등)의 마스터 파일과 디지털화 사업을 통해 생성된 고해상도 원본 자료(TIFF 이미지, WAV 음원, 무손실 영상 파일 등)가 저장된다. 보존 마스터 계층은 포맷 식별 결과, 기술 메타데이터, 보존 메타데이터와의 연계를 전제로 하며, 삭제 및 덮어쓰기를 제한하는 WORM(Write Once Read Many) 정책을 적용하여 데이터의 안정성을 확보한다. 또한 정기적인 무결성 검증(Fixity Check), 즉 체크섬 기반 검증을 수행하고 오류가 발생할 경우 자동 복구가 가능하도록 설계된다.

다음으로 서비스·액세스 계층(Access Tier)은 이용자 서비스 제공을 위한 계층으로 파일 스토리지(NAS)와 캐시 구조를 결합하여 구성된다. 이 계층에는 이용자 열람과 다운로드를 목적으로 생성된 파생본(access copy), 예를 들어 JPEG 이미지, MP3 음원, MP4 영상 파일 등이 저장된다. 웹서비스 및 검색 시스템이 빠르게 접근할 수 있도록 파일 단위 접근에 최적화된 스토리지 구조를 갖추며, 필요 시 보존 마스터 계층의 오브젝트 스토리지로부터 원본 파일을 불러와 새로운 액세스 파일을 재생성할 수 있는 연계 구조를 포함한다. 이를 통해 보존 원본을 직접 서비스에 노출하지 않으면서도 안정적인 서비스 제공이 가능해진다.

업무・운영 계층(Operations Tier)은 내부 업무 시스템과 운영을 지원하는 계층으로 블록 스토리지(SAN 또는 DAS)를 중심으로 구성된다. 이 계층에는 서지 및 메타데이터 관리 데이터베이스, 이용자 정보 데이터베이스, 검색 인덱스 서버, 로그・통계・분석용 데이터베이스 등이 포함되며, 고성능 I/O가 요구되는 워크로드를 안정적으로 처리하는 것이 주요 목적이다. 블록 스토리지는 짧은 지연 시간과 높은 입출력 성능을 제공하므로, 대규모 서비스 트래픽과 내부 업무 처리를 동시에 지원하는 데 적합하다.

마지막으로 아카이브・콜드 계층(Cold/Archive Tier)은 조회 빈도가 극히 낮은 장기보존 자원을 대상으로 하는 계층이다. 이 계층은 저비용 오브젝트 스토리지 또는 테이프 라이브러리를 활용하여 구성할 수 있으며, 오브젝트 스토리지 내에서도 저성능・저비용 티어로 데이터를 이동시키는 방식이 가능하다. 필요에 따라 테이프 또는 오프라인 미디어와 연계한 이중 아카이브 전략을 적용함으로써, 장기적인 비용 절감과 재해 대비를 동시에 고려할 수 있다.

이와 같은 계층형 스토리지 구조는 디지털자원의 성격과 이용 패턴에 따라 저장 계층을 구분함으로써, 보존 안정성과 서비스 효율성을 동시에 확보할 수 있는 기반을 제공한다. 특히 보존 마스터 계층을 중심으로 한 오브젝트 스토리지 기반 구조는 대규모 비정형 디지털자원의 장기보존 요구와 높은 정합성을 가지며, 국립중앙도서관의 디지털 장기보존전략을 기술적으로 뒷받침하는 핵심 인프라로 기능할 수 있다.



[그림 12] 데이터 접근 방식에 따른 스토리지 구분

❑ 파일 스토리지 방식 적용방안

파일 스토리지 방식으로 계층형 디렉토리 구조를 형성한 경우, 추가적인 이벤트가 발생한 경우 정보를 추가할 수 없게 되지만, 오브젝트 스토리지의 경우 계층형 디렉토리 구조가 형성되지 않고 개별 파일과 메타데이터만 존재하기 때문에 메타데이터에 필요한 정보를 추가하는 방식으로 쉽게 확장할 수 있다. 즉 오브젝트 스토리지는 데이터의 추가 정보를 손쉽게 태깅(Tagging)하여 데이터 구조화하고 이를 기반한 데이터 검색 편의성을 제공한다



[그림 13] 파일 스토리지와 오브젝트 스토리지 비교

국립중앙도서관 디지털자원을 보존하기 위해서는 문서, 이미지, 영상, 음원 등 대량의 비정형 데이터를 원본 그대로 저장하고 이를 장기적으로 안정적으로 제공할 수 있는 스토리지 체계가 필요하다.

현재 국립중앙도서관에서는 디지털 형식으로 생산·수집된(Born Digital) 자료와 아날로그를 디지털로 변환한 자료를 서비스용 파일은 KOLIS를 통해 일괄 등록 관리하는 데 비해, 보존이 필요한 원본 파일의 경우 스토리지 용량의 한계로 인하여 각 부서에서 저장·관리하고 있다. 이로 인해 디지털자원을 보존 부서로 이관하고 체계적으로 축적할 수 있는 통합 스토리지 플랫폼이 부재하며, 부서 간 데이터 공유와 재이용이 비효율적인 구조를 보이고 있다.

또한 디지털자원을 외부 저장소(상용 클라우드, 외부 호스팅 등)에 분산 저장할 경우, 장기적으로 비용 부담이 크게 증가할 우려가 있으며, 외부 서비스 장애나 사업자 변경에 따른 접근성 저하 및 데이터 유출 위험도 존재한다. 물리적인 이동식 저장장치(외장하드, USB 등)는 분실·도난 시 보안 위협과 자료 유실 위험을 증대시킨다.

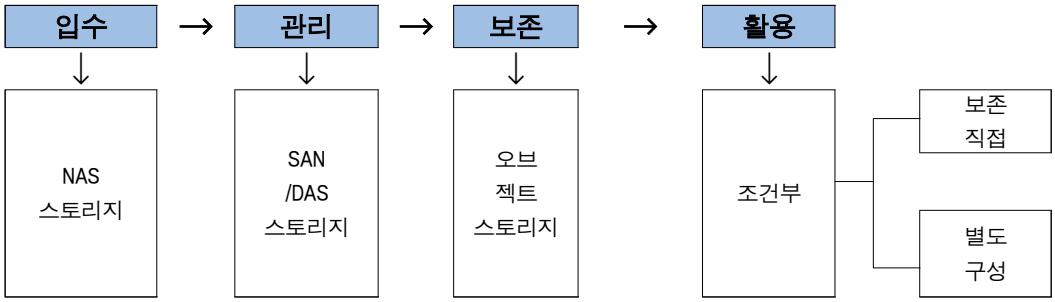
보존 목적의 스토리지와 활용 목적의 스토리지를 물리적으로 완전히 분리할 경우, 하드웨어・소프트웨어・운영 인력 등 이중 투자로 인한 비용 증가가 발생할 수 있다. 따라서 운영 효율성과 비용 효율성을 함께 고려하여, 단일 통합 스토리지 플랫폼을 기반으로 디지털자원을 등록・관리하고 동일한 기반 위에서 보존과 활용을 동시에 지원할 수 있는 공동 활용 인프라를 구축할 필요가 있다.

국립중앙도서관은 디지털자원의 생산・수집 단계부터 보존・서비스 단계까지의 전 과정을 통합 관리 할 수 있는 디지털 스토리지 플랫폼을 구축할 필요가 있다. 이 플랫폼은 물리적으로는 하나의 통합 스토리지 인프라를 기반으로 하되, 논리적으로는 보존용 마스터 계층과 서비스・업무용 계층을 구분하여 관리함으로써, 보존의 안정성과 서비스의 효율성을 동시에 달성하는 구조를 지향해야 한다. 이를 통해 부서 간 개별 저장 관행에서 발생하는 중복 투자와 비효율을 줄이고, 디지털자원의 체계적인 축적과 장기적인 안전 보존, 그리고 국민 대상 서비스 확대를 위한 기반을 동시에 마련할 수 있다.

대규모 비정형 데이터 관리에 최적화된 오브젝트 스토리지 방식으로 통합 스토리지 플랫폼을 구축해서 클라우드 환경에서 효율적인 데이터 관리를 가능하게 한다.

❑ 디지털자원 생애주기별 스토리지 적용방안

국립중앙도서관의 디지털자원은 생산·입수 단계부터 보존과 서비스 단계에 이르기까지 서로 다른 기술적 요구를 지닌다. 이에 따라 디지털자원의 생애주기(Lifecycle)에 적합한 스토리지 기술을 단계별로 선택적으로 적용하는 것이 필수적이다. 입수・관리 단계에서는 업무 효율성과 협업 편의성이 중요하므로 전통적 파일 스토리지와 블록 스토리지 기반의 저장 환경을 활용하고 보존 단계에서는 장기보존과 안정성이 핵심 요건이므로 오브젝트 스토리지를 중심으로 한 통합 보존 스토리지 플랫폼을 구축하여 운영해야 한다. 활용 단계에서는 민원・서비스 제공 요구와 접근 패턴을 고려하여 보존 스토리지를 직접 연계하여 사용할지, 또는 별도의 활용 스토리지 계층을 구성할지를 결정하는 방식으로 차별화된 스토리지 전략을 마련한다.



[그림 14] 디지털자원 생애주기별 스토리지 기술

우선, 개별 부서에서 디지털자원을 입수하고 관리하는 단계에서는 다양한 경로를 통해 생산되거나 외부에서 전달된 디지털 파일을 신속하게 저장하고 처리해야 한다. 온라인자료(전자책, 웹툰, 웹소설, 오디오북 등)는 납본시스템 또는 이메일 등을 통해 전달되며, 국립중앙도서관 내부의 원문 DB 구축 사업에서 생성된 파일이나 자료보존연구센터의 매체 변환 작업을 통해 생성된 고해상도 원본 자료 역시 이 단계에서 처리된다. 이 단계의 특징은 업무 과정에서 파일을 열람하고 수정하거나 여러 직원이 협업하는 경우가 잦고, 임시 저장 및 반복 검토가 필요한 경우가 많다는 점이다. 이러한 특성상 입수 단계에서는 파일 스토리지(NAS)를 활용하여 협업 편의성과 접근성을 높이는 것이 효과적이며, 관리 단계에서는 통합 자료관리 시스템을 통해 서지 메타데이터와 기술 메타데이터가 데이터베이스(DBMS) 형태로 저장되므로 SAN 또는 DAS 기반의 블록 스토리지를 통해 고성능 I/O 요구를 충족하는 것이 적합하다.

디지털자원이 현용 단계에서 준현용 또는 보존 단계로 전환되면 보존 부서로 이관되며, 이 시점부터는 데이터의 변경 가능성이 거의 없고 장기보존을 기본 전제로 관리가 이루어진다. 보존 단계에서는 대규모 비정형 데이터(이미지, 음원, 영상, 웹아카이브 등)를 안정적으로 저장하고 무결성을 검증할 수 있어야 하므로 오브젝트 스토리지 기반의 통합 보존 스토리지 플랫폼을 적용하는 것이 바람직하다. 오브젝트 스토리지는 Erasure Coding 기반의 고내구성, 체크섬 기반 무결성 검증, 버전 관리, 장기적 확장성 및 경제성 등 장기보존 환경에 필요한 주요 요건을 충족한다. 또한 S3 API(Application Programming Interface) 호환성을 통해 다양한 업무 시스템과의 연동이 용이하며, AIP(Archival Information Package) 형태의 아카이브 패키지를 체계적으로 관리하기에 적합한 구조를 제공한다.

보존 스토리지에 저장된 자원은 이후 이용자 서비스나 연구용 제공 등 다양한 활용 단계로 이어지며, 이 요구에 따라 두 가지 방식을 선택 적용한다. 활용

목적의 파일 변환이 필요 없고 데이터 손상 위험이 낮으며 비교적 접근 빈도가 낮은 경우에는 보존용 오브젝트 스토리지를 직접 연계하여 사용하는 것이 가능하다. 이 방식은 스토리지 중복 구성을 최소화하여 운영비용을 절감할 수 있으며, 변환 작업 없이 즉시 활용할 수 있어 효율적이다. 또한 동일한 파일을 보존과 서비스에서 공유할 수 있으므로 메타데이터 및 버전 일관성 유지에도 유리하다.

반면, 파일을 자주 열람하거나 대량 조회가 발생하는 서비스 환경에서는 보존 스토리지를 직접 활용할 경우 성능 저하 또는 원본 데이터에 부담이 발생할 수 있다. 특히 이미지·영상 처리 시스템처럼 I/O가 많은 처리 환경에서는 별도의 활용 전용 스토리지가 필수적이다. 또한 서비스 제공 목적으로 고압축·저용량 파일(JPEG, MP4, MP3 등)을 생성하거나, OCR·STT 분석 결과물 또는 썸네일 이미지처럼 파생본을 제공해야 하는 경우에도 활용용 오브젝트 스토리지 또는 파일 스토리지(NAS)를 별도로 구축하는 것이 적합하다. 이는 보존 정책에서 원본 변환을 제한하거나 변환본의 보존 스토리지 저장을 금지하는 경우 정책적 요구와도 일치한다. 별도의 활용 계층을 구성함으로써 보존 스토리지의 안정성을 확보하고 서비스 성능을 최적화하며, 파생본의 수명주기 관리도 보다 유연하게 운영할 수 있다.

이와 같이 국립중앙도서관의 디지털자원 생애주기별 스토리지 적용은 입수·관리 단계에서는 파일·블록 스토리지, 보존 단계에서는 오브젝트 스토리지, 활용 단계에서는 활용 패턴에 따른 보존 스토리지 직접 연계 또는 별도 활용 스토리지 구성을 선택하는 형태로 운영되어야 한다. 이러한 생애주기 기반 스토리지 전략은 업무 효율성과 보존 안정성을 동시에 확보하며, 장기적으로 서비스 품질 향상과 비용 효율성 확보에 기여할 수 있다.

❑ 스토리지 기반 3단계 디지털자원 보존 구조

디지털자원의 안정적 보존과 효율적 활용을 동시에 달성하기 위해 스토리지를 1단계 Hot Tier(SSD), 2단계 Warm Tier(HDD), 3단계 Cold Tier(LTO Tape)의 3계층으로 구분하여 운영할 수 있다. 이 구조는 데이터의 접근 빈도, 보존 기간, 처리 성능 요구, 비용 구조, 재난 대비 수준 등을 종합적으로 고려하여 설계된 계층형(HSM: Hierarchical Storage Management) 보존전략이다.

○ 1단계 Hot Tier(SSD 계층)

먼저 1단계 Hot Tier(SSD 계층)는 디지털자원의 검색·색인·AI 기반 처리 및

서비스 제공을 위한 가장 빠른 작업용 공간으로 운영된다. 이 계층은 특히 메타데이터·색인 데이터와 같이 처리 빈도가 높은 자료를 고속으로 다루기 위해 NVMe SSD(Solid State Drive) 기반으로 구성되며, RAID 1+0을 적용해 성능과 안정성을 동시에 확보한다. 또한 AI/ML 서비스를 운영하는 환경에서는 GPU 서버의 로컬 스토리지로도 활용되어 AI 추론 과정에서 생성되는 중간 데이터를 처리한다.

이 계층의 주요 기능은 고속 메타데이터 처리(DB 및 색인 서버), AI 기반 자동 태깅·포맷 식별·OCR/STT 분석·품질 진단, 웹 기반 검색·열람 서비스에서 높은 접근 빈도를 보이는 데이터를 캐싱하는 역할이다. 또한 디지털자원이 수집되어 들어오는 순간, 원본 파일이 1차적으로 임시 저장되는 Ingest Buffer의 역할도 수행한다.

SSD(Solid State Drive) Tier에 저장되는 데이터는 주로 메타데이터 DB, 검색엔진(ElasticSearch, Solr, OpenSearch)의 색인 데이터, AI 처리 과정에서 생성되는 임시 파일(OCR/STT 결과, 태깅 결과, 임베딩 벡터 등)이다. 특히, 임베딩(Embedding)은 텍스트·이미지·오디오 등 비정형 데이터를 인공지능 모델이 처리할 수 있도록 고차원 벡터 공간에 수치적으로 표현한 결과물이다. 이는 원본 데이터로부터 알고리즘적으로 생성된 파생 메타데이터(derived metadata)에 해당하며, 자동 메타데이터 생성 및 의미 기반 검색(semantic search)에 활용된다. 이러한 임베딩 데이터는 PREMIS(Preservation Metadata: Implementation Strategies) 데이터 사전에서 객체의 기술적 특성(Object Characteristics) 요소를 통해 관리할 수 있다.

SSD(Solid State Drive) 계층의 보존 목적은 무엇보다 고속 처리와 서비스 품질을 확보하는 것이다. 메타데이터 수정이나 색인 재생성 같은 핵심 작업은 즉각적으로 반영되어야 하므로 빠른 입출력 성능이 필수적이다. 다만 SSD(Solid State Drive)는 비용이 높기 때문에 용량을 최소화하여 운영하며, 데이터는 주기적으로 스냅샷을 생성한 뒤 HDD(Hard Disk Drive) 계층으로 자동 오프로드한다.

주요 기능	• 고속 메타데이터 처리 (DB, 색인 서버 등) • AI/ML 기반 파일 자동 태깅, 포맷 식별, OCR/STT, 품질 진단 • 웹 뷰어 · 검색 서비스에서 가장 많이 사용되는 핫 데이터 캐싱 • 디지털자원 수집 시 1차 임시 저장(Ingest Buffer)
기술 요소	• NVMe SSD (PCIe 4/5 기반) • SSD RAID 1+0 구성(고성능 + 안정성) • AI/ML 서비스를 제공하는 경우 AI 추론 서버(GPU)의 로컬 스토리지로도 사용
특징	• 속도 최우선 • 주기적 스냅샷 및 HDD 계층으로 자동 오프로드 • 비용이 가장 비싸므로 용량 최소화 운영
저장 대상	• 메타데이터 DB : 도서 · 기록 · 영상 · OCR 텍스트 메타데이터 등의 저장 • 검색엔진(ElasticSearch, Solr, OpenSearch 등)을 위한 검색 색인 데이터(Indexed Data) • AI 처리용 중간 데이터 : OCR/STT 결과, 임베딩(Embedding), 태깅 임시 파일 ※ 임베딩 : 텍스트·이미지·오디오 등 복잡한 비정형 데이터를 수학적으로 정의된 고차원 벡터 형태(숫자 배열)로 변환한 표현 방식(AI 기반 검색 및 자동 메타데이터 생성에 활용할 수 있으며, 임베딩은 파생 메타데이터(derived metadata)로 분류해서 OAIS PREMIS의 “Object Characteristics” 영역에 저장 가능) • 디지털자원의 Ingest(수집) 파일의 임시 저장 공간
보존 목적	• 고속 처리와 서비스 성능 확보 • 핫데이터(고빈도 데이터) 중심 운영 • 메타데이터 변경 시 즉각적 반영

[표 17] 디지털자원 보존 구조 1단계 Hot Tier

❍ 2단계 Warm Tier(HDD 계층)

HDD(Hard Disk Drive) 기반의 2단계 계층은 국립중앙도서관 디지털자원의 장기보존과 상시 활용을 동시에 지원하는 메인 저장소로서, 대규모 디지털 파일을 안정적으로 보관하고 서비스하기 위한 핵심 인프라이다. 이 계층에는 이미지(TIFF, JPEG2000), 영상(MKV, FFV1, MP4 서비스본), 음원(WAV, FLAC), 문서(PDF/A, EPUB, XML) 등 다양한 유형의 디지털자원 원본 파일과 서비스본 파일이 저장되며, 웹 서비스나 API(Application Programming Interface)를 통해 제공되는 대부분의 콘텐츠가 바로 이 계층에 기반한다.

이 계층은 PB급까지 수평 확장이 가능한 HDD(Hard Disk Drive) 기반 오브젝트 스토리지(Ceph, MinIO, Scality 등)를 중심으로 구성되며, RAID 6 또는 이레이저 코딩(EC 4+2, 8+3 등)을 적용하여 높은 고가용성과 장애 허용성을 확보한다. 오브젝트 스토리지 특성상 데이터와 메타데이터를 분리해 저장하고 모든 접근을 표준화된 S3 API(Application Programming Interface)로 처리하기 때문에 서비스의

확장성, 외부 시스템과의 호환성이 뛰어나며 AI 학습·분석 워크로드에도 효과적으로 활용할 수 있다.

HDD(Hard Disk Drive) 계층의 가장 큰 장점은 비용 대비 용량 효율성으로 수십 PB 규모의 장기 저장 용도로 적합하다. 특히 24/7 온라인 접근을 보장하도록 설계되어 있어, 연중무휴로 중단 없이 데이터를 읽고 쓸 수 있다는 점은 기관의 온라인 서비스 안정성과 직결된다. 다만 HDD(Hard Disk Drive)는 기계식 저장장치라는 특성상 주기적인 디스크 교체와 재구축이 필요하며, LTO(Linear Tape-Open)와 같은 오프라인 장기보존 매체보다 물리적 안정성이 낮기 때문에 정기적인 무결성 검증(Fixity Check)이 반드시 수행되어야 한다.

이 계층에는 디지털 원본뿐 아니라 DB 및 메타데이터 백업, 파일 백업도 함께 저장된다. 국립중앙도서관의 디지털자원 보존은 단순히 장기보존만을 목표로 하지 않고 이용자 서비스·내부 업무·AI 자동화 등 지속적 활용을 전제한 보존체계를 지향한다. 따라서 원본 파일과 서비스본 파일을 서로 다른 저장소에 분리하여 운영하는 방식은 효율성과 관리성을 저하시킬 수 있으며, 동일한 HDD(Hard Disk Drive) 계층 안에서 논리적 구분을 통해 원본과 서비스본을 통합적으로 관리하는 방식이 기관 운영 환경에 보다 적합하다. 또한 보존부서에서 직접 디지털자원을 생산하는 경우에는 원본용 고용량 비압축 파일과 서비스용 저용량 압축 파일을 별도로 만들기도 하지만, 외부에서 수집된 자원의 경우 원본과 서비스본의 파일 구성이 동일하게 입수되는 경우가 많다. 이러한 다양한 형태의 디지털자원을 일관된 구조로 관리하기 위해서도, 본 계층을 중심으로 한 통합 저장 전략이 유효하다.

결과적으로 HDD(Hard Disk Drive) 기반 2단계 계층은 디지털자원의 온라인 보존, 제공, 백업, AI 활용을 모두 아우르는 기관의 핵심 저장 계층으로 기능하며, SSD(Solid State Drive) 계층의 고속 처리 작업과 LTO(Linear Tape-Open) Tape 계층의 장기보존 역할 사이에서 중심적 연결축을 형성한다.

주요 기능	• TIFF, JPEG2000, WAV, FLAC, MKV, PDF/A 등 원본 파일의 온라인 장기 저장 • 백업, 버전 관리, PREMIS 이벤트 관리 기록 저장 • SSD에서 처리된 파일의 다음 단계 아카이브 공간 • 내부 및 외부 이용자에게 제공되는 API, 다운로드 서비스의 원본 저장소
기술 요소	• 대용량 HDD 기반 Object Storage (Ceph, MinIO, Scality 등) • RAID 6 또는 이레이저 코딩(EC 4+2/8+3 등) ※ 이레이저 코딩(Erasure Coding, EC) : 파일을 N개 블록으로 나누면서 N의 절반인 M개의 Parity 블록을 생성하여 Fault-tolerance를 보장하는 방식 • Metadata와 Data 분리 저장 구조 • 오브젝트 스토리지인 경우 S3 API 기반 접근 (AI 학습/분석에도 활용 가능)
특징	• 비용 대비 저장 효율이 높으며, 수십 PB까지 확장 가능 • 24/7 온라인 접근 보장(하루 24시간, 일주일 7일 내내(연중무휴) 끊김 없이 데이터를 읽고 쓸 수 있도록 서비스가 항상 가동되는 상태를 유지) • 디스크 교체 및 재구축에 따른 관리 비용 발생 • LTO와 달리 물리적 장기보존성은 낮으나 정기 검증(Fixity Check)에는 필수적임
저장 대상	• 디지털자원 원본 파일 : 이미지 파일(TIFF, JPEG2000), 영상 파일(MKV, FFV1, MP4 서비스본), 음원 파일(WAV, FLAC), 문서 파일(PDF/A, EPUB, XML 등) • 서비스용 변환본 : 썸네일, 웹 이미지, 스트리밍 파일(저용량 파일) • 백업 데이터 : DB 및 메타데이터 백업, 파일 백업
보존 목적	• 기관의 온라인 서비스 전체를 지탱하는 핵심 계층 • 고가용성, 대용량, 상대적 저비용 등이 조화를 이루는 계층 • 서비스본과 원본의 통합 보존 : 국립중앙도서관의 디지털자원 보존은 단순한 장기보존 목적으로 이루어지지 않고 상시적으로 활용이 이루어지는 구조이므로 서비스본과 원본을 분할 보존하는 것은 바람직하지 않음

[표 18] 디지털자원 보존 구조 2단계 Warm Tier

○ 3단계 Cold Tier(LTO Tape 계층)

LTO(Linear Tape-Open) Tape 계층은 디지털자원을 수십 년 단위로 안정적으로 보존하기 위해 설계된 최종 보존 저장소로서, 온라인 기반의 SSD(Solid State Drive)·HDD(Hard Disk Drive) 계층과 완전히 분리된 독립적인 오프라인 콜드 스토리지로 운영된다. 이 계층의 핵심 목적은 단순한 저장이 아니라 디지털자원을 기술 변화·재난·사고·보안 위협으로부터 장기적으로 보호하는 것에 있으며, 디지털 장기보존체계 전체에서 ‘마지막 방어선(last line of defense)’에 해당하는 역할을 수행한다.

LTO(Linear Tape-Open) 테이프는 모든 저장매체 중 Gb당 비용이 가장 낮아 대규모 아카이브 구축에 적합할 뿐 아니라 적정한 온·습도 조건을 준수할 경우 실질적인 수명이 30~50년에 달하는 매우 안정적인 장기보존 매체이다. 물리적으로 전원이 공급되지 않는 오프라인 매체라는 특성 덕분에, 랜섬웨어 감염, 해킹 침입, 서버 장비

장애 등 온라인 스토리지가 겪을 수 있는 다양한 디지털 위협으로부터 본질적으로 격리되어 있다.

이 계층에는 기관이 장기적으로 반드시 보존해야 하는 AIP(Archival Information Package) 형태의 영구보존본, HDD(Hard Disk Drive) 계층 전체의 정기 백업 이미지, 그리고 데이터베이스 및 보존 메타데이터의 장기보존본이 저장된다. 이는 OAIS(Open Archival Information System, ISO 14721) 모델에서 정의하는 Archival Storage 기능을 실질적으로 수행하는 영역으로 디지털자원의 완전성과 재현성을 장기적으로 보장하는 핵심 저장소 역할을 한다.

LTO(Linear Tape-Open) Tape 계층은 Tape Library(LTO-8/9/10), LTFS 또는 TAR 기반 파일 시스템, 그리고 IBM Spectrum Archive, Quantum Scalar, HPE DMF 등 전문 Tape Management Software를 통해 체계적으로 관리된다. 이러한 시스템은 테이프 로딩·언로딩 자동화, 데이터 무결성 검사, 카트리지 상태 모니터링, 교체 주기 관리 등 고급 기능을 제공하여 대규모 장기보존 환경 운영을 효율화한다. 또한 LTO 기술은 약 10~15년 주기로 세대가 진화하기 때문에, 세대 간 이전 작업(Tape Refresh 작업)을 통해 지속적으로 장기호환성을 확보해야 한다.

보존 안정성을 한층 더 강화하기 위해 LTO(Linear Tape-Open) 테이프는 최소 두 개 이상의 복제본을 제작하는 것이 일반적이며, 하나는 내부 보관(on-site), 다른 하나는 지리적으로 분리된 원격 소산센터(off-site)에 보관한다. 이러한 다중 카피 전략은 자연재해, 화재, 대규모 정전 등 예기치 않은 재난 상황에서도 디지털 자원이 안전하게 유지될 수 있도록 하는 핵심 요소이다.

LTO(Linear Tape-Open) 테이프는 순차 접근 방식으로 인해 읽기 및 검색 속도가 느리기 때문에, 온라인 서비스나 실시간 분석에는 직접 활용되지 않는다. 대신 이용자 요청이 있을 경우 사서 또는 시스템 관리자가 Tape Library에서 해당 테이프를 로드하여 데이터를 추출하는 방식으로 접근하게 된다. 오프라인 운용 방식은 서비스 효율성에서는 다소 제한점이 있으나, 바로 그 점이 오히려 LTO(Linear Tape-Open)를 장기보존·재난복구·무결성 유지 등 최종 보존 목적에 적합한 매체로 만들어주는 중요한 장점이다.

결론적으로 LTO(Linear Tape-Open) Tape 계층은 디지털자원의 영구적 보존, 무결성 유지, 재난 대비 기능을 종합적으로 제공하는 최종 보존 저장소이며, 기술 세대교체 시점에서 포맷 변환 및 매체 마이그레이션도 대부분 이 계층을 중심으로 이루어진다. SSD(Solid State Drive)·HDD(Hard Disk Drive) 계층이 제공하는 고속 처리 및 온라인 서비스 환경을 보완하고 장기적으로 안정적 보존을 실현하는 필수 인프라로서 디지털자원의 생애주기 전체에서 핵심적인 역할을 수행한다.

주요 기능	• 디지털 원본의 장기보존(30-50년) • 재해, 사고, 랜섬웨어 등으로부터 격리된 안전한 저장 • 디지털자원의 보안 및 무결성 유지 수단 • 보다 수준 높은 오프라인 콜드 스토리지 체계를 구성하기 위해서 다중 카피(At least 2 copies)를 제작해서 내부 보관 및 외부 보관(소산백업)을 수행함
기술 요소	• LTO-8/9/10 Tape Library • LTFS 또는 TAR 기반 저장 • Tape Management Software (IBM Spectrum Archive, Quantum Scalar, HPE DMF 등) • 정기적 Tape Refresh Cycle (10~15년 단위)
특징	• Gb 당 가장 낮은 비용 • 실제 수명 30~50년(보관 조건 준수 시) • 접근 속도는 느리지만 검색 메타데이터는 온라인 관리 • 재난복구(DR)를 위한 디지털자원 보존 필수 구성 요소
저장 대상	• 디지털자원의 영구보존용 원본 파일(Archival Master) • HDD 계층 전체의 정기 백업 이미지 • 메타데이터 및 DB의 주기적 백업본 • 국립중앙도서관 기준에 따라 보존 가치가 높은 원본 데이터
보존 목적	• 재난 · 사고 · 랜섬웨어에 대비한 최후의 보존 수단 • 온라인 계층과 분리된 독립적 아카이브 • 기술 세대교체 시점에 맞춰서 파일 포맷 전환 및 저장매체의 마이그레이션 작업 수행

[표 19] 디지털자원 보존 구조 3단계 Cold Tier

❑ 디지털자원 접근정기 점검

안정적인 디지털자원의 보존을 위해서는 보존 파일, 매체, 메타데이터 등에 대한 정기적인 점검체계 마련이 필수적이다. 정기적인 점검체계를 구성하기 위한 항목은 NDSA(National Digital Stewardship Alliance) 레벨 모델의 요구사항에서 제시된 항목을 기반으로 제시하고자 한다.

구분 \ 수준	Level1 최소 보존 수준 (Protect Your Data)	Level2 기본적 관리 수준 (Know Your Data)	Level3 고도화된 안정성 수준 (Monitor Your Data)	Level4 완전한 장기보존 수준 (Repair Your Data)
저장 및 지리적 중복성	· 최소 2개의 데이터 복제본 보관 · 서로 다른 물리적 장소에 저장되어야 함 (same building 가능)	· 복제본이 서로 독립적인 시스템에서 운영되어야 함 · 동일한 장애 지점(SPOF)이 없는 구조	· 지리적으로 완전히 분리된 공간에 복제본 보관(다른 건물 · 도시 · 광역 지역 등) · 온라인/오프라인 복제 조합 가능	· 최소 3개 이상의 복제본을 완전히 독립적으로 운영 · 하나는 반드시 오프라인 또는 클라우드/원격지에 저장 · 저장 기술의 이질성(heterogeneity) 확보 (예: HDD + Tape + Cloud)
고정성 및 데이터 무결성	· 체크섬 생성(MD5/SHA 등) · 파일 이동이나 복제 시 Fixity 확인	· 정기적 Fixity 검사 수행 · 자동 또는 반자동 검증 도입	· 정기적이고 자동화된 Fixity 검사 · 손상 발견 시 즉시 복구 가능한 절차 마련 · 매체 상태(Healthy 상태) 점검 수행	· Fixity 검사 자동화 + 모니터링 시스템 구축 · 오류 발생 시 자동 복구 또는 반자동 복구 · 매체 교체 · Refresh 주기적으로 수행
정보보안	· 기본적인 접근 제어(권한 관리) · 인증된 사용자만 데이터 수정 가능	· 보안 정책 수립 및 문서화 · 파일 시스템 및 접근 로그 유지	· 침입 탐지 및 보안 관리 체계 적용 · 보안 사고 대응 절차 마련	· 보안 정책을 DevOps 수준으로 통합 · 데이터 무결성 및 접근 패턴 지속적 모니터링 · 보안 인증(ISO 27001 등) 기반 운영 가능
메타데이터	· 최소한의 기술 메타데이터(파일 크기, 생성일, 포맷 등) 보유 · 기초 수준의 보존 메타데이터 확보	· 생산 맥락(context) 정보를 포함한 메타데이터 유지 · 파일 생성자, 생산 시스템 등 관리	· PREMIS 기반 보존 메타데이터 유지 · 이벤트(Event) 기록 관리(Fixity, Migration 등)	· 자동화된 보존 메타데이터 생성(PREMIS Event 자동 기록) · 메타데이터의 장기보존을 위한 별도 복제 존재 · 메타데이터 변형 시 버전 관리 수행
파일 형식	· 파일 포맷 목록화 · 어떤 포맷이 있는지 인지하고 위험 포맷 식별 가능해야 함	· 위험 포맷 정의 및 관리 · 포맷 정책 문서화	· 장기보존에 적합한 포맷 우선 보유(PDF/A, TIFF 등) · 포맷 유효성 검사(TIFF validator, JHOVE 등) 실시	· 정기적 포맷 마이그레이션 수행 · 신규 포맷 등장 시 검증 · 수용 정책 운영 · 장기보존 포맷 유지 전략 자동화 가능

[표 20] NDSA 레벨 모델의 요구사항

○ 무결성 검증(Fixity Check)

NDSA 카테고리 및 요구사항		디지털자원 점검 항목	
고정성 및 데이터 무결성	Level1. 파일 이동 시 비교를 위한 체크섬 생성	체크섬(SHA-256/512) 생성 및 초기 무결성 기준값 설정	무결성 검증
	Level2. 정기적 Fixity 검사 수행	월/분기 단위 정기 Fixity 검사, 불일치 시 원인 추적	
	Level3. 자동화된 Fixity 검사 및 손상 시 복구 절차 보유	자동화된 HDD · LTO Fixity 검사, 복원 경로 마련	
	Level4. Fixity 모니터링 체계 및 자동 · 반자동 복구	PREMIS Fixity Event 자동 기록, 자동화된 복구 프로세스	

[표 21] NDSA 카테고리 기준 무결성 점검 근거

디지털자원의 장기보존 환경에서 무결성 점검은 디지털 객체의 원본성을 보장하고 저장・전송・마이그레이션 과정에서 발생할 수 있는 비트 단위 손상(Bit Rot)을 조기에 탐지하기 위한 핵심 관리 활동으로 정의된다. 이러한 점검은 장기보존 스토리지 환경에서 데이터의 변조나 손상을 예방하는 목적을 가지며, OAIS(Open Archival Information System) 참조모델에서 보존저장영역(Archival Storage) 기능에 포함된 오류검사(Error Checking) 및 복구(Recovery) 요구사항을 충족하기 위한 필수 절차로 위치한다.

무결성 점검은 다양한 기술적 항목을 중심으로 수행된다. 우선 SHA-256 또는 SHA-512와 같은 체크섬 알고리즘을 활용하여, 디지털 객체가 원본 상태와 동일하게 유지되고 있는지를 비교 검증한다. 아울러 파일 내부에 포함된 CRC(Cyclic Redundancy Check) 값을 비교함으로써 구조적 오류 발생 여부를 탐지하며, HDD(Hard Disk Drive) 기반 오브젝트 스토리지 환경에서는 이레이저 코딩이나 RAID 구성에서 발생하는 재구축(Rebuild) 또는 복구(Healing) 상태를 점검한다. LTO(Linear Tape-Open) 테이프 환경에서는 ECC 또는 CRC(Cyclic Redundancy Check) 오류 블록의 증가 여부를 분석하여 매체 열화의 초기 신호를 파악한다. 또한 무결성 불일치가 발생한 경우에는 원본, 복제본, 백업본 간의 교차 비교를 통해 손상 원인을 진단하는 절차를 수행한다.

이러한 무결성 검증 과정은 자동화 도구를 통해 효율적으로 운영될 수 있다. Fixity(AVPreserve)는 표준 체크섬을 생성하고 검증하는 경량 도구로, 정기적인 스케줄링을 통해 대량 파일에 대한 자동 검증에 적합하다. iRODS와 Rule Engine을 결합한 방식은 데이터 이동 및 복제 과정에서 무결성 검사를 자동으로 수행할 수 있어, 온프레미스와 클라우드를 포함한 분산 객체 환경에서 높은 신뢰성을

제공한다. 또한 Rsync의 체크섬 기반 검증 기능이나 ZFS 파일 시스템의 self-healing 기능은 파일 시스템 수준에서 변경 여부를 확인하고 자동 복구를 지원한다. HDD(Hard Disk Drive) 기반 오브젝트 스토리지 환경에서는 Ceph나 MinIO의 자동 Scrub 기능을 활용하여 파일 접근성 점검과 비트 손상 여부를 검사할 수 있으며, Ceph의 Deep Scrub은 실제 파일 블록 읽기 테스트를 수행하고 MinIO의 Healing 및 무결성 체크(Integrity Check) 기능은 이레이저 코딩을 기반으로 손상 또는 접근 실패 발생 시 자동 복구를 지원한다.

❍ 포맷 유효성 검사(Format Validation)

NDSA 카테고리 및 요구사항		디지털자원 점검 항목	
File Formats	Level1. 포맷 목록화	보유 포맷 유형 조사 · 관리	포맷 유효성 검사
	Level2. 위험 포맷 식별 가능	비표준 · 충돌 가능 포맷 파악	
	Level3. JHOVE 등 포맷 유효성 검증 도구 사용	포맷 유효성 검사 및 오류 분석	
	Level4. 정기적 포맷 유효성 검사 및 포맷 마이그레이션 자동화	포맷 변환 전략 수립 · 선제적 변환 실행	

[표 22] NDSA 카테고리 기준 포맷 유효성 검사 근거

파일 포맷 유효성 점검은 디지털자원의 장기보존 가능성을 확보하기 위한 필수 관리 활동으로서 파일 구조의 손상 여부를 판단하고 국제 표준 포맷 준수 상태를 검증하는 것을 주요 목적으로 한다. 이 점검은 TIFF, PDF/A, WAV, MKV/FFV1 등과 같은 국제적으로 통용되는 보존 포맷이 해당 규격을 충실히 따르고 있는지를 확인함으로써, 장기보존 가능성이 낮은 포맷을 사전에 식별하고 개선 경로를 마련하는 데 기여한다. 또한 파일 포맷 유효성 점검 결과는 향후 포맷 마이그레이션 대상의 사전 분류와 우선순위 도출을 위한 기초 자료로 활용된다.

점검 항목은 포맷 유형별로 세분화되어 수행된다. 문서 및 이미지 파일의 경우 PDF/A-1, PDF/A-2, PDF/A-3 규격에 따른 유효성 여부와 TIFF Baseline 및 Extension 구조의 적합성을 검증한다. 영상 및 음원 파일에 대해서는 코덱 접근 가능성, 스트림 오류 발생 여부, 헤더 값의 비정상 여부를 점검하며, 파일 내부에 포함된 IPTC, XMP, EXIF 등의 메타데이터가 손상되지 않았는지도 함께 분석한다. 아울러 각 파일이 포맷별 Recommended Practice에 부합하는지를 확인함으로써 형식적 유효성뿐 아니라 보존 관점에서의 적합성을 점검한다.

이러한 파일 포맷 유효성 점검은 자동화 도구를 통해 체계적으로 수행될 수

있다. 문서 및 이미지 유형의 파일에 대해서는 FITS, JHOVE, Apache Tika를 결합하여 파일 접근성과 포맷 분석을 자동화할 수 있다. FITS는 파일 메타데이터를 추출하고 포맷 분석을 수행하는 과정에서 파싱 실패나 파일 열기 불가 시 오류를 발생시키므로, 접근성 검사 자동화에 활용될 수 있다. JHOVE는 TIFF, JPEG2000, PDF 파일에서 접근성 오류가 발생할 경우 이를 정형성 위반(Not Well-Formed) 또는 유효하지 않음(Not Valid)로 보고하며, 접근 불가능 상태를 실질적인 접근성 오류로 식별할 수 있다. Apache Tika는 파일 파싱 실패 여부를 기준으로 접근성 문제를 자동으로 식별하고 로그를 기록하는 방식으로 문서, 이미지, 오디오 등 다양한 유형에 범용적으로 적용될 수 있다.

영상 및 음원 파일의 경우에는 FFmpeg 및 FFprobe를 활용하여 오디오 스트림과 비디오 코덱의 접근 가능 여부를 자동으로 검사할 수 있으며, 이는 대형 시청각 아카이브에서 자동 품질검사(QC) 도구로 활용되고 있다. 이와 함께 ePADD Access Checker는 이메일 아카이빙 용도로 개발되었으나 파일 접근 가능성과 MIME 파싱 여부의 자동 점검에 활용될 수 있다. 방송 및 영화 분야에서는 QCTA나 Baton QC Tools와 같은 자동화 품질검사 도구를 통해 영상 및 음원 파일을 실제로 열어 손상 여부, 스트림 접근성, 재생 시간 불일치 여부 등을 자동으로 보고한다. 또한 Preservica Content Checker는 클라우드와 온프레미스 아카이브 환경을 모두 지원하며 접근성, 무결성, 메타데이터 정합성을 통합적으로 검사할 수 있는 도구로 활용된다. MediaConch는 Matroska(MKV), FFV1 등 보존 포맷의 유효성을 검증하고 PREMIS(Preservation Metadata : Implementation Strategies)와의 연계를 지원하는 도구로 제시되며, 파일 포맷 유효성 점검과 보존 메타데이터 관리의 연계를 가능하게 한다.

○ 저장매체 상태 점검(Media Health Check)

NDSA 카테고리 및 요구사항		디지털자원 점검 항목	
Storage & Geographic Redundancy	Level1. 기본 백업/복제 수준(최소 2개 복제본)	매체 상태 기본 점검(SSD/HDD/LTO)	저장매체 상태 점검
	Level2. 저장소 구조 이해 · 독립성 확보(2개 복제본이 서로 다른 시스템에 존재)	스토리지 유형별 점검 실행	
	Level3. 매체 상태 모니터링 및 오류점검 필수(지리적으로 분리된 2개 이상 복제본 유지)	HDD SMART/LTO ECC · CRC 오류 검사	
	Level4. 예측 기반 유지보수(최소 3개 복제본, 그중 1개 원격 소산)	매체 교체 주기 · 수명 기반 예측 정비 구축	

[표 23] NDSA 카테고리 기준 저장매체 상태 점검 근거

저장매체 상태 점검은 디지털자원의 장기보존 신뢰성을 확보하기 위한 핵심 관리 활동으로서 SSD(Solid State Drive), HDD(Hard Disk Drive), LTO(Linear Tape-Open) Tape와 같은 다양한 저장매체의 물리적 열화를 조기에 감지하고 장애를 예방하는 것을 주요 목적으로 한다. 특히 오브젝트 스토리지 환경에서 발생할 수 있는 silent corruption을 탐지하고 테이프 라이브러리 환경에서는 헤드 마모나 테이프 장력 저하와 같은 물리적 문제를 조기에 발견함으로써 장기보존 환경의 안정성을 유지하는 데 기여한다. 또한 동일한 디지털자원의 복제본들 간 상태 차이를 비교·분석함으로써 보존본의 신뢰성을 점검하는 기능도 함께 수행된다.

점검 항목은 저장매체 유형별 특성을 반영하여 세분화된다. SSD(Solid State Drive)의 경우 Wear Level, TBW(Total Bytes Written), Bad Block 증가 패턴을 중심으로 매체 수명과 열화 상태를 분석한다. HDD(Hard Disk Drive)에 대해서는 SMART 속성 중 온도(Temperature), 재할당 섹터(Reallocated Sector), 대기 섹터(Pending Sector) 등의 지표와 함께 UER(Unrecoverable Error Rate)를 점검하여 디스크 오류 가능성을 진단한다. LTO(Linear Tape-Open) Tape 매체의 경우 ECC 및 CRC(Cyclic Redundancy Check) 오류 블록 발생 여부, 테이프 드라이브의 헤드 마모율, 로드·언로드 사이클을 기반으로 한 재장전 테스트를 수행하여 미디어 열화 상태를 판단한다. 아울러 테이프 라이브러리 단위에서는 슬롯과 드라이브 간 교차 로딩 테스트 및 로봇 암(Arm)의 동작 상태를 점검함으로써 물리적 운용 안정성을 확인한다. 이러한 저장매체 상태 점검은 자동화 도구를 활용하여 정기적으로 수행될 수 있다. HDD(Hard Disk Drive)와 SSD(Solid State Drive)의 경우 smartctl과 같은 SMART

모니터링 도구를 이용하여 온도, 배드섹터, UER, Wear Level 등의 상태를 자동 진단할 수 있으며, cron 기반 스케줄링을 통해 주기적 점검이 가능하다. 파일 시스템 수준에서는 ZFS나 btrfs가 자체 체크섬 기반 검사를 통해 매체 오류를 자동으로 탐지하며, self-healing 기능을 포함하여 오류 복구를 지원한다.
LTO(Linear Tape-Open) Tape 환경에서는 테이프 파일시스템 구조 점검(LTFS consistency check, LTFSck)과 매체 가독성 점검(Read test)을 결합하여 파일 접근성을 점검할 수 있다. 테이프 파일시스템 구조 점검(LTFS consistency check, LTFSck)은 LTFS(Linear Tape File System) 구조에 대한 접근성을 검사하는 방식으로 테이프 라이브러리 내에서 자동 로딩 테스트를 수행하고 ECC 및 CRC(Cyclic Redundancy Check) 오류 블록을 기준으로 접근성을 판단한다. LTFS(Linear Tape File System)는 자기 테이프를 디스크나 USB와 유사하게 파일 단위로 접근·관리할 수 있도록 하는 개방형 표준 파일 시스템으로 대용량 데이터 보관 및 아카이빙 환경에서 활용된다. 또한 IBM, Quantum, HPE에서 제공하는 테이프 드라이브 로그 분석 도구(Tape Drive Log Analyzer)는 드라이브 헤드 상태, 블록 오류(ECC/CRC), 카트리지 열화를 분석하고 Tape Refresh 시기를 자동으로 판정하는 데 사용된다. 이와 함께 IBM Spectrum Archive, Quantum Scalar iLayer와 같은 테이프 라이브러리 자동화 소프트웨어를 통해 자동 로딩 테스트와 오류 블록 모니터링을 수행할 수 있다.

❍ 메타데이터 오류점검

NDSA 카테고리 및 요구사항		디지털자원 점검 항목	
Meta data	Level1. 기본 기술 메타데이터 보유	파일 속성 기반 메타데이터	메타 데이터 오류점검
	Level2. 생산 맥락(Context) 메타데이터 관리	생산 · 관리 메타데이터 정합성	
	Level3. PREMIS 기반 보존 메타데이터 유지	PREMIS Event 점검	
	Level4. 자동화된 메타데이터 기록 및 버전 관리	메타데이터 자동 수집 체계	

[표 24] NDSA 카테고리 기준 메타데이터 오류점검 근거

메타데이터 오류점검은 디지털 장기보존 환경에서 보존 메타데이터의 신뢰성과 AIP(Archival Information Package) 구조의 안정성을 확보하기 위한 핵심 관리 활동이다. 이 점검의 주요 목적은 METS(Metadata Encoding and Transmission Standard)·PREMIS(Preservation Metadata: Implementation Strategies) 기반 보존 메타데이터가 정확하게 기술·관리되고 있는지를 확인함으로써 장기적인 인출

가능성을 유지하는 데 있다. 특히 기술 메타데이터와 실제 파일 내용 간의 불일치를 조기에 탐지하고 메타데이터 누락으로 인해 발생할 수 있는 검색 및 식별 오류를 예방하는 것을 중요한 목표로 한다.

점검 항목은 보존 패키지 구조와 메타데이터 구성 요소를 중심으로 체계적으로 설정된다. 우선 METS(Metadata Encoding and Transmission Standard) 및 PREMIS (Preservation Metadata: Implementation Strategies) XML(Extensible Markup Language) 문서에 대해 스키마 유효성 검사를 수행하여 문법적·구조적 오류 여부를 확인한다. 이와 함께 식별자(Identifier), 이벤트(Event), 권한(Rights) 등 보존 관리에 필수적인 필드의 누락 여부를 점검함으로써 보존 메타데이터의 기본 완결성을 검증한다. 또한 FITS나 ExifTool 등을 통해 추출된 기술 메타데이터 값과 PREMIS(Preservation Metadata: Implementation Strategies)에 기록된 기술 정보 간의 일치 여부를 비교하여, 메타데이터와 실제 파일 상태 간의 불정합을 식별한다. AIP(Archival Information Package) 또는 BagIt 기반 패키지의 경우에는 manifest 파일을 포함한 전체 구조의 일관성을 검증하여 패키지 무결성을 점검하며, 메타데이터의 버전 간 변경 이력을 비교함으로써 기록 관리의 연속성과 추적 가능성을 확인한다.

이러한 메타데이터 오류점검은 자동화 도구를 활용하여 효율적으로 수행될 수 있다. XML(Extensible Markup Language) Schema Validator 및 Schematron Validator는 METS(Metadata Encoding and Transmission Standard), PREMIS (Preservation Metadata: Implementation Strategies), 더블린 코어(Dublin Core) 등 보존 관련 스키마에 대한 자동 검증 기능을 제공하며, 다수의 아카이브 시스템에서 기본 기능으로 내장되어 활용되고 있다. iRODS Metadata Catalog(iCAT)는 정책 기반 메타데이터 검증 기능을 통해 필수 메타데이터가 누락될 경우 자동으로 경고를 발생시킨다. BagIt Validity Check 도구는 bag-info, manifest, tag 파일의 누락 여부를 검사하고 구조적 오류를 자동으로 보고함으로써 패키지 구조 점검을 지원한다. 또한 Archivematica의 Metadata Microservice는 PREMIS(Preservation Metadata: Implementation Strategies) 이벤트를 자동으로 생성하고 오류를 감지하여 METS(Metadata Encoding and Transmission Standard) 구조를 자동 검증하는 기능을 제공한다.

○ 스토리지 단계별 점검 항목

무결성 검증, 포맷 유효성 검사, 저장매체 상태 점검, 메타데이터 오류점검 등을 스토리지 저장 단계별로 정기적으로 실시하고 가능한 경우 자동화해서 NDSA Level 3~4단계를 달성할 수 있도록 디지털자원에 대한 스토리지 전략으로 도입할 필요가 있다.

디지털자원 스토리지 점검 항목	1단계 SSD Hot Tier	2단계 HDD 오브젝트 스토리지 Warm Tier	3단계 LTO Taple Cold Tier
무결성 검증 (Fixity Check)	매일 ~ 주 1회	월 1회	반기 1회 또는 테이프 당 12~18개월
포맷 유효성 검사 (Format Validation)	필요 시	반기 1회	마이그레이션 시
저장매체 상태 점검 (Media Health Check)	주 1회	월 1회	연 1회
메타데이터 오류점검	월 1회	월 1회	연 1회

[표 25] 디지털자원 스토리지 점검 항목별 점검 주기

○ 디지털자원 오류 및 접근 정기 점검 필요성

디지털 파일은 시간이 지날수록 자연적으로 손상되기 마련이다. 저장매체의 자기 물질·전자셀은 일정한 확률로 오류를 발생시키며, 비트 단위 손상(Bit Rot), Silent Corruption, 파일 구조 손상 등은 사용자가 인지하지 못한 채 누적될 수 있다. 이러한 특성 때문에 장기보존 환경에서는 정기적인 무결성 검증(Fixity Check)이 필수적이다. 체크섬(SHA-256/512), CRC(Cyclic Redundancy Check) 값 비교, 객체 스토리지의 EC·RAID 상태 확인, LTO(Linear Tape-Open) 테이프의 ECC/CRC(Cyclic Redundancy Check) 오류 블록 검사 등을 통해 파일이 온전한지 지속적으로 검증해야 한다.

저장매체는 절대 영구적이지 않다. SSD(Solid State Drive)는 셀 마모, HDD(Hard Disk Drive)는 기계적 부품의 마모와 베드 섹터 증가, LTO(Linear Tape-Open) 테이프는 물리적 열화와 헤드 마모에 노출되어 시간이 지날수록 신뢰도가 떨어진다. 또한 Object Storage 역시 Silent Data Corruption에 취약할 수 있어 정기 Scrub(블록 단위 읽기 테스트)가 필수적이다. 이러한 물리적 특성을 고려하면, 장기보존에서는 매체 고장 이전에 미리 교체하는 매체 교체(Migration·Refresh) 전략이 수립되어야 한다. HDD(Hard Disk Drive)의 SMART 상태 기반 교체, SSD(Solid State Drive) Wear Level 기반 교환 주기, LTO(Linear Tape-Open) Refresh(10~15년 주기) 등이 여기에 해당한다.

디지털 포맷 역시 영구적이지 않다. 기술 생태계가 빠르게 변화함에 따라 특정 포맷이 더 이상 지원되지 않거나, 해당 포맷을 열 수 있는 소프트웨어·코덱이 사라질 위험이 존재한다. 장기보존 기관이 지속적으로 포맷 유효성 검사(JHOVE, FITS, MediaConch)를 수행하고 보존 포맷(PDF/A, TIFF, WAVE, MKV/FFV1 등)으로 포맷 마이그레이션을 수행해야 하는 이유가 여기에 있다.

보존 환경에서는 예기치 않은 재난이 언제든 발생할 수 있다. 화재, 지진, 침수, 시스템 장애, 랜섬웨어, 내부 실수 등 다양한 요인이 디지털 아카이브를 위협한다. 따라서 장기보존체계에서는 최소한 2~3개 이상의 복제본을 다른 저장환경에서 운영해야 하며, 그중 하나 이상은 반드시 오프사이트(원격 소산) 환경에 보관해야 한다. 이는 NDSA Level 4와 OAIS(Open Archival Information System)의 Archival Storage 요구사항이 강조하는 기본 원칙이다.

디지털 보존의 목적은 단순히 파일을 저장하는 것이 아니라 미래에도 해당 파일을 재현(Reproducibility)할 수 있도록 보장하는 데 있다. 이를 위해서는 데이터(AIP의 Content Information)뿐 아니라 해당 데이터를 이해하고 처리하는 데 필요한 기술적·관리적 메타데이터(Representation Information, PREMIS(Preservation Metadata: Implementation Strategies) Events 등)가 반드시 함께 보존되어야 한다. 파일만 존재하고 메타데이터가 소실되면 장기보존은 실패한 것이나 다름없다.

또한 모든 보존 과정은 기록되어야 한다. 누가, 언제, 어떤 점검을 했고, 어떤 오류가 있었으며, 어떻게 조치되었는지에 대한 기록은 향후 보존 신뢰성을 보장하는 필수 근거가 된다. PREMIS(Preservation Metadata: Implementation Strategies)는 이러한 보존 활동(Event), 권리(Rights), 객체(Object) 정보를 구조화하여 관리하기 위한 국제 표준이며, 장기보존 기관에서 반드시 적용해야 하는 핵심 규격이다.

마지막으로 현대 디지털 라이브러리·아카이브는 PB~EB급 데이터를 다루기 때문에 자동화 없이는 장기보존이 불가능하다. 무결성 검증(Fixity Check), 포맷 확인, 매체 상태 점검(SMART, ECC/CRC, Scrub), 메타데이터 검증, DR 테스트 등 모든 보존 절차는 자동화되어야 하며, 이상 탐지 시 경고를 발생시키고 사람의 개입을 최소화하는 체계가 요구된다. 자동 모니터링 시스템은 장기보존의 신뢰성과 지속가능성을 좌우하는 핵심 요소이다.

이를 기반으로 국립중앙도서관 디지털자원 스토리지 관리 방안을 다음과 같이 제시한다.

1. 디지털 파일은 시간이 지날수록 자연히 손상될 수 있으므로, 정기적으로 자동화된 무결성 검증이 필수적이다.
2. 저장매체는 영구적이지 않으므로 매체 교체(Migration・Refresh)가 필요하다.
3. 포맷은 기술 변화로 인해 사라질 수 있으므로 포맷 마이그레이션이 필요하다.
4. 재난은 언제든 발생할 수 있으므로 최소 2~3개 복제본을 생성하고 오프사이트 보관은 필수이다.
5. 보존은 저장이 아니라 재현성(Reproducibility) 확보이므로, AIP(Archival Information Package) 구성요소 (데이터+메타데이터)가 완전해야 한다.
6. 보존 과정은 항상 기록되어야 하므로 PREMIS(Preservation Metadata: Implementation Strategies) 기반 보존 메타데이터가 필요하다.
7. 자동화 없이는 PB급 장기보존은 불가능하므로 자동 Fixity, 자동 모니터링, 자동 DR 테스트 등이 필요하다.

❑ 검증 대상 및 절차

○ 무결성 검증(Fixity Check) 절차

무결성 검증(Fixity Check)은 디지털 장기보존 환경에서 파일의 원본성이 유지되고 있는지를 확인하기 위한 핵심 절차로서, 저장・전송・마이그레이션 과정에서 발생할 수 있는 비트 단위 손상을 탐지하고 보존 신뢰성을 확보하는 데 목적이 있다. 무결성 검증 절차는 사전 준비 단계, 검증 및 검사 단계, 결과 기록 단계로 구성된다.

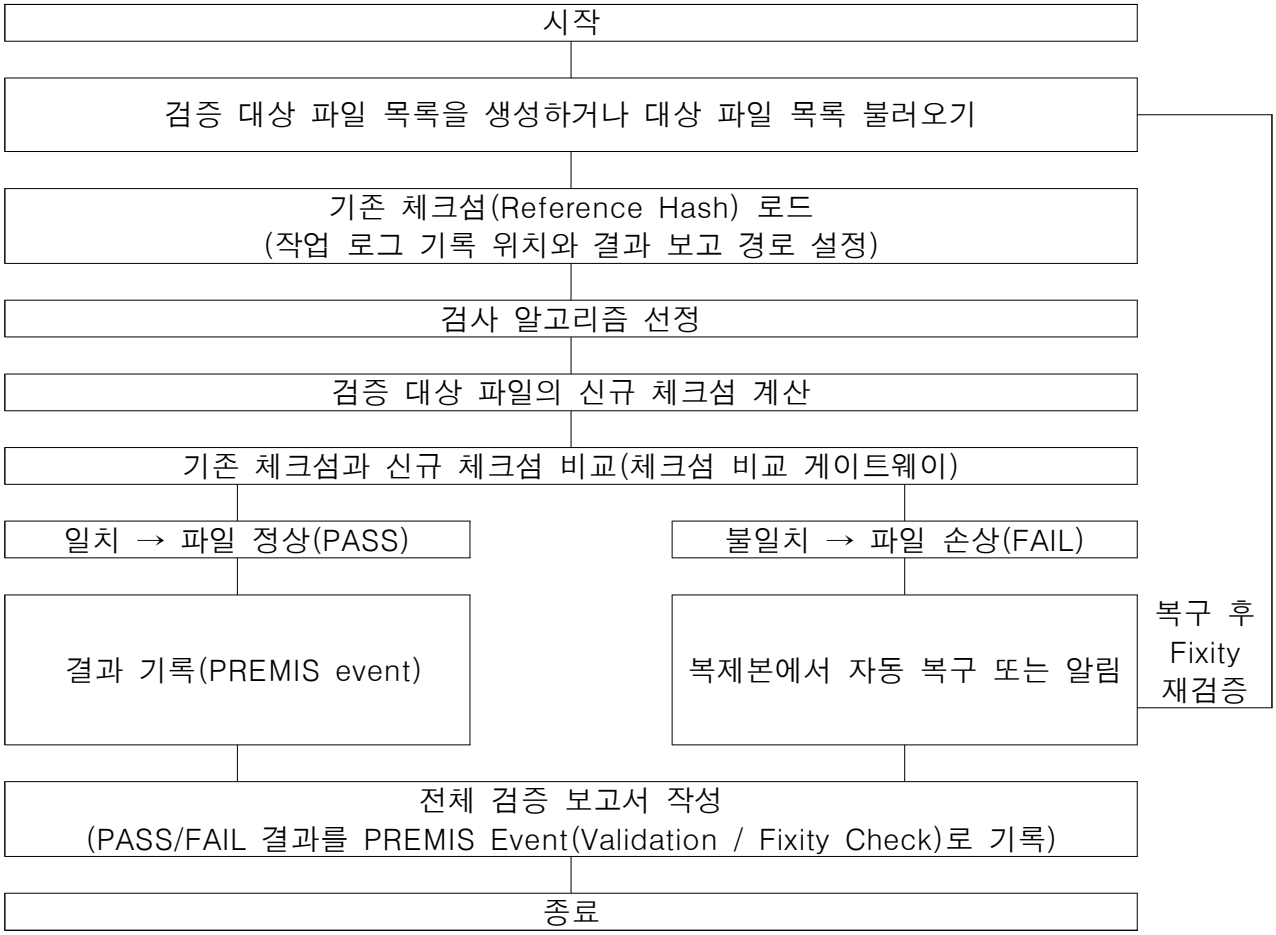
사전 준비 단계에서는 우선 검증 대상이 되는 파일 목록 또는 IP(Information Package)나 BagIt 기반의 패키지를 로드하고 기존에 생성・관리되고 있던 체크섬 값을 확보한다. 이때 활용되는 기존 체크섬 정보는 PREMIS(Preservation Metadata: Implementation Strategies) 이벤트 기록이나 Manifest-md5, Manifest-SHA256 파일 등에 포함된 값이다. 또한 검증 대상 파일의 경로와 검사 결과를 기록할 로그 경로를 설정하며, 무결성 검증에 사용할 해시 알고리즘을 선정한다. 알고리즘은 SHA-256 또는 SHA-512와 같이 충돌 가능성이 낮고 장기보존 환경에서 권장되

는 방식이 사용된다.

검증 및 검사 단계에서는 현재 시점의 파일에 대해 새롭게 체크섬을 계산하고 이를 기존에 기록된 체크섬 값과 비교한다. 두 값이 일치하지 않을 경우 해당 파일은 무결성 오류가 발생한 것으로 판정되며, 일치할 경우에는 무결성이 유지된 상태로 기록된다. 오류로 판정된 파일에 대해서는 복제본을 활용하여 자동 또는 수동 방식으로 복구를 수행하며, 복구가 완료된 이후에는 동일한 절차에 따라 재검증을 실시하여 복구 결과의 유효성을 확인한다.

무결성 검증 과정에는 다양한 도구가 활용될 수 있다. Fixity, BagIt, ACE Audit Manager, Checkm과 같은 전용 무결성 검증 도구를 비롯하여, SHA-256sum 또는 SHA-512sum과 같은 해시 계산 도구가 사용된다. 또한 검증 결과와 관련된 이력은 PREMIS(Preservation Metadata: Implementation Strategies) 기반의 무결성 이벤트 기록 도구를 통해 체계적으로 관리된다.

검사 결과는 PASS 또는 FAIL 형태로 정리된 결과표로 산출되며, 무결성 오류가 발생한 파일의 목록과 해당 파일의 경로가 함께 기록된다. 아울러 오류 발생 시 수행된 복구 여부와 후속 조치 내용이 함께 문서화된다. 이러한 검증 활동은 PREMIS(Preservation Metadata: Implementation Strategies) 이벤트로 기록되며, 이때 Event Type은 ‘무결성 검증(Fixity Check)’로, 결과는 Success 또는 Failure로 명시된다. 최종적으로 무결성 검증 결과는 일간 또는 월간 단위의 관리자 보고서 형태로 집계되어, 디지털자원의 장기보존 상태를 지속적으로 점검하고 관리하는 근거 자료로 활용된다.



[그림 15] 무결성 검증(Fixity Check) 절차

❍ 포맷 유효성 검사(File Format Validation) 절차

포맷 유효성 검사(File Format Validation)는 디지털자원이 장기보존 환경에서 안정적으로 재현되고 이용될 수 있는지를 확인하기 위한 필수 절차로서, 파일이 해당 포맷의 규격과 정책을 충족하는지를 체계적으로 검증하는 것을 목적으로 한다. 이 절차는 사전 준비 및 데이터 수집 단계, 검증 및 검사 단계, 검사 결과 기록 단계로 구성된다.

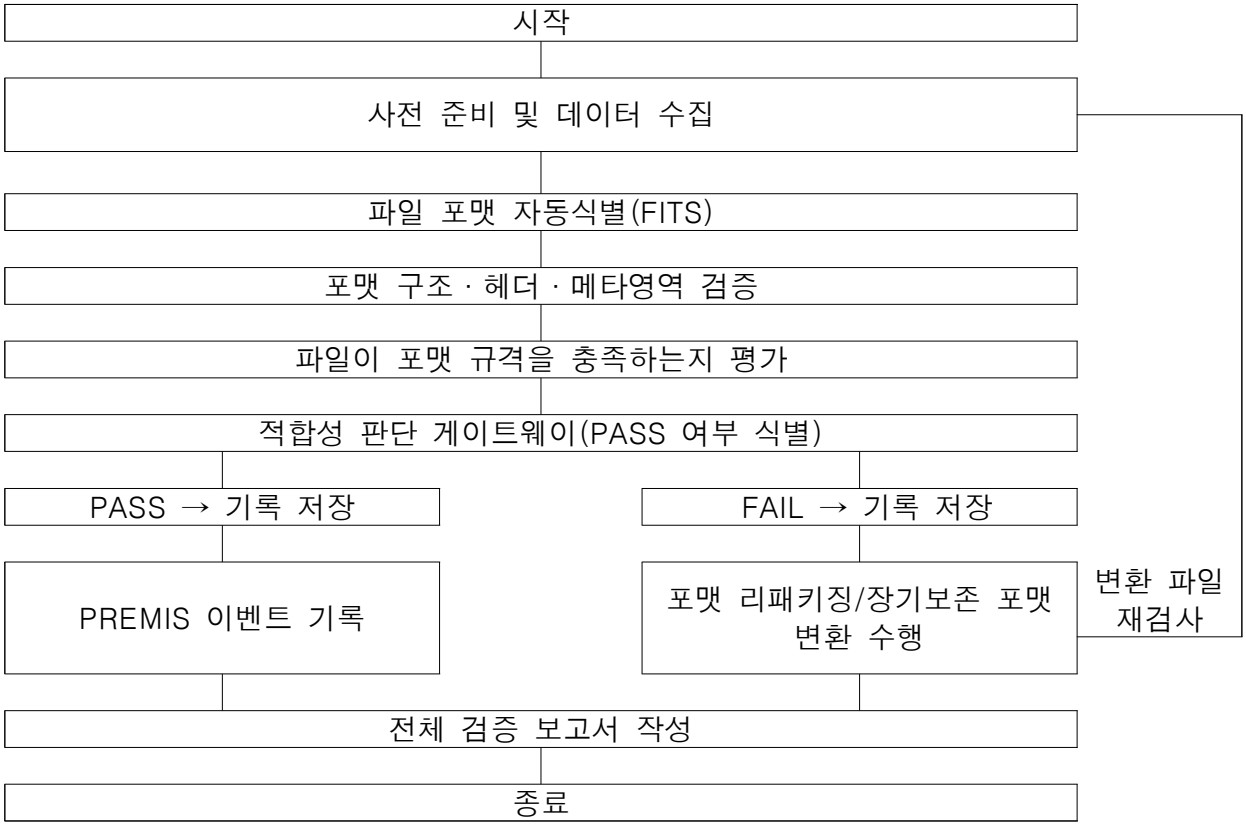
사전 준비 및 데이터 수집 단계에서는 우선 검사 대상이 되는 파일 목록을 생성하고 장기보존 대상 포맷 정책을 확인한다. 이때 PDF/A, TIFF, WAV, FFV1 등과 같이 기관에서 채택한 장기보존 포맷이 기준으로 활용된다. 또한 각 포맷에 대한 프로필 및 사양 문서를 확보하여 검사 기준을 명확히 하며, JHOVE나 FITS와 같은 포맷 검증 도구의 설정을 완료한다.

검증 및 검사 단계에서는 파일 포맷의 자동식별을 수행한 후, 파일의 구조, 헤더, 메타데이터 영역을 중심으로 포맷 규격 충족 여부를 검증한다. 이 과정에서

파일이 해당 포맷의 기술적 규격을 충족하는지 여부를 평가하며, 오류가 발생한 경우에는 포맷 리패키징을 수행하거나 장기보존에 적합한 포맷으로 변환하는 조치를 취한다. 포맷 규격을 충족한 파일은 정상 파일로 판정하여 PASS 처리하며, 변환이 이루어진 파일에 대해서는 동일한 절차에 따라 재검사를 수행하여 변환 결과의 유효성을 확인한다.

포맷 유효성 검증에는 다양한 자동화 도구가 활용된다. JHOVE는 TIFF, PDF, WAV 등의 포맷 유효성 검증에 사용되며, FITS는 종합적인 포맷 식별과 분석을 수행하는 도구로 활용된다. DROID는 파일 형식 식별을 위한 도구로 사용되고, MediaInfo는 영상 및 음원 파일의 구조 분석에 활용된다. 또한 VeraPDF는 PDF/A 규격 준수 여부를 검증하는 데 사용된다.

검사 결과는 파일 단위로 PASS 또는 FAIL 형태로 정리되며, 포맷 오류가 발생한 경우에는 헤더 오류, 태그 누락, 규격 위반 등 오류 유형이 함께 기록된다. 아울러 해당 파일에 대해 포맷 변환이 필요한지 여부와 장기보존 적합성에 대한 판정 결과가 포함된다. 이러한 검증 활동은 PREMIS(Preservation Metadata : Implementation Strategies) 이벤트로 기록되며, Event Type은 ‘Validation’으로 Outcome은 Success 또는 Failure로 명시된다. 최종적으로 포맷 유효성 검사 결과는 포맷 변환 보고서와 조치 이력으로 정리되어, 디지털자원의 장기보존 관리와 품질 통제를 위한 근거 자료로 활용된다.



[그림 16] 포맷 유효성 검사(File Format Validation) 절차

❍ 저장매체 상태 점검(Storage Media Health Check) 절차

저장매체 상태 점검은 SSD(Solid State Drive), HDD(Hard Disk Drive), LTO(Linear Tape-Open) 테이프, 오브젝트 스토리지 등 디지털자원의 장기보존에 활용되는 물리적 저장 인프라의 안정성을 확보하기 위한 필수적인 관리 절차이다. 본 점검 절차는 사전 준비 및 데이터 수집 단계, 검증 및 검사 단계, 검사 결과 정리 단계로 구성된다.

사전 준비 및 데이터 수집 단계에서는 우선 점검 대상이 되는 스토리지 유형을 정의하고 SSD(Solid State Drive), HDD(Hard Disk Drive), Tape, Object Storage로 구성된 저장장치 목록을 확정한다. 이어서 SMART 정보, Scrub 로그, 드라이브 로그 등 점검에 필요한 운영 데이터에 접근할 수 있는 권한을 확보하며, 점검 대상 클러스터, 노드, 디스크 간의 매핑 관계를 명확히 설정한다. 또한 월 단위, 분기 단위, 연 단위 등 점검 주기를 설정하여 정기적인 상태 점검이 가능하도록 준비한다.

검증 및 검사 단계에서는 저장매체 유형별 특성에 따라 점검 절차를 구분하여 수행한다. SSD(Solid State Drive)와 HDD(Hard Disk Drive)의 경우 SMART 지표를

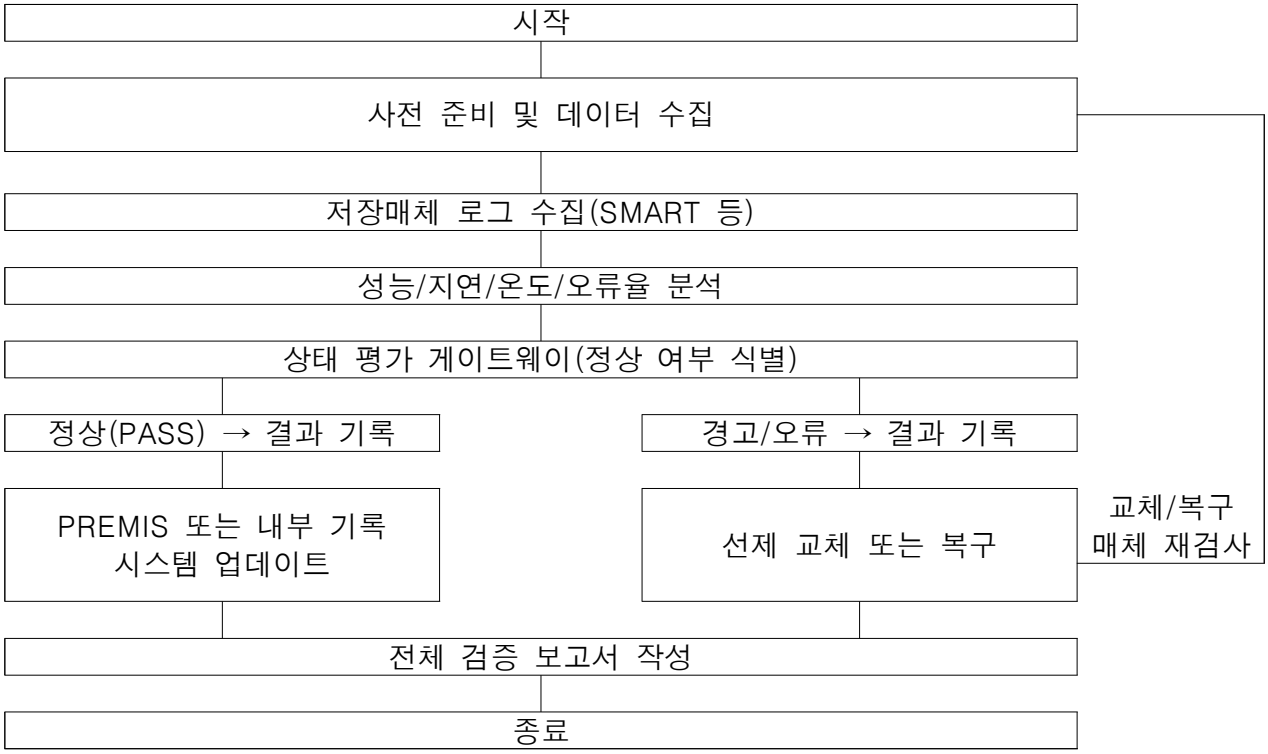
수집하고 배드 섹터 및 재할당 섹터 발생 여부를 점검한다. 아울러 입출력 응답 지연과 같은 성능 저하 징후를 확인하고 RAID 구성 환경에서는 디그레이드 상태 여부를 점검한다. 이러한 점검 결과에 따라 필요 시 디스크 교체 또는 데이터 마이그레이션을 수행한다.

LTO(Linear Tape-Open) 테이프의 경우에는 Tape Drive 로그를 수집하여 읽기 및 쓰기 오류율을 분석하고 카트리지의 물리적 상태와 오류 블록 발생 여부를 점검한다. 또한 일부 데이터를 대상으로 Read-Verify 방식의 샘플 검사를 수행하여 접근성을 확인하며, 이상이 발견될 경우에는 테이프 재기록 또는 카트리지 교체를 통해 보존 안정성을 유지한다.

오브젝트 스토리지 환경에서는 Ceph 또는 MinIO와 같은 시스템의 클러스터 상태를 점검한다. OSD 장애 여부, PG 상태 등 클러스터 헬스 상태를 확인하고 Scrubbing 및 Deep Scrubbing을 실행하여 데이터 블록의 접근성과 무결성을 점검한다. 이와 함께 이레이저 코딩 기반 복원 상태를 확인하고 노드 장애 예측을 수행하며, 이상 데이터 블록이 발견될 경우 Healing 절차를 통해 자동 복구를 수행한다.

이러한 점검 과정에는 SMARTctl, iostat, nvme-cli와 같은 저장장치 진단 도구가 활용되며, 오브젝트 스토리지 환경에서는 Ceph health, ceph pg scrub, ceph osd tree, MinIO Healthcheck 및 Prometheus Exporter 등이 사용된다. LTO(Linear Tape-Open) 환경에서는 Tape Drive Diagnostic Log와 ITDT와 같은 진단 도구가 활용되며, RAID Controller Utility 및 각 제조사에서 제공하는 모니터링 도구도 함께 사용된다.

검사 결과는 저장장치 상태 보고서 형태로 정리되며, 각 저장매체에 대해 정상, 경고, 심각 상태로 구분하여 기록된다. 또한 성능 저하의 원인 분석 결과, 정상 및 비정상 디스크 목록, 교체가 필요한 하드웨어 식별 정보가 포함된다. 이를 통해 장기보존 데이터의 안전성을 평가하고 필요한 후속 조치를 체계적으로 관리할 수 있도록 한다. 이러한 저장매체 상태 점검 활동은 PREMIS(Preservation Metadata: Implementation Strategies) 이벤트로 기록되며, 해당 이벤트의 유형은 Hardware Check로 명시된다.



[그림 17] 저장매체 상태 점검(Storage Media Health Check) 절차

❍ 메타데이터 오류점검(Metadata Validation) 절차

메타데이터 오류점검(Metadata Validation)은 디지털자원의 장기보존을 위해 생성·관리되는 메타데이터의 정확성과 일관성을 확보하기 위한 필수 절차이다. 본 점검 절차는 사전 준비 및 데이터 수집 단계, 검증 및 검사 단계, 검사 결과 정리 단계로 구성된다.

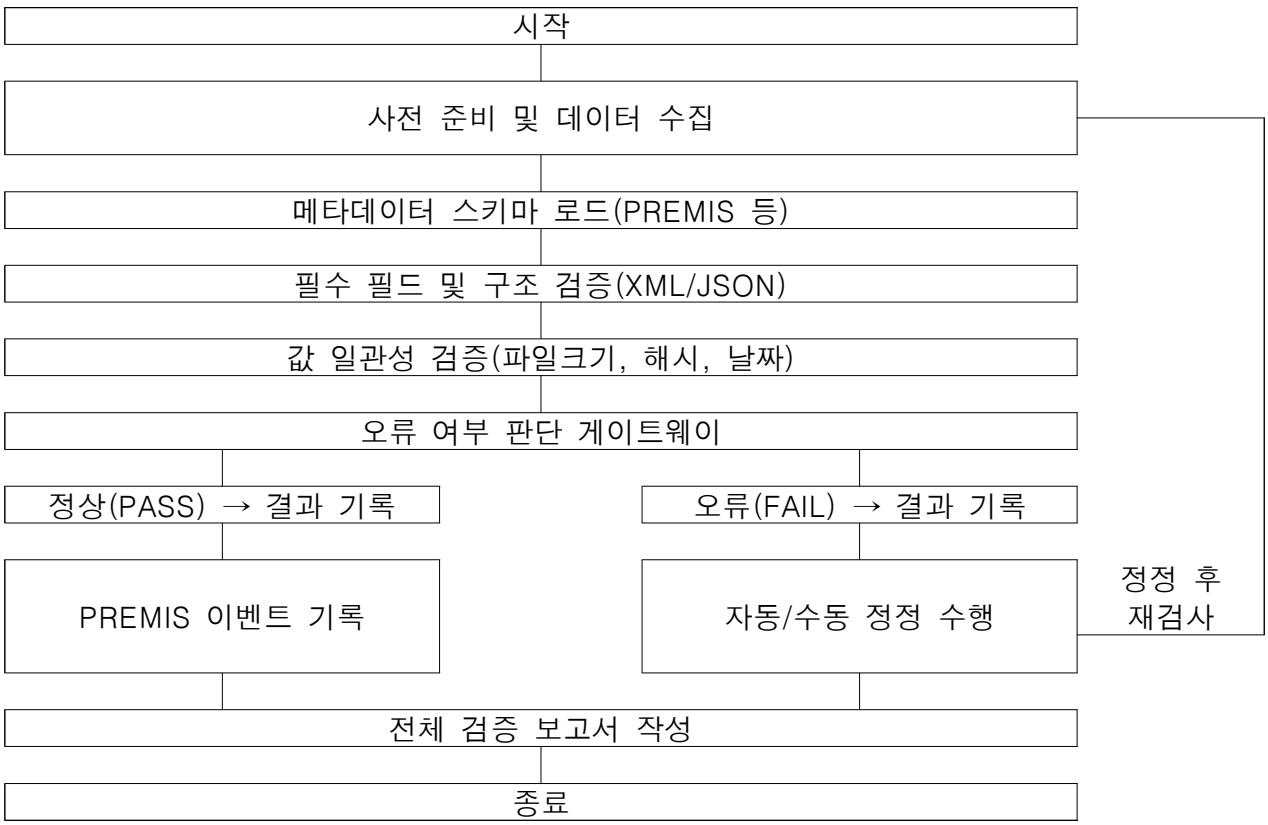
사전 준비 및 데이터 수집 단계에서는 우선 PREMIS(Preservation Metadata: Implementation Strategies), METS(Metadata Encoding and Transmission Standard), 더블린 코어(Dublin Core) 등 점검에 필요한 메타데이터 스키마를 확보한다. 이어서 검사 대상이 되는 AIP(Archival Information Package) 목록을 정의하고 해당 AIP(Archival Information Package)에 포함된 메타데이터 파일(XML 또는 JSON 형식)을 수집한다. 또한 XML(Extensible Markup Language) Schema Validator 등 메타데이터 유효성 검증을 수행하기 위한 스키마 검증 도구를 준비하여 점검 환경을 구성한다.

검증 및 검사 단계에서는 메타데이터의 필수 필드 존재 여부를 우선적으로 확인하고 파일 크기, 해시값, 날짜 정보 등 메타데이터 값의 정확성과 일관성을 검증한다. 아울러 METS(Metadata Encoding and Transmission Standard) 구조맵을 중심으로 구조 메타데이터의 정합성을 점검하며, XML(Extensible Markup

Language) 또는 JSON(JavaScript Object Notation) Schema 기반의 자동 유효성 검사를 수행한다. 점검 과정에서 오류가 탐지될 경우에는 자동 정정 또는 수동 검수를 통해 메타데이터를 수정하고 수정된 메타데이터에 대해 재검증을 수행한다.

이러한 점검 과정에는 XML(Extensible Markup Language) Schema Validator, JSON(JavaScript Object Notation) Schema Validator와 같은 스키마 검증 도구가 활용되며, 기술 메타데이터 확인을 위해 JHOVE가 사용된다. 또한 이미지 및 시청각 자료의 메타데이터 분석을 위해 ExifTool이 활용되고, PREMIS(Preservation Metadata: Implementation Strategies) Event Manager를 통해 점검 결과를 보존 메타데이터로 기록한다. 필요에 따라 Python 기반의 Custom QA Scripts를 사용하여 점검 절차를 자동화할 수 있다.

검사 결과는 필드 누락 및 오류 목록, 메타데이터 스키마 불일치 보고서, METS(Metadata Encoding and Transmission Standard) ID 매칭 오류와 같은 구조 오류 내역, 그리고 정정 조치 결과로 정리된다. 이러한 메타데이터 오류점검 활동은 PREMIS(Preservation Metadata: Implementation Strategies) 이벤트로 기록되며, 해당 이벤트의 유형은 Metadata Validation으로 명시된다.



[그림 18] 메타데이터 오류점검(Metadata Validation) 절차

3.2.1.5. 오브젝트 스토리지 기술 검토

오브젝트 스토리지는 데이터를 오브젝트 단위로 저장하며, 각 오브젝트는 데이터, 메타데이터, 고유 식별자로 구성하는 방식으로 Ceph, MinIO, Scality, Hadoop/HDFS 등이 활용되고 있다. 첫 번째 Ceph는 RADOS 기반 오브젝트 스토어 위에 RBD(블록), CephFS(파일), RGW(S3) 등을 제공하고 CRUSH 알고리즘으로 데이터를 배치하고 확장한다. 이로 인하여 가장 유연하지만 가장 복잡한 분산 스토리지이다. 다만 설계・튜닝 난이도가 높고 메타데이터/저지연 워크로드는 주의가 필요하다. 두 번째, MinIO는 단일 실행파일 기반으로 경량화된 구조로 S3 전용 오브젝트 스토리지이다. Erasure Coding 기반으로 고성능 데이터 보호를 지원하고 쿠버네티스, DevOps 등 클라우드 네이티브 환경에 최적화되어 있다. 분산 모드에서 노드/디스크가 증가할수록 처리량이 선형에 가깝게 증가하고 AI/ML・데이터레이크의 High Throughput S3에 매우 적합하다. 세 번째, 엔터프라이즈 상용 오브젝트/파일 스토리지인 Scality는 대표 제품이 RING이고, 최근에는 경량 S3 스토리지인 ARTESCA도 활용하고 있다. 완전 상용 SDS(Software-Defined Storage, 소프트웨어 정의 스토리지, 소프트웨어 계층을 사용하여 기본 물리적 스토리지 하드웨어 인프라에서 스토리지 리소스를 분리하는 데이터 스토리지 방법론(IBM))로 EC・복제・지오분산으로 고가용성을 구현하고 단일 글로벌 네임스페이스를 제공한다. 이를 통하여 엔터프라이즈급 안정성 및 관리기능 제공하고 있다. 마지막으로 Hadoop 기반 오브젝트 스토리지인 Ozone는 HDFS 파일시스템으로 구동하지만 기술적 성숙도가 낮고 네임노드에 의존하고 있으므로 네임노드 장애 시에는 접근할 수 없으므로 대규모 아카이브에는 위험 요소가 있다.

오브젝트 스토리지	무결성 점검	접근성 및 가용성	성능	운영유지보수 난이도	용량확장성 및 TCO	적합 분야
Ceph	백그라운드 스크러빙, 체크섬, OSD 자체 healing	강한 일관성, 수백 OSD 장애 허용	고성능 가능하지만 튜닝 필수, 복잡함	모니터/OSD /MDS 등 데몬 다수로 인하여 가장 높음	• 무제한 확장이 가능함 • HW 비용이 낮지만, 운영 인건비는 높음	• 오브젝트 · 블록 · 파일을 단일 SDS로 통합할 필요 • 통합 인프라/클라우드 구축 목적 • 숙련된 DevOps/인프라 팀이 있는 환경
MinIO	Bit-rot 검출, EC self-healing, WORM	강한 일관성, 장애 시 자동 복구	업계 최고 수준의 read/write throughput, NVMe · 100GbE 환경에서 최대 성능	단일 바이너리, 구성 단순으로 인하여 가장 낮음	• 선형(scale-out) 확장이 가능함 • x86/JBOD(Just a Bunch Of Disks) 어느 HW나 사용이 가능하여 TCO 최저이므로 가장 경제적인 PB~EB 규모 운영이 가능함	• AI/ML · OCR · STT · 영상처리 등 S3 기반 고성능 처리 필요 • 운영 인력 적고, 안정적 S3 저장소가 필요 • 장기보존, 무결성 검증, WORM 필수 • PB에서 EB로 확장할 계획 • TCO 최적화가 중요
Scality	엔터프라이즈 self-healing, 지오분산	액티브-액티브 다중 사이트	대규모 병렬 성능 우수, 일정한 안정성	벤더가 관리하고 UI 기반을 제공하여 낮은 편임	• 노드 추가 시 총 용량 증가하는 방식으로 수평 확장 가능하지만 완전한 선형 확장은 아님 • 상용 제품으로 고가의 상용 라이선스 비용이 발생하나 안정성과 관리편의는 가장 좋음	• SLA, 엔터프라이즈 지원, 안정성 최우선 • 랜섬웨어 보호, 금융/통신 수준 보안 요구 • 예산 충분하고 대형 벤더 지원 필요
Hadoop	블록 체크섬, 복제 기반 복구	3중 복제로 데이터 안정성이 높으나 노임노드 장애 시 위험	대용량 순차 처리 강점, 랜덤 I/O 매우 약함	네임노드 관리 위험도로 인하여 높은 편임	• 구축 및 운영비용이 매우 낮으나, 3중 복제 방식 때문에 용량 효율 떨어짐	• 분석 중심(빅데이터 처리) 워크로드 • 아카이브보다는 분석 플랫폼 핵심일 때

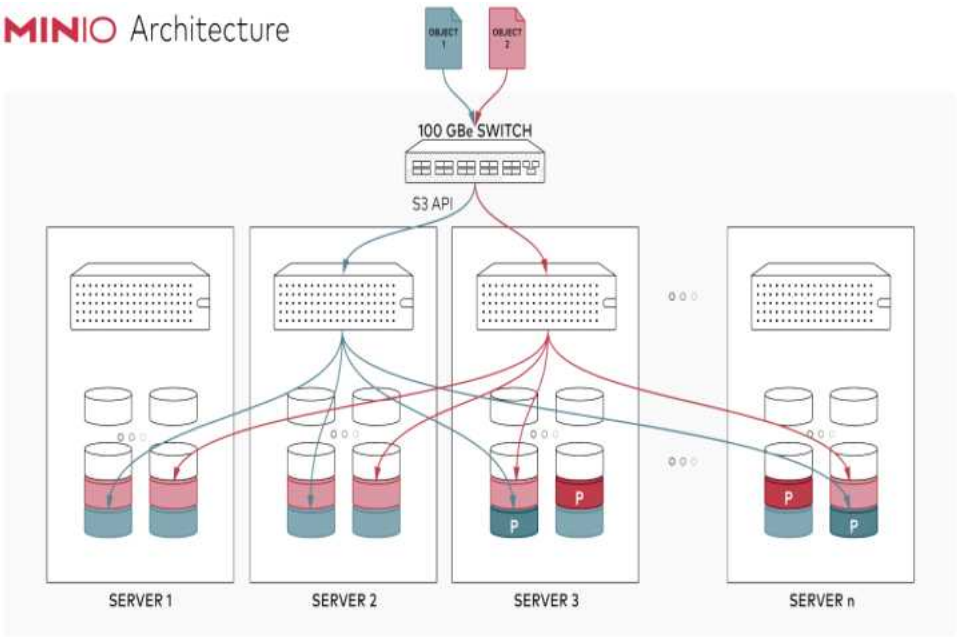
[표 26] 오브젝트 스토리지 기술요소 비교

오픈소스를 기반으로 제공되어 구축비용이 낮고, 하드웨어 구조 및 운영을 위한 기술적 수준이 낮아 운영 유지보수를 위한 전문 인력이 필요하지 않아 운용이 낮아 TCO 관점에서 최적의 대안으로 디지털자원 보존관점의 오브텍트 스토리지는 MinIO를 적용하는 것이 바람직할 것으로 판단된다. 만약 통합 스토리지 플랫폼이 목표라면 Ceph, 엔터프라이즈급 SLA가 필요하면 Scality, 빅데이터 분석 중심이라면 Hadoop을 적용하는 것이 바람직할 것이다.

오브젝트 스토리지	구축비용	운영비용	확장성	운영난이도	안정성	합계
Ceph	3 낮음	2 보통	3 높음	1 높음	2 보통	11
MinIO	3 낮음	3 낮음	3 높음	3 낮음	2 보통	14
Scality	1 높음	1 높음	2 보통	1 낮음	3 높음	8
Hadoop	3 낮음	3 낮음	2 보통	2 보통	1 낮음	11

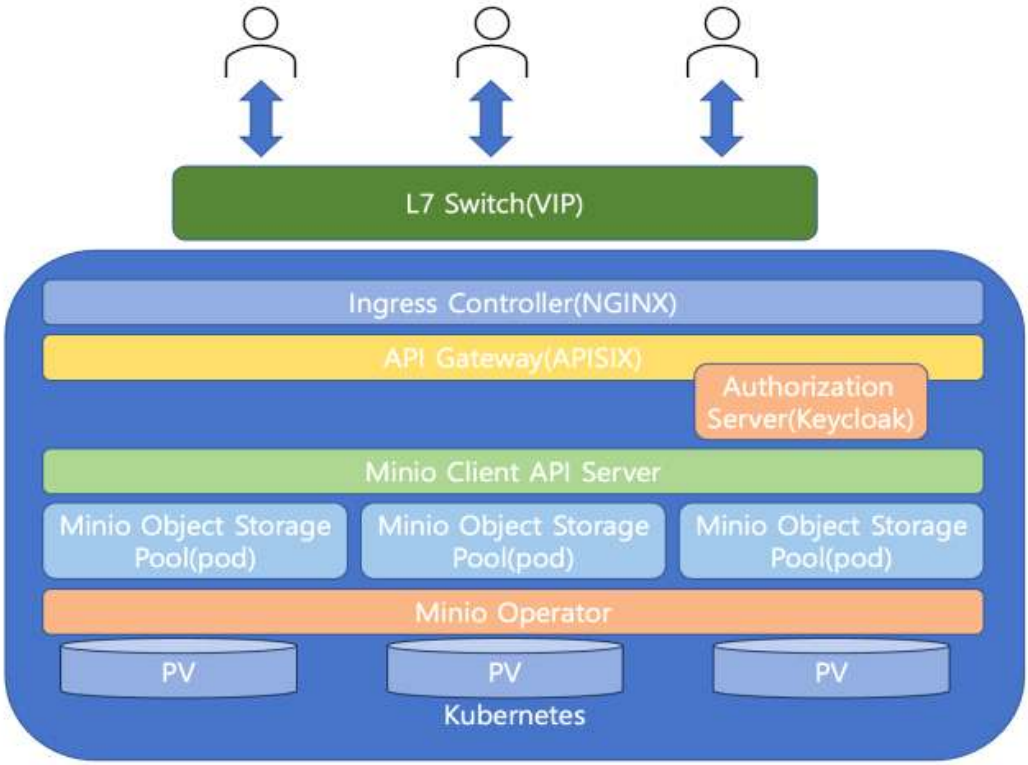
[표 27] 오브젝트 스토리지 기술요소 선정

이러한 검토 결과를 반영하여 간편한 설치 및 운영, 확장 및 고가용성 등의 측면에서 MinIO 오픈소스 솔루션이 가장 적합한 것으로 판단할 수 있다.



[그림 19] MinIO 아키텍처

이를 위해 오픈소스 오브젝트 스토리지인 MinIO와 컨테이너 오케스트레이션 플랫폼인 Kubernetes를 결합하여, 하나의 통합 오브젝트 스토리지 클러스터를 구축하고 그 위에 보존용 마스터, 서비스용 파생본, 업무용 중간 산출물 등을 버킷·정책·권한·수명주기 관리로 논리적으로 분리하는 구조를 설계할 수 있다. Kubernetes는 MinIO 클러스터의 배포·확장·장애 복구를 자동화하여 인프라 운영 효율을 높이고, MinIO는 S3 호환 API(Application Programming Interface), Erasure Coding, 버전 관리, 복제 기능을 통해 장기보존에 필요한 내구성과 확장성을 제공한다.



[그림 20] MinIO + Kubernetes 기반 통합 스토리지 플랫폼 구조

3.2.1.6. 스토리지 마이그레이션 방안

스토리지 마이그레이션 방안의 수립을 위해서는 우선 데이터의 중요도에 따른 분류가 선행되어야 한다. 데이터는 보존 필요성과 손상 발생 시 영향 수준에 따라 중요 데이터와 일반 데이터로 구분할 수 있다.

중요 데이터는 영구 또는 준영구적인 보존이 요구되며, 데이터 손상이 발생할 경우 기관 운영이나 서비스 제공에 중대한 영향을 미칠 수 있는 데이터를 의미한다. 이러한 데이터에 대해서는 복구 가능성을 전제로 한 백업 체계의 구축과 교차 검증, 이중화가 필수적으로 요구된다. 중요 데이터의 범주에는 보존용 원본 파일, 마스터 파일, 데이터베이스의 핵심 테이블, 규제 기록물, 공공 기록물, 장기보존이 필요한 디지털자원 등이 포함된다.

반면 일반 데이터는 일부 손실이 허용되거나, 손실이 발생하더라도 재생성 또는 재수집이 가능한 데이터를 의미한다. 이 범주의 데이터는 보존 안정성보다는 접근성과 비용 효율성이 주요 고려 요소가 된다. 일반 데이터의 예로는 시스템 로그, 임시 파일, 분석 및 처리 과정에서 생성되는 중간 산출물, 그리고 재수집이 가능한 데이터 등이 해당된다.

❑ SSD(Solid State Drive) 스토리지 마이그레이션

SSD(Solid State Drive) 스토리지는 플래시 메모리의 쓰기 내구성 한계로 인해 일반적으로 약 5~10년의 유효 수명을 가지므로, 중요 데이터는 전체 수명의 약 50%인 3~5년을 선제적 마이그레이션 주기로 설정하는 것이 바람직하다. 반면, 일반 데이터는 SSD(Solid State Drive) 수명의 약 60% 수준을 적용하여 4~6년 주기로 스토리지 마이그레이션을 적용한다.

구분	중요 데이터	일반 데이터
마이그레이션 전략	**3~5년 주기**로 신규 SSD 또는 상위 계층 스토리지로 데이터 이동	**4~6년 주기**로 선제 이전
점검 주기	**월 1회** SMART, Wear Level, ECC Error Rate 점검	**분기 1회** SMART 점검
교체 주기	3~5년 또는 Wear Leveling 80~90% 도달 시 선제 교체	5년 또는 성능저하 감지 시
백업 및 복제 전략	이중화 필수(미러링 + 이중노드), Object Storage나 Tape에 별도 백업	RAID 1/5/1+0 등 일반적 수준
추가 전략	DB/트랜잭션 계층은 1~2년 마다 성능 · 지연 · 내구성 평가	고속 캐시 · VM 스토어 등에 사용 후 필요시 HDD/Tape로 덤프

[표 28] SSD 스토리지 마이그레이션 주기 설정

❑ HDD(Hard Disk Drive) 스토리지 마이그레이션

HDD(Hard Disk Drive) 스토리지는 기계식 부품의 마모와 회전계 부하로 인해 약 5~7년 정도의 안정적 수명을 갖는 것으로 알려져 있다. 이에 따라 중요 데이터는 HDD(Hard Disk Drive) 예상 수명의 약 70%에 해당하는 4~5년을 선제적 마이그레이션 주기로 설정하고 일반 데이터는 전체 수명 범위인 5~7년을 기준으로 스토리지 마이그레이션 주기를 적용한다.

구분	중요 데이터	일반 데이터
마이그레이션 전략	4~5년 주기로 신규 HDD 또는 오브젝트 스토리지/테이프로 마이그레이션	5~7년 주기 또는 용량 증설 시 이전
점검 주기	월 1회 SMART 점검 및 연 1회 전체 배드섹터 스캔	분기 1회 SMART 점검
교체 주기	4~5년 또는 SMART 경고(Seek Error, Reallocated Sector 증가 등) 발생 시	6~7년 주기
백업 및 복제 전략	RAID 6, 오브젝트 스토리지, Tape의 3중 백업 권장	RAID 5 또는 6 정도면 충분
추가 전략	고가용 SAN 환경은 듀얼 패스, 핫스페어 구성 필수	로그 · 중간 데이터를 일정 기간 후 Tape나 삭제로 관리

[표 29] HDD 스토리지 마이그레이션 주기 설정

❑ 테이프 드라이브 스토리지 마이그레이션

테이프 스토리지는 매체의 물리적 안정성과 규격상 장기보존 특성으로 인해 약 30년의 보존 수명을 가지나, 실제 운영에서는 드라이브 호환성·규격 변화(LTO 세대 교체 주기 5~7년) 등을 고려할 필요가 있다. 이 특성을 반영해 중요 데이터는 전체 수명의 약 30% 수준인 5~7년을 선제적 마이그레이션 주기로 설정하고 일반 데이터는 약 50% 수준인 10~12년 주기로 스토리지 마이그레이션을 적용한다.

※ LTO 세대교체 주기 : 일반적으로 LTO 세대 간 공식적인 발표 간격은 약 2~3년 정도이지만, 읽기 호환성, 드라이브 단종 주기, 장비 유지보수 기간 등 때문에 실제 기관·기업의 운영 환경에서 적용하는 교체 주기는 5~7년이 가장 일반적이다.

구분	중요 데이터	일반 데이터
마이그레이션 전략	LTO 세대 2세대 증가 시(약 8~10년) 또는 5~7년 주기로 규칙적 Rewriting	10~12년 주기(장비 지원 기간 내)
점검 주기	연 1회 무결성 검증(Read-Verify), 카트리지 상태 점검	연 1회 무결성 점검 또는 표본 검사
교체 주기	LTO-9 기준 15~20년, 하지만 실제 운영에서는 7~10년 권장	10~15년 주기
백업 및 복제 전략	최소 2카피(onsite + offsite) 또는 3카피(DR 포함)	1~2카피면 충분
추가 전략	LTFS 메타데이터 주기적 업데이트, 온습도 안정 관리(서고 기준)	비핵심 백업 또는 로그 · 분석 원본의 장기보관에 적합

[표 30] 테이프 드라이브 스토리지 마이그레이션 주기 설정

3.2.2. 백업 인프라 구성 방안

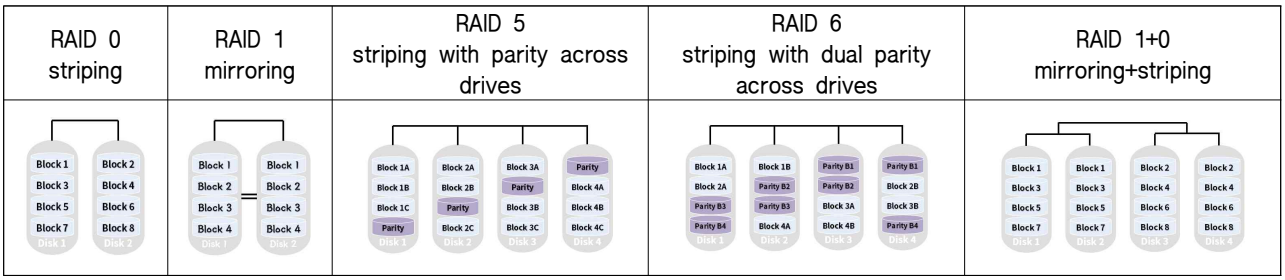
3.2.2.1. 백업 인프라 구성 요소

일반적으로 컴퓨터에 저장된 데이터를 백업해서 필요시에 복구할 수 있는 인프라는 컴퓨터 노드 내부 데이터 보호를 위한 RAID 구성, 디스크 기반 백업 (D2D: Disk-to-Disk), VTL(Virtual Tape Library) 기반 백업, LTO(Linear Tape-Open) 기반 백업 (Tape Backup), 소산백업(오프사이트 백업), 원격지 백업(재해복구 DR 백업) 등을 주로 활용하고 있다.

디지털자원 장기보존을 수행하는 기관 환경에서는 디스크 기반 백업을 통한 신속한 운영 복구, 테이프 기반 백업을 통한 장기보존, 소산백업을 통한 물리적 재난 대비, 그리고 원격지 백업을 통한 업무 연속성 확보를 상호 보완적으로 결합한 다계층 백업 전략이 필요하다. 이러한 다층적 백업 체계는 단일 장애나 재해가 디지털자원의 영구적 손실로 이어지는 것을 방지하는 핵심적인 안전장치로 기능한다.

❑ 컴퓨터 노드 내부 데이터 보호를 위한 RAID 구성

컴퓨터 노드 내부에 장착된 디스크는 물리적 고장, 노후화, 전원 문제 등으로 인해 장애가 발생할 수 있으며, 이 경우 저장된 데이터가 손실될 위험이 있다. 이러한 위험을 최소화하기 위해 여러 개의 디스크를 논리적으로 묶어 하나의 저장장치처럼 구성하는 RAID(Redundant Array of Independent Disks) 기술을 적용한다. RAID는 디스크를 스트리핑, 미러링, 패리티 등 다양한 방식으로 조합하여 성능 향상 또는 데이터 보호를 구현하며, 구성 방식에 따라 RAID 0, RAID 1, RAID 5, RAID 6, RAID 1+0 등의 유형으로 구분된다.



[그림 21] 노드 내부 데이터 보호 RAID 종류

RAID 유형	구분	내용
RAID 0(Stripping)	구성 방식	데이터를 2개 이상의 디스크에 순차적으로 분산 저장
	최소 디스크 수	2
	장점	모든 디스크를 동시에 사용하여 성능이 가장 높음
	단점	디스크 1개라도 고장 시 전체 데이터 손실 → 데이터 보호 기능 없음
	특징 및 적용 환경	고성능이 중요하나 데이터 보호가 불필요한 임시 처리 영역
RAID 1 (Mirroring)	구성 방식	동일한 데이터를 두 개의 디스크에 복제 저장
	최소 디스크 수	2
	장점	디스크 1개 고장 시에도 즉시 데이터 유지 가능
	단점	전체 디스크 용량의 50%만 사용 가능 → 스토리지 효율 낮음
	특징 및 적용 환경	운영체제, 중요 설정 파일 등 안정성 중시 환경
RAID 5 (Stripping + Single Parity)	구성 방식	데이터와 1개의 패리티 정보를 디스크 전체에 분산 저장
	최소 디스크 수	3
	장점	디스크 1개 고장 시 패리티로 데이터 복구 가능
	단점	디스크 2개 이상 동시 고장 시 복구 불가
	특징 및 적용 환경	성능 · 용량 · 안정성의 균형이 필요한 일반적 서버 환경
RAID 6 (Stripping + Dual Parity)	구성 방식	RAID 5 구조에서 패리티를 2개로 확장하여 분산 저장
	최소 디스크 수	4
	장점	디스크 2개까지 동시 고장 허용 → 보호 수준 높음
	단점	RAID 5 대비 쓰기 성능 저하
	특징 및 적용 환경	고가용성 · 고신뢰성이 요구되는 핵심 시스템
RAID 0+1 / 1+0 (Stripping + Mirroring)	구성 방식	스트리핑과 미러링 결합 구조
	최소 디스크 수	4
	장점	높은 성능과 높은 안정성 동시 확보
	단점	전체 디스크 용량의 50%만 사용 가능
	특징 및 적용 환경	데이터베이스, 고성능 · 고가용성 엔터프라이즈 환경

[표 31] RAID 유형

❑ 디스크 기반 백업 (D2D: Disk-to-Disk)

디스크 기반 백업(Disk-to-Disk, D2D)은 스토리지 장애나 사용자 오류에 대비한 단기 복구, 즉 운영 백업(Operational Backup)을 목적으로 활용된다. 이 방식은 백업된 데이터에 대한 복구가 필요한 경우 수분 또는 수 시간 이내에 신속한 1차 복구가 가능하다는 점에서 파일 손상, 사용자 실수, 운영 시스템 장애 대응에 효과적이다. 일반적으로 NAS 또는 D2D 전용 어플라이언스를 활용하여 별도의 백업용 스토리지를 구축하고 Veritas NetBackup, Veeam, CommVault, Bacula 등과 같은 백업 솔루션을 통해 스케줄 기반의 전체·증분 백업을 수행한다. 디스크 기반 백업은 복구 속도가 매우 빠르고 운영 중 오류 발생 시 즉각적인 대응이 가능하며, 관리와 테스트 복구가 용이하다는 장점을 가진다. 그러나 디스크 매체 특성상 장기 보관 시 비용 부담이 크고, 재해 대응을 위해서는 반드시 오프사이트 백업을 추가로 구성해야 한다는 한계를 지닌다.

❑ VTL(Virtual Tape Library) 기반 백업

VTL(Virtual Tape Library) 기반 백업은 디스크 기반 백업과 테이프 기반 백업을 연결하는 중간 계층으로서의 역할을 수행한다. VTL은 물리적 LTO(Linear Tape-Open) 테이프 환경을 디스크 기반으로 가상화하여, 백업 소프트웨어가 이를 실제 테이프 라이브러리처럼 인식하도록 제공하는 기술이다. 디스크 기반 저장 영역 위에 가상의 테이프 드라이브를 구성함으로써, 기존 테이프 기반 백업 정책과 운영 방식을 유지하면서도 백업 및 복구 속도를 향상시킬 수 있다. 또한 필요 시 실제 LTO(Linear Tape-Open) 테이프 라이브러리로 테이프를 생성하고 이관하는 자동화가 가능하다. 이러한 특성으로 인해 VTL은 테이프 환경을 유지해야 하는 기관에서 운영 효율성을 높이는 데 기여한다. 다만 VTL 자체는 디스크 기반이므로 장기보존 매체로는 적합하지 않으며, LTO(Linear Tape-Open) 테이프 대비 비용이 높고 대규모 장기 데이터 축적 시 확장비용이 증가할 수 있다는 제약이 있다.

❑ LTO(Linear Tape-Open) 테이프 기반 백업(Tape Backup)

LTO(Linear Tape-Open) 테이프 기반 백업(Tape Backup)은 가장 긴 수명과 가장 낮은 단위 비용을 제공하는 장기보존 매체로 평가된다. LTO(Linear Tape-Open) 테이프는

도서관, 기록관, 방송사, 연구기관 등에서 디지털자원의 수십 년 단위 장기보존을 위해 사실상 표준처럼 활용되고 있다. 일반적으로 LTO-8((Linear Tape-Open-8), LTO-9(Linear Tape-Open)-9 등의 테이프와 로봇 암, 슬롯, 테이프 드라이브로 구성된 테이프 라이브러리를 통해 자동화된 백업이 이루어지며, WORM(Write Once Read Many) 테이프를 적용할 경우 원본 데이터의 변조 방지 또한 가능하다. LTO(Linear Tape-Open) 테이프는 30년 이상 장기 보관이 가능하고 디스크 대비 단위 저장 비용이 10~20배 이상 저렴하며, 오프라인 보관이 가능해 랜섬웨어나 보안 위협으로부터 자유롭다는 장점을 가진다. 또한 비회전형 미디어 특성상 에너지 비용이 거의 발생하지 않아 장기보존 정책에 매우 적합하다. 반면 실시간 서비스나 고빈도 접근에는 부적합하고 검색 및 액세스 속도가 느리며, 로봇 장비 운영과 테이프 슬롯 관리, 물리 매체 교체 및 이관에 따른 운영 부담이 존재한다.

❑ 소산백업(오프사이트 백업)

소산백업(오프사이트 백업)은 테이프 백업본 중 일부를 외부 물리적 보관소(타 지역 분관 또는 민간 보관소 등)로 이관하여 보관하는 방식으로 지진·화재 등으로 전산센터 전체가 물리적으로 파손되는 상황에 대비하기 위한 최종 보호 계층이다. 이 방식은 데이터 보호의 마지막 방어선으로서 내부 네트워크나 서버 장애의 영향을 전혀 받지 않으며, 랜섬웨어나 악성코드가 접근할 수 없는 완전한 에어갭(air-gap) 환경을 제공한다. 오프라인 보관 특성상 장기 보관비용이 매우 낮고, 테이프 데이터가 물리적으로 분리되어 있어 복구의 최후 보루 역할을 수행한다. 다만 테이프 이관·운송·보관을 위한 인력과 운영비용이 발생하며, 긴급 복구 시 테이프 반입이 필요해 복구 시간이 길어질 수 있고, 외부 보관 절차에 대한 관리 부담과 자동화 수준의 한계가 존재한다.

❑ 원격지 백업(재해복구, DR 백업)

원격지 백업(재해복구, DR 백업)은 테이프 소산과 달리, 물리적으로 분리된 지역의 DR센터 또는 타 지역 데이터센터에 2차 스토리지를 구축하여 데이터를 온라인·자동화 방식으로 복제하는 구조이다. 이 방식은 시스템 장애나 지역 단위 재해 발생 시 업무 연속성(BCP)과 신속한 서비스 복구를 목적으로 하며, 실시간 또는 주기적 복제를 통해 RPO(Recovery Point Objective)와 RTO(Recovery Time Objective) 요구사항에 맞는 복구 체계를 설계할 수 있다. DR센터에는 오브젝트

스토리지, VTL, 테이프 라이브러리 등 다양한 스토리지를 조합하여 구축할 수 있으며, S3 Replication, 스토리지 미러링, 백업 솔루션의 복제 기능을 활용한 자동화된 데이터 동기화가 가능하다. 원격지 백업은 지역 단위 장애 발생 시 빠른 서비스 재개와 완전한 이중화를 가능하게 하지만, DR센터 구축과 네트워크 회선 확보에 따른 비용 부담이 크고, 실시간 복제 시 대역폭 증가로 추가 비용이 발생할 수 있다. 또한 원본 데이터에 오류가 발생할 경우 해당 오류가 그대로 복제될 위험이 있으므로, 별도의 무결성 검증 체계가 병행되어야 한다.

3.2.2.2. 디지털자원 백업 인프라 구조 적용 방안

국립중앙도서관에서 관리하는 디지털자원은 문서, 이미지, 음원, 영상, 웹 아카이브 등 비정형 데이터가 중심을 이루며, 국가 지식자산으로서 수십 년 이상 장기적으로 보존되어야 하는 특성을 지닌다. 이러한 디지털자원의 특성상 단일 백업 수단이나 단일 저장 기술만으로는 데이터 손실, 시스템 장애, 사이버 위협, 물리적 재난 등 다양한 위험에 충분히 대응하기 어렵다. 따라서 국립중앙도서관의 디지털자원 보존을 위해서는 디스크 장애부터 데이터센터의 물리적 파괴에 이르기까지 모든 위험 시나리오를 포괄할 수 있는 다계층 보호 구조, 즉 6단계 통합 백업 아키텍처의 구축이 필수적이다.

이 다계층 백업 구조는 1단계 RAID 기반 스토리지 보호, 2단계 디스크 기반 백업(D2D), 3단계 VTL 기반 백업, 4단계 LTO(Linear Tape-Open) 테이프 기반 장기보존, 5단계 소산백업, 6단계 원격지 DR 복제로 구성되며, 각 단계는 서로 다른 장애 유형과 복구 요구를 담당함으로써 전체적으로 데이터 무결성, 장기보존 안정성, 재해복구 능력, 업무 연속성을 동시에 확보하도록 설계된다.

우선 1단계 RAID 기반 기본 스토리지 보호는 디지털자원이 저장되는 서버 또는 스토리지 노드 내부에서 발생할 수 있는 가장 기본적인 장애에 대응하는 단계이다. 이는 개별 디스크의 물리적 고장, 노후화, 전원 이상 등 일상적인 하드웨어 장애로부터 데이터를 보호하는 것을 목적으로 한다. RAID 0, RAID 1, RAID 5, RAID 6, RAID 1+0과 같은 스트리핑・미러링・패리티 기반 기술을 적용함으로써 단일 디스크 고장으로 인한 데이터 손실을 방지하고 서비스 중단 없이 디스크 교체가 가능한 환경을 제공한다. 보존 계층에서는 높은 장애 허용성을 가진 RAID 6 구성이 권장되며, 서비스 계층에서는 성능과 안정성을 동시에 고려한 RAID 10 구성이 적합하다. 또한 디스크 장애 발생 시 자동 재빌드가 가능한 Hot Spare 정책, 정기적인 S.M.A.R.T 진단, 생산계 디스크의 주기적 교체 정책을 통해 기본적인 스토리지

신뢰성을 지속적으로 유지해야 한다.

2단계 디스크 기반 백업(Disk-to-Disk, D2D)은 운영 환경에서 발생하는 사용자 오류, 파일 손상, 서비스 장애에 신속하게 대응하기 위한 단기 복구 계층이다. 이 단계의 핵심 목적은 수분에서 수 시간 이내의 빠른 복구를 가능하게 하는 것으로 NAS 기반 백업 스토리지나 D2D 전용 어플라이언스를 활용하여 백업 솔루션 기반의 전체·증분 백업을 수행한다. 디스크 기반 백업은 가장 빠른 복구 성능을 제공하며, 테스트 복구와 운영 환경에서의 민첩한 대응이 가능하다는 장점이 있다. 일반적으로 증분 백업은 일 단위로 수행하고 전체 백업은 주 1회 또는 월 1회 주기로 수행하며, 보관 기간은 30~90일 수준의 단기 보관 정책이 적절하다. 다만 이 계층은 장기보존보다는 운영 복구를 목적으로 하므로, 상위 계층과의 연계를 전제로 설계되어야 한다.

3단계 VTL(Virtual Tape Library) 기반 백업은 디스크 기반 백업과 테이프 기반 장기보존체계를 연결하는 중간 계층으로 기능한다. VTL은 디스크 스토리지 위에 가상의 테이프 드라이브와 카트리지를 구성하여, 백업 소프트웨어가 이를 실제 테이프 라이브러리처럼 인식하도록 한다. 이를 통해 기존 테이프 기반 백업 정책을 유지하면서도 디스크 기반의 빠른 백업 성능과 자동화된 테이프 이관 기능을 동시에 활용할 수 있다. 일반적으로 1차 백업은 VTL로 수신한 후 일정 기간 저장하고 이후 LTO(Linear Tape-Open) 테이프로 자동 이관하는 방식이 적용된다. 이 단계는 백업 정책의 표준화와 운영 효율성 측면에서 중요한 역할을 하며, 중간 계층으로서 3~6개월 수준의 보관 기간을 설정하는 것이 적절하다.

4단계 LTO(Linear Tape-Open) 테이프 기반 장기보존 계층은 국립중앙도서관 디지털자원의 수십 년 단위 장기보존을 담당하는 핵심 계층이다. LTO(Linear Tape-Open) 테이프는 낮은 단위 저장 비용과 긴 매체 수명, 높은 안정성을 동시에 제공하는 장기보존 매체로서, 도서관과 기록관 환경에서 사실상 표준으로 활용되고 있다. LTO-8(Linear Tape-Open-8), LTO-9(Linear Tape-Open-9) 등의 테이프와 로봇 암, 슬롯, 테이프 드라이브로 구성된 테이프 라이브러리를 통해 자동화된 백업이 이루어지며, WORM 테이프를 적용할 경우 원본 데이터의 변조를 구조적으로 차단할 수 있다. 이 계층에서는 최소 10년 이상의 장기보존 정책을 적용하되, 권고 수준으로는 30년 이상의 보존 기간을 설정하는 것이 바람직하다. 또한 LTO(Linear Tape-Open) 세대교체 주기(약 7~10년)에 맞춘 계획적인 마이그레이션과 GFS(Grand-Father/Father/Son) 로테이션 정책을 통해 장기적 안정성을 확보해야 한다.

5단계 소산백업(오프사이트 백업)은 테이프 백업본 중 일부를 외부 물리적 보관소로 이관하여 보관하는 단계로, 데이터센터 전체가 파괴되는 극단적 재난 상황에 대비하는 최후의 방어선이다. 이 방식은 본관 시스템과 물리적으로 완전히 분리된 환경에서 데이터를 보관함으로써, 내부망 장애나 사이버 공격의 영향을 전혀 받지 않는 완전한 에어갭 구조를 제공한다. 일반적으로 월 1회 또는 분기 1회 주기로 테이프 사본을 외부로 이관하며, 1차·2차 테이프 중 최소 하나는 반드시 외부 보관 정책을 적용한다. 소산 백업 매체는 최소 30년 이상 보관을 원칙으로 하고 반입·반출 이력의 전자적 관리와 정기적인 매체 상태 점검을 통해 보존 신뢰성을 유지해야 한다.

마지막으로 6단계 원격지 백업(DR 복제)은 본관과 물리적으로 분리된 DR센터에 데이터를 실시간 또는 주기적으로 자동 복제함으로써, 지역 단위 재해 발생 시에도 서비스와 업무를 지속할 수 있도록 하는 온라인 기반 재해복구 계층이다. DR센터에는 오브젝트 스토리지, VTL, 테이프 라이브러리 등 2차 스토리지를 구축하고 S3 Replication, 스토리지 미러링, 백업 솔루션의 복제 기능을 활용하여 RPO와 RTO 요구사항에 맞는 복제 정책을 수립한다. 일반적으로 데이터 손실 허용량(RPO)은 24시간 이하, 권고 수준으로는 1~4시간 이내로 설정하고 서비스 복구 목표 시간(RTO)은 24~72시간 범위로 설정하는 것이 현실적이다. 또한 연 1~2회 이상의 DR 복구훈련을 통해 실질적인 복구 가능성을 검증하고 DR센터는 본관에서 최소 30km 이상 이격된 위치에 구축함으로써 지리적 재해 대응 능력을 확보해야 한다.

종합하면, 국립중앙도서관의 디지털자원 백업 인프라는 단일 기술이나 단일 매체에 의존하는 방식이 아니라 RAID-디스크 백업-VTL-테이프-소산-DR 복제로 이어지는 6단계 다계층 보호 구조를 통해 설계되어야 한다. 이러한 통합 백업 아키텍처는 일상적 장애부터 대규모 재난에 이르기까지 모든 위험 시나리오에 대응할 수 있는 가장 안정적인 디지털자원 보존·복구 체계로서, 국가대표 디지털 보존기관으로서의 국립중앙도서관 역할을 기술적으로 뒷받침하는 핵심 기반이 된다.

단계	계층	목적
1단계	RAID	서버 내부 디스크 장애 보호(기본 가용성 확보)
2단계	Disk-D2D	빠른 단기 복구 · 운영 오류 대응
3단계	VTL	디스크-테이프 연계 및 자동화
4단계	LTO Tape	수십 년 장기보존 핵심 계층
5단계	소산백업	물리적 · 지리적 분리 기반 최후 보호
6단계	원격지 DR	온라인 기반 재해복구 · 업무연속성

[표 32] 국립중앙도서관 디지털자원 6단계 백업 아키텍처

3.3. 디지털자원 보존을 위한 인공지능 기술 적용방안

3.3.1. 생성형 AI 기술

생성형 AI를 이해하기 위해서는 먼저 인공지능, 머신러닝, 딥러닝이라는 세 가지 핵심 개념의 층위를 짚고 넘어갈 필요가 있다. 가장 상위의 개념인 인공지능(Artificial Intelligence, AI)은 인간만이 수행할 수 있다고 여겨졌던 사고, 추론, 판단, 학습 기능을 컴퓨터가 모방하거나 대체하도록 하는 모든 기술을 포괄한다. 체스를 두거나, 이미지 속 객체를 식별하거나, 긴 글을 읽고 요약하는 것처럼 복잡한 인지적 행위를 자동화하는 기술 모두가 AI에 해당한다.

이러한 AI를 구현하는 여러 방법 중 하나가 머신러닝(Machine Learning, ML)이다. 머신러닝은 사람이 규칙을 직접 작성하던 기존의 프로그래밍 방식과 달리, 컴퓨터가 데이터로부터 스스로 패턴을 찾아내고 그 패턴을 기반으로 규칙을 형성하는 학습 중심 접근 방식이다. 그러므로 전통적 프로그래밍이 사람이 규칙을 만들고 컴퓨터가 이를 따른다는 구조였다면, 머신러닝은 데이터와 목표를 제시하면 컴퓨터가 규칙을 스스로 찾아낸다는 구조로 발전하였다. 이러한 변화는 인간이 소프트웨어와 기계를 다루는 방식을 근본적으로 바꾸며, 데이터 중심의 지능형 시스템이 본격적으로 등장하는 기반이 되었다.

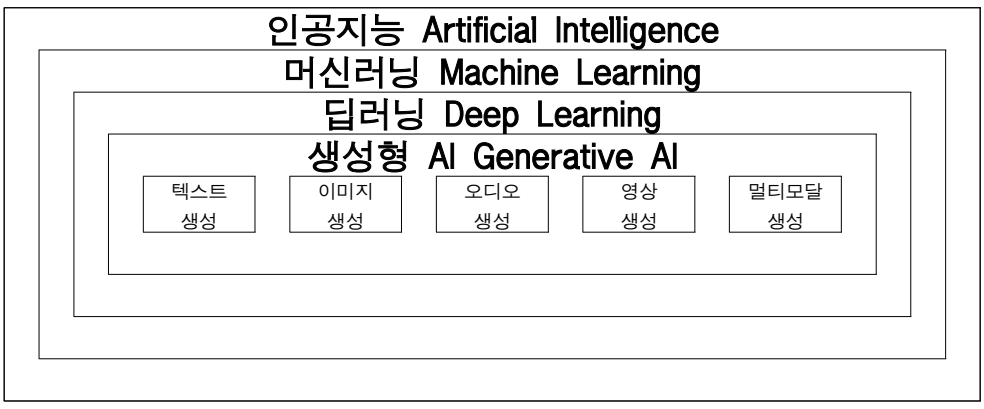
머신러닝 안에는 다양한 기법이 존재하는데, 그중에서도 최근의 인공지능 혁신을 견인한 중심 기술이 바로 딥러닝(Deep Learning)이다. 딥러닝은 인간의 뇌 신경망 구조에서 영감을 얻은 인공신경망(Artificial Neural Network)을 깊고 복잡하게 쌓아 구성한 모델을 활용한다. 기존의 머신러닝에서는 사람이 데이터를 분석해 특징을 추출한 뒤, 이를 분류 알고리즘에 입력하는 절차가 별도로 이루어졌다면, 딥러닝은 특징 추출과 분류를 하나의 통합된 신경망이 동시에 수행할 수 있도록 설계된다. 그 결과 이미지 인식, 음성 인식, 기계 번역, 자연어 처리 등에서 인간 수준 또는 그 이상에 가까운 성능을 보이며, AI가 본격적으로 대중의 주목을 받는 계기가 되었다.

이러한 기술적 흐름 속에서 생성형 AI(Generative AI)가 등장하였다. 생성형 AI는 단순히 기존 데이터를 분류하거나 예측하는 데 그치지 않고 새로운 텍스트, 이미지, 음성, 영상 등을 직접 만들어내는 능력을 갖추고 있다는 점에서 이전의 AI와 구별된다. 예를 들어, 사용자가 한 문장을 입력하면 이어서 자연스러운 문장을 생성하거나, 텍스트 설명을 제공하면 그림이나 디자인 시안을 만들어내며, 짧은 멜로디를 들려주면 그 분위기에 맞는 음악을 작곡하는 등 창작적 산출물을 만들어낸다.

대표적인 사례가 2022년 11월 공개된 OpenAI의 ChatGPT이다. ChatGPT는 대규모 언어 모델(GPT)을 기반으로 사용자의 질문을 이해하고 관련 정보를 바탕으로 내용 요약, 재구성, 자연스러운 문장 생성까지 수행하는 대화형 생성형 AI이다. 또한 대화의 이전 맥락을 기억해 적절한 답변을 제공함으로써, 사람이 컴퓨터와 상호작용하는 방식에 새로운 변화를 가져왔다. 이를 통해 인공지능은 전문가와 개발자만의 도구를 넘어, 누구나 일상적으로 활용할 수 있는 대화형 지식 도구로 자리매김하게 되었다.

생성형 AI는 매우 다양한 생성 능력을 갖고 있으며, 그 활용 범위도 계속 확장되고 있다. 텍스트 생성 영역에서는 문서 작성, 요약, 번역, 질의응답, 스토리·시나리오 작성 등이 가능하며, 이미지 생성 분야에서는 삽화, 포스터, 예술 작품, 디자인 시안 등 다양한 시각 콘텐츠를 생성할 수 있다. 또한 나레이션·안내 음성·배경 음악과 같은 오디오 생성, 간단한 영상·애니메이션 제작 등 영상 생성 능력도 빠르게 고도화되고 있다. 최근에는 텍스트, 이미지, 음성, 영상을 하나의 모델이 동시에 이해하고 생성하는 멀티모달(Multimodal) AI가 확산되면서, 생성형 AI는 단일 기능을 넘어 복합적 창작을 수행하는 종합 지능체로 발전하고 있다.

결국 생성형 AI는 인공지능의 기술적 진화를 집약한 결과물이며, 데이터 기반 학습과 딥러닝 기술이 결합된 새로운 형태의 창작·지식 생성 도구로 자리 잡고 있다. 이는 공공·산업·문화 등 다양한 분야에서 혁신적 변화를 일으키고 있으며, 앞으로 AI 생태계의 중심적 역할을 담당할 것으로 전망된다.



[그림 22] AI 개념도

국립중앙도서관은 방대한 양의 디지털자원이 빠른 속도로 증가하고 문서・사진・음성・영상 등 다양한 형식의 자료가 동시에 생산되는 환경에 놓여 있다. 이처럼 복잡한 환경에서 사람이 모든 관리・보존・활용 업무를 수작업으로 처리하는 것은 점점 더 어려워지고 있다.

따라서 국립중앙도서관은 생성형 AI와 멀티모달 AI 기술을 결합하여, 디지털자원의 생산-관리-보존-활용 전 과정에 지능형 자동화를 도입하는 방향으로 전략을 전환할 필요가 있다.

3.3.2. 디지털자원 인공지능 서비스 요소

도서관이나 기록관에서 인공지능 기술을 적용한 대표적인 사례로 메타데이터 생성 및 키워드 추출 자동화 서비스를 들 수 있다. 아날로그 자료의 재디지털화(Retrodigitization) 자료와 본디지털(Born-Digital) 자료가 어떻게 AI를 활용한 분석 절차를 거쳐 메타데이터와 검색 인덱스를 형성하게 되는지를 단계적으로 보여준다.

필름, 사진, 문서, 음반 등 다양한 유형의 아날로그 자료가 재디지털화 과정을 통해 1차 데이터(Primary Data)로 전환된다. 이 과정에서는 스캔, 영상 변환, 오디오 디지털화 등의 기본적인 디지털 변환 작업이 수행된다.

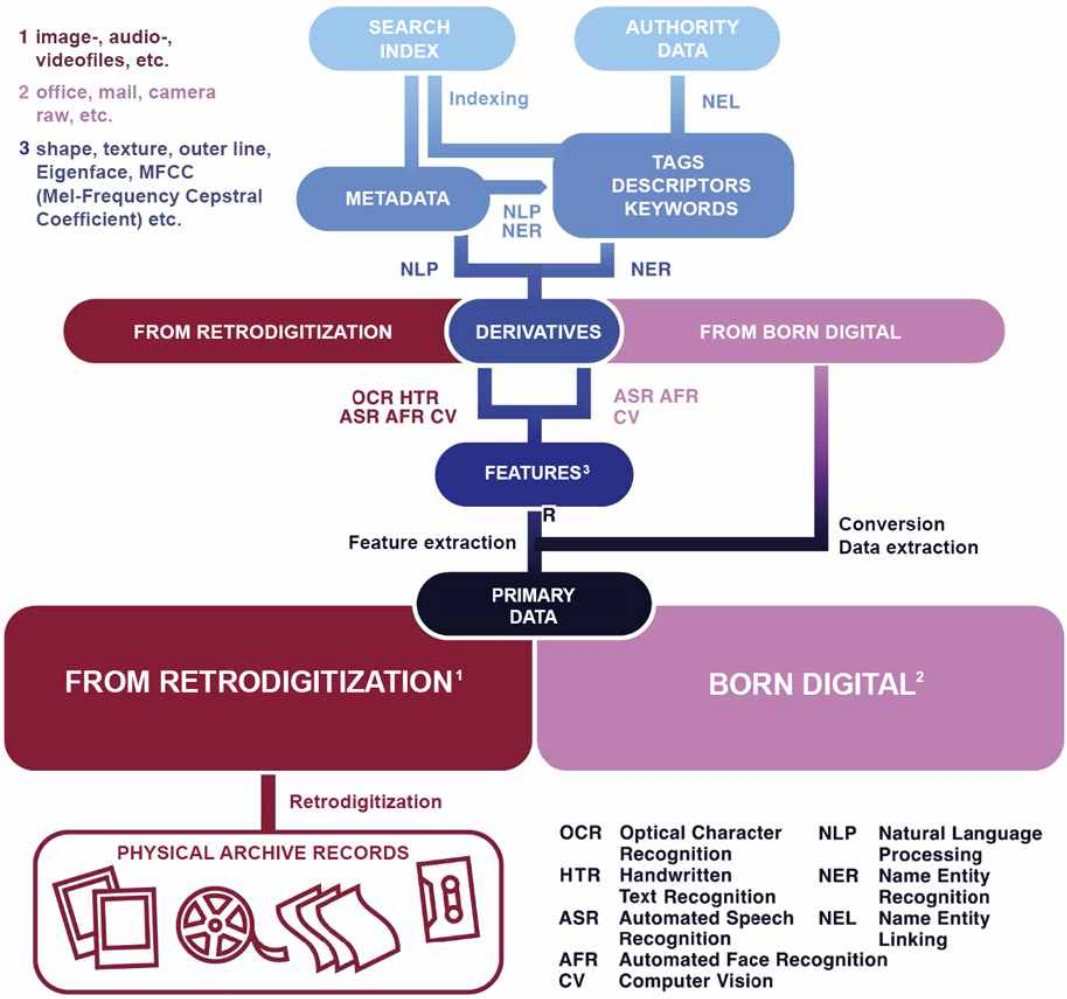
1차 데이터는 이후 특징(Feature) 추출 단계로 이동한다. 이 단계에서 AI 기술(OCR・HTR・ASR・AFR・CV)이 활용되어 문자인식(OCR), 필기체 인식(HTR), 음성인식(ASR), 얼굴 인식(AFR), 영상 분석(CV)을 통해 자료의 구조적・내용적 특징이 자동으로 추출된다. 본디지털자료 또한 구조 파악과 데이터 변환을 통해 동일한 특징 추출 단계로 연결된다. 특징 추출은 자료의 외형, 텍스처, 음향 특성(MFCC) 등 다양한 속성을 포함한다.

이렇게 도출된 특징들은 파생 데이터(Derivatives)로 집계되며, 이는 메타데이터 생성의 핵심 기반이 된다. 예를 들어 문서에서 인식된 텍스트, 이미지에서 추출된 객체 정보, 음원에서 추출된 발화 단위 등이 모두 파생 데이터의 범주에 속한다.

파생 데이터는 다시 상위의 메타데이터(Metadata) 계층으로 전달된다. 이 단계에서는 NLP(자연어 처리)와 NER(개체명 인식) 기술이 적용되어 문서의 주제, 키워드, 개체명, 관계 정보를 자동으로 식별한다. 이렇게 생성된 태그(Tags)・설명어(Descriptors)・색인어(Keywords)는 검색・분류・정책 서비스를 위한 기반 데이터를 제공한다.

메타데이터 계층에서 생성된 기술 정보는 검색 인덱스(Search Index)와 전거 데이터

(Authority Data) 시스템으로 연계된다. 검색 인덱스는 이용자가 디지털자원을 효과적으로 검색할 수 있도록 색인 구조를 구성하며, 전거 데이터는 인물·기관명 등에 대한 식별·정규화 된 정보를 제공하여 검색품질과 데이터 일관성을 강화한다.[58]



[그림 23] AI 기반 디지털자원 심층 인덱싱 자동화 사례

인공지능 기술은 방대한 디지털 컬렉션을 보다 정확하고 효율적으로 다루기 위한 핵심 기반으로 자리 잡아가고 있다. 특히 제목 생성, 메타데이터 기술, 분류·색인, 요약, 포맷 분석, 전사, 중복 식별, 파일 진단 등 다양한 단계에서 AI가 적용되면서, 디지털자원의 생산부터 보존·활용에 이르는 전체 흐름이 근본적으로 변화하고 있다. 변화 속에서 디지털자원 관리·보존 관점에서 적용 가능한 인공지능 서비스 모델과, 훼손 또는 저품질 디지털자원을 복원하기 위한 복원 관점의 인공

58) Jonas Arnold etc., Machine learning in the archive: depth indexing in the service of access to archives, June 2024

지능 서비스 모델을 체계적으로 제시하고자 한다.

우선, 디지털자원이 생산되거나 수집된 직후 가장 먼저 수행되는 업무는 제목 부여와 메타데이터 작성이다. AI는 자료의 본문, OCR・STT 결과, 요약, 목차 등 다양한 텍스트를 분석하여 자료의 성격에 맞는 제목을 자동으로 제안할 수 있으며, 기관 규칙과 형식까지 반영하여 품질을 일정하게 유지할 수 있다. 이어서 AI는 작성자, 발행일, 기관명, 문서유형, 주제, 분류기호 등 방대한 메타데이터 필드를 자동으로 추출・생성하고 담당자는 이를 검토하는 방식으로 업무를 크게 경감할 수 있다. 이러한 기술은 디지털자원의 등록 과정에서 발생하던 오류와 편차를 줄이고, 메타데이터 품질을 일정 수준 이상으로 유지하는 데 큰 역할을 한다.

주제 분류와 색인 과정에서도 인공지능 모델이 중요한 역할을 한다. KDC・LCC・DDC 등 분류체계를 학습한 모델은 문서와 영상・음성 자료의 내용을 분석해 적합한 분류기호를 추천하고 시소러스 기반 색인어를 자동으로 제안함으로써 자료의 조직화와 검색 효율성을 개선한다. 또한 텍스트뿐 아니라 이미지와 영상, 음성까지 함께 처리하는 멀티모달 요약 기술을 활용하면, 이용자는 방대한 자료를 일일이 열람하지 않고도 핵심 내용을 신속하게 파악할 수 있다. 검색 결과에서도 요약문을 즉시 제공해 이용 만족도를 높일 수 있다.

디지털자원의 장기보존 측면에서도 AI 기술은 중요한 도구로 활용된다. 파일 포맷을 분석해 장기보존에 적합하지 않은 자료를 식별하고 PDF/A・TIFF・FFV1 등 보존형 포맷으로 변환해야 할 필요성을 제시할 수 있다. 또한 파일 내부 손상, 프레임 누락, 오디오 끊김, 인코딩 오류 등 보존 위험을 조기에 파악하여, 담당자가 미리 조치할 수 있도록 경고를 제공한다. 이런 기능은 디지털자원의 건강 상태를 지속적으로 모니터링하고 보존 인프라의 안정성을 확보하는 데 도움이 된다.

이미지・스캔본・음성・영상 등에서 텍스트를 추출하는 OCR/STT 기술도 AI 기반 전처리로 크게 개선되고 있다. AI는 전사된 텍스트의 오탈자와 띄어쓰기를 자동 보정하고 문장 구조를 자연스럽게 다듬음으로써, 검색・요약・분류 등 후속 서비스의 정확성을 높인다. 더 나아가 중복・유사 자료 식별 기술을 통해 방대한 컬렉션에서 중복 파일을 제거하거나 관계성을 정리할 수 있으며, AI 기반 포맷 식별 기술은 기존 시그니처 기반 도구의 한계를 보완해 포맷・버전・코덱을 더 정밀하게 판단할 수 있게 한다.

디지털자원의 관리뿐 아니라 복원 과정에서도 AI의 역할은 점점 확대되고 있다. 문서 복원 기술은 희미한 글자와 오염, 낙장 등을 보정하고 레이아웃과 구조를

복구하여 열람 가능한 상태로 되살린다. 이미지 복원 기술은 저해상도 사진의 화질을 향상시키고, 노이즈를 제거하고 손상된 영역을 자연스럽게 보완하여 서비스용 이미지로 활용할 수 있게 한다. 영상 복원 역시 노이즈 제거, 프레임 보간, 색 보정, 흔들림 완화 등을 통해 저화질 영상의 품질을 크게 높일 수 있으며, 오래된 필름 아카이브의 활용도를 높이는 데 효과적이다. 음원 복원에서는 테이프 잡음이나 왜곡된 주파수를 개선하여, 구술 기록이나 현장 녹음 자료의 가독성(청취 가능성)을 높이고 전사 정확도를 향상시킨다.

이와 같이, 인공지능 기반의 다양한 서비스 모델은 디지털자원의 생성, 기술, 분류, 요약, 분석, 진단, 보존, 복원에 이르기까지 전 생애주기를 아우르며, 기존의 수작업 중심 업무를 지능화·자동화하는 방향으로 발전하고 있다. 특히 대규모 디지털 컬렉션을 관리하는 기관에서는 이러한 기술을 활용함으로써 업무 효율성을 높일 뿐 아니라 자료의 품질과 접근성을 향상시키고, 장기적으로는 국민에게 더 나은 디지털 지식 서비스를 제공할 수 있는 기반을 마련하게 된다.

구분	AI 서비스 요소	설명
디지털자원 관리 및 보존 관점 인공지능 서비스 모델	제목 자동 생성	디지털자원의 본문 · 전사 내용을 분석해 기관 규칙에 맞는 제목 후보를 자동 생성하여 기술 품질을 높이고 작업 부담을 경감하는 서비스
	메타데이터 자동 생성	작성자 · 발행일 · 주제 · 분류기호 등 핵심 메타데이터를 AI가 자동 추출 · 생성하여 등록 · 관리 효율을 높이는 기술
	자동 분류 · 색인(주제 분류, 색인어 추천)	문헌 · 영상 · 음성 등의 내용을 분석해 KDC/LCC/DDC 분류기호와 색인어를 자동 추천하여 검색 · 편목 업무를 자동화
	멀티모달 요약(문서 · 이미지 · 영상 · 음성)	텍스트 · 이미지 · 영상 · 음성 전체를 분석해 핵심 내용을 요약문으로 생성하여 빠른 이해와 활용을 지원
	포맷 분석 · 장기보존형 변환 · 손상 · 위험 탐지	파일 포맷 · 무결성 · 저장 상태를 자동 분석해 보존 위험을 예측하고 필요 시 보존형 포맷 변환을 제안
	OCR/STT 전사 및 자동 정제	스캔본 · 영상 · 음성에서 텍스트를 추출하고 오탈자 · 띄어쓰기를 자동 보정하여 검색 가능성과 품질을 확보
	중복 · 유사 디지털자원 식별	텍스트 · 이미지 · 영상 · 음원을 임베딩 기반으로 비교하여 중복 또는 유사 자료를 자동 탐지해 저장 최적화
	포맷 식별 자동화	바이트 패턴 · 헤더 · 코덱 등을 분석하여 기존 도구보다 정밀하게 포맷과 버전을 식별하는 AI 기반 기술
	파일 진단	문서 · 이미지 · 영상 · 음원의 손상 여부, 프레임 누락, 오디오 오류 등을 자동 탐지하여 사전 경고를 제공
디지털자원 복원 관점 인공지능 서비스 모델	문서 복원	희미한 글자 · 오염 · 낙장 등을 AI로 복원하고 OCR과 레이아웃 복구를 통해 문서를 열람 가능한 상태로 개선
	이미지 복원	저해상도 · 손상 사진을 고해상도 · 무노이즈 이미지로 복원하고 색상 · 명암을 재구성하여 활용 품질을 향상
	영상 복원	오래된 필름이나 저화질 영상의 노이즈 제거, 프레임 보간, 색 보정, 해상도 향상을 수행하여 가독성을 높임
	음원 복원	테이프 잡음, 환경 소음 등이 섞인 음원을 정제하고 음성 · 음악의 명료도를 향상시켜 전사 · 서비스 품질 개선

[표 33] 디지털자원 보존을 위한 AI 서비스 요소

3.3.3. 디지털자원 관리 및 보존을 위한 인공지능 서비스 제안

❑ 제목 자동 생성

디지털자원 관리・보존 과정에서 제목은 이용자와 시스템이 가장 먼저 마주하는 핵심 식별 정보이다. 제목이 부정확하거나 규칙에 맞지 않으면 검색・분류・보존・서비스 전 과정에 영향을 준다. 지금까지는 생산자나 담당자가 제목을 직접 입력하는 방식이 일반적이었고, 이 과정에서 오타, 표현 불균형, 내용과 맞지 않는 제목이 자주 발생했다. 생성형 AI를 활용한 제목 자동 생성은 이러한 문제를 줄이고 생산 단계부터 일관되고 품질 높은 제목을 확보하기 위한 기술이다.

생성형 AI 기반 제목 자동 생성 시스템을 구축하려면 먼저 학습 데이터를 정의해야 한다. 가장 중요한 데이터는 디지털자원의 원문 텍스트이다. 문서 본문, OCR 결과, 영상・음성 전사(STT) 텍스트 등 자료의 내용과 주제를 직접 보여주는 자연어 텍스트가 기본이 된다. 여기에 기존 자료의 제목 데이터세트를 함께 학습시켜 내용과 제목 간 패턴을 익히도록 한다. 부제, 요약문, 목차 등 제목과 밀접한 텍스트도 함께 활용하면 자료의 핵심을 더 정확히 파악할 수 있다.

또한 국립중앙도서관이 보유한 제목 작성 규칙과 형식 정보도 중요하다. 문서 유형, 연도, 주제, 부서명 요소를 어떤 순서와 형식으로 배열하는지, 괄호와 기호를 어떻게 사용하는지, 길이를 어떻게 제한하는지에 대한 내부 규칙이 있다면 이를 데이터로 정리해 모델에 반영한다. KORMARC 245 필드에 규정된 제목・책임표시 형식도 주요 기준이 된다. 이를 통해 모델은 단순 생성이 아니라 기관의 기술 규범을 준수하는 제목을 만들 수 있다.

전문가가 검토・수정한 고품질 제목 사례도 핵심 학습 자원이다. 잘못된 제목과 수정된 제목을 쌍으로 구성하면 모델은 표현 축소, 정보 추가, 배열 순서 조정 방식까지 학습할 수 있다. 자연어 텍스트, 형식 규칙, 골드세트가 함께 결합되어 제목 자동 생성 모델의 기반을 이룬다.

알고리즘 측면에서는 여러 모델이 결합된다. 중심 모델은 GPT, LLaMA 계열 대규모 언어모델(LLM)과 T5, BART 같은 Sequence-to-Sequence 모델이다. 이 모델들은 긴 문서나 전사 텍스트를 읽고 한두 줄의 제목으로 압축하는 데 강점이 있다. 그러나 생성 능력만으로는 기관 규칙을 자동 준수하기 어렵기 때문에 내용 이해 모델과 형식 제어 모델을 함께 구성해야 한다.

이를 보완하기 위해 자연어 이해(NLU) 모델을 활용한다. BERT, RoBERTa 같은

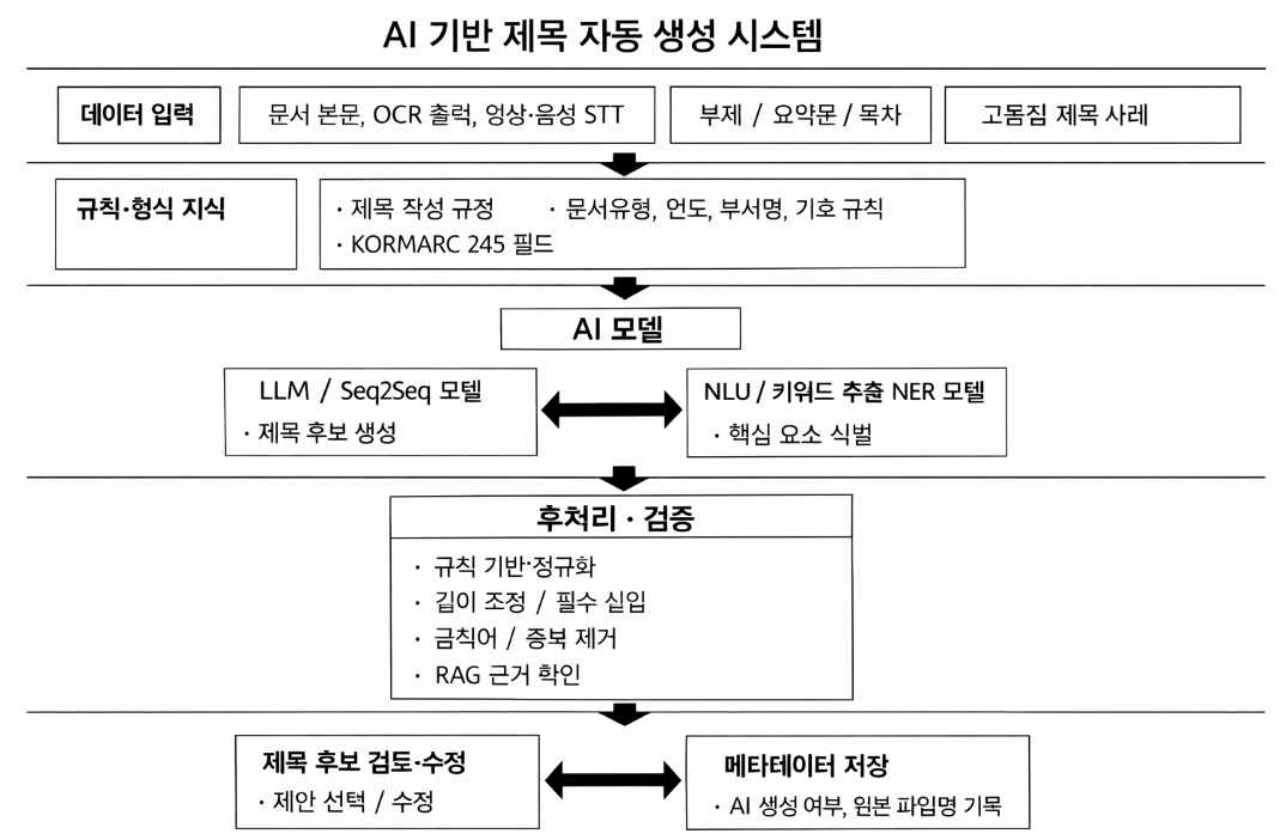
의미 분석 모델은 핵심 주제와 키워드를 추출한다. NER 모델은 인명, 기관명, 지명, 연도를 식별해 제목 구성 요소를 정리한다. 예를 들어 기관 보고서라면 기관명과 연도를 포함하고, 인명이나 지명이 핵심인 자료라면 이를 제목 요소로 반영한다.

생성 이후에는 스타일과 형식 제약을 적용하는 후처리 단계가 필요하다. 제목 길이를 제한하고 기관명·부서명·연도 등 필수 요소를 규칙에 맞게 자동 삽입한다. 괄호, 슬래시, 특수기호를 정규화하고 불필요한 기호를 제거한다. 이 과정을 통해 내용과 형식을 모두 충족하는 표준화된 제목을 완성한다.

업무 절차는 다음과 같다. 디지털자원이 시스템에 이관되면 문서·이미지·영상·음원 파일을 등록한다. 이미지와 영상에는 OCR을 적용해 텍스트를 추출하고, 영상과 음원에는 STT를 적용해 내용을 확보한다. 확보된 텍스트를 기반으로 LLM이 주제, 내용, 문맥, 자료 유형을 분석해 제목 후보 2~5개를 생성한다. 의미 중심 자연어 제목과 기관 규칙을 반영한 규칙형 제목을 함께 제시하면 선택 폭을 넓힐 수 있다.

생성된 제목은 후처리를 거친다. 길이, 금칙어 포함 여부, 중복 표현을 자동 점검한다. 이후 담당자가 화면에서 후보를 검토하고 선택하거나 일부 수정해 최종 제목을 확정한다. 확정된 제목은 메타데이터에 자동 기록되고 검색·분류·보존·서비스 전 과정에서 일관되게 사용된다. 생산자는 제목을 처음부터 작성하는 대신 AI 제안을 검토·조정하는 수준으로 업무 부담을 줄일 수 있다.

제목 자동 생성에는 몇 가지 제약이 있다. 대표적인 문제는 환각(Hallucination)이다. 생성형 AI는 원문에 없는 정보를 추가할 수 있다. 예를 들어 보고서에 없는 기관명이나 연도를 제목에 삽입할 수 있다. 이를 방지하려면 생성 결과가 원문 근거에 기반하도록 설계해야 한다. RAG 방식처럼 근거 문장을 함께 제시하는 구조도 유효하다. 무엇보다 사람의 최종 검토 단계를 반드시 유지해야 한다.



[그림 24] 제목 자동 생성 단계

또 하나의 고려사항은 기관 규칙과의 충돌 문제이다. 생성형 AI가 제안한 제목이 내용 측면에서는 적절하더라도, 국립중앙도서관의 제목 작성 규칙이나 KORMARC 245 필드 형식과 맞지 않을 수 있다. 따라서 제목 자동 생성 시스템에는 규칙 기반의 후처리·검증 모듈을 함께 두어 AI가 만든 결과물을 한 번 더 필터링하고 정규화하는 과정이 포함되어야 한다. 예를 들어 필수 요소 누락, 순서 오류, 허용되지 않는 기호 사용 등을 자동으로 잡아내고 수정하도록 할 수 있다.

마지막으로 원본성의 관리도 중요한 이슈이다. 자동 생성된 제목은 사람이 직접 작성한 제목과 구별할 수 있도록 메타데이터에 제목 자동 생성 여부 필드를 두어 명시적으로 기록하는 것이 바람직하다. 또한 파일 원래의 이름(원본 파일명)과, 메타데이터로서의 제목을 구분해 기록해야 나중에 출처를 추적하기 쉽다. 이렇게 해 두면, 연구자나 이용자가 필요할 경우 원래 파일은 어떤 이름이었고, AI가 부여한 제목은 무엇인지를 명확히 알 수 있고, 향후 제목 생성 모델을 개선할 때도 피드백 자료로 활용할 수 있다.

❑ 메타데이터 자동 생성

디지털자원은 단순히 파일 하나만 보관된다고 해서 잘 관리되고 활용되는 것이 아니다. 누가, 언제, 어떤 목적으로 만들었는지, 어떤 주제에 속하는지, 어떤 형식인지, 어디에 어떻게 연계되어 있는지 등을 설명해 주는 메타데이터가 함께 있어야 비로소 제대로 된 관리와 서비스가 가능하다. 그러나 지금까지 메타데이터 작성은 대부분 생산자나 소수의 전문가가 일일이 입력하는 노동집약적인 방식에 의존해 왔고, 이 과정에서 누락・오류・불균형이 자주 발생해 왔다. 생성형 AI와 각종 분석 모델을 결합하면, 문서・사진・영상・음원 등 다양한 유형의 디지털 자원에 대한 메타데이터를 자동으로 생성하고 사람은 최종 검토와 수정에 집중하는 방식으로 업무 구조를 바꿀 수 있다.

이를 위해서는 먼저 AI가 참고할 학습 데이터를 잘 준비하는 것이 중요하다. 메타데이터 자동 생성 모델은 작성자, 기관명, 발행일, 문서유형, 언어, 주제, 분류기호(KDC/LCC) 등과 같이 이미 사람이 입력해 놓은 구조형 메타데이터를 학습한다. 여기에 문서 본문, OCR・STT 결과, 캡션, 도표・표 구조, 이미지 안에 등장하는 지도나 인물 같은 시각 요소까지 함께 이용하면, AI가 자료의 전체적인 맥락을 이해하는 데 도움이 된다. 또한 KORMARC, 더블린 코어(Dublin Core), PREMIS (Preservation Metadata: Implementation Strategies), METS(Metadata Encoding and Transmission Standard) 등 각종 표준 메타데이터 규정과 기관 내부 규칙, 시소러스와 표준 분류체계 같은 도메인 지식을 함께 제공하면, AI가 단지 비슷한 말을 찍어내는 수준을 넘어 도서관에 맞는 형식으로 메타데이터를 만들어낼 수 있다. 마지막으로 사람이 잘 작성해 놓은 메타데이터와 원문을 쌍으로 연결한 골드세트는 모델이 어떤 자료에 어떤 메타데이터가 적절한지를 배우는 데 기준점 역할을 한다. 여기서 골드세트(Gold Set)란 전문가가 정확성을 검증한 원문–메타데이터 쌍으로 구성된 기준 학습 데이터 집합을 의미한다.

알고리즘 측면에서는 여러 유형의 AI 모델이 함께 쓰인다. 우선, NER(명명된 개체 인식)을 포함한 정보 추출 모델이 문장 속에서 사람 이름, 기관명, 지명, 연도 등 중요한 정보를 자동으로 찾아낸다. 문서의 레이아웃 구조를 분석하는 LayoutLM・Donut 같은 모델은 제목, 본문, 표, 각주, 서명란 등을 구분해내어, 어느 위치의 내용이 어떤 메타데이터에 해당하는지 파악한다. 여기에 문서 유형과 주제를 자동으로 분류하는 텍스트 분류 모델, KDC/LCC 등 여러 개의 분류기호나 색인어를 동시에 추천하는 멀티레이블 분류 모델, 그리고 요약문을 만들어 주거나 언어・길이・유형 등을 자동 생성해 주는 생성형 모델이 결합된다. 더 나아가 시맨틱(의미 기반)

분석 모델은 문맥을 읽고 관련 주제를 추론하거나, 온톨로지를 이용해 특정 개체를 표준 용어와 연결하는 역할을 한다.

서비스 절차를 실제 업무 흐름으로 보면 다음과 같이 이해할 수 있다. 먼저 디지털 자원이 업로드되거나 이관되면, OCR·STT 등을 통해 텍스트를 최대한 확보하고 이미지와 표·도표 구조 등도 자동으로 분리·분석한다. 그 다음 단계에서는 NER와 규칙을 이용해 작성자, 기관명, 작성일, 언어, 문서유형과 같은 기본 메타데이터가 자동으로 추출된다. 이어서 주제·분류·색인어 단계로 넘어가면, KDC/LCC 분류 기호가 자동 추천되고, 시소러스와 연계된 색인어 후보가 제시된다. 본문을 기반으로 한 요약문도 자동으로 생성되며, 영상·음성 자료는 STT를 거친 뒤 멀티모달 요약을 적용할 수 있다. 마지막으로 각 필드별로 신뢰도 점수를 부여하고 충돌·누락을 자동 검출한 후, 담당자가 이를 검토·수정해 확정된 메타데이터를 PREMIS(Preservation Metadata: Implementation Strategies)/METS(Metadata Encoding and Transmission Standard) 구조로 저장하고 검색·추천·챗봇 등 서비스에 반영한다.

다만 이런 자동화에도 여러 제약과 주의사항이 존재한다. 우선 원문 품질이 낮으면 메타데이터 추출의 정확도도 함께 떨어진다. OCR·STT에서 발생한 오류는 그대로 누적될 수 있으므로, 후처리 모델과 사람 검토 단계가 반드시 필요하다. 또한 KORMARC, DC, PREMIS(Preservation Metadata: Implementation Strategies) 등 각종 표준의 필드 형식과 제어어휘를 엄격히 지켜야 하기 때문에, AI가 제안한 값이 형식 요건을 만족하는지 검증하는 절차가 필요하다. 행정기록과 역사기록처럼 분야별로 구조가 다른 자료에 대해서는 하나의 모델만으로는 한계가 있으며, 분야별 특화 모델이나 추가 규칙이 요구된다. 마지막으로 자동 생성된 메타데이터에는 개인정보·민감정보가 포함될 수 있으므로, 비식별화 기준을 적용하고 민감한 항목을 필터링·마스킹하는 절차를 함께 설계해야 한다.

구분	오픈소스 /클라우드 명칭	설명	적용사례
파일 · 미디어의 기술 메타데이터 (Technical metadata) 추출 툴	Apache Tika[59]	• 1,000개+ 파일 유형에서 텍스트 · 메타데이터를 추출하는 범용 툴킷	• NASA JPL 환경[60] • 미국 OSTI(Office of Scientific and Technical Information)[61]
	Tesseract OCR	• 텍스트가 없는 이미지 기반 자원을 검색 가능하게 만들고, 이후 LLM/키워드/NER 기반 의미 메타데이터 생성의 입력을 제공	• 미국 의회도서관(LOC) – Chronicling America 신문 컬렉션[62] • Biodiversity Heritage Library (BHL) – Internet Archive 디지털화 파트너를 통해 생성되는 OCR에 적용[63]
	ExifToo[64]	• 이미지 · 오디오 · 비디오 · PDF 등 다양한 파일의 메타 정보를 읽고/쓰고/편집하는 커맨드라인/라이브러리	• AVPreserve가 – MDQC (Metadata Quality Control)[65]
	MediaInfo[66]	• 비디오 · 오디오 파일의 기술/태그 데이터(컨테이너, 코덱, 길이, 비트레이트 등)를 보여주는 도구	
문서 인지(IDP) 기반 구조 메타데이터 자동 추출	Amazon Textract[67]	• 텍스트뿐 아니라 폼 · 테이블 · 레이아웃 요소까지 추출하는 ML 서비스	• AWS – Textract로 스캔 문서 텍스트를 추출한 뒤 Amazon Comprehend로 엔터티(개체) 추출을 수행[68]
	Azure Document Intelligence [69]	• 일반 문서 모델에서 텍스트 · 테이블 · 키–값 등 구조 추출을 제공	• Yeeflow – Azure Document Intelligence를 활용한 인보이스 인식 자동화[70]
	Google Document AI[71]	• 커스텀 추출기에서 스키마 정의 엔터티를 추출하는 옵션을 제공(파운데이션/커스텀 등)	• FibroGen – Document AI로 인보이스 처리 자동화를 구현하여 ROI를 얻음[72]

[표 34] 메타데이터 자동 생성을 위한 오픈소스 및 클라우드

59) https://tika.apache.org
60) https://pdfa.org/wp-content/uploads/2020/07/Evaluating-Text-Extraction-at-Scale-TAllison.pdf
61) https://www.osti.gov/servlets/purl/1897834
62) https://blogs.loc.gov/headlinesandheroes/2025/04/ocr-reprocessing
63) https://about.biodiversitylibrary.org/ufaqs/what-is-optical-character-recognition-ocr-and-how-does-bhl-use-it
64) https://exiftool.org
65) https://www.weareavp.com/mdqc-case-studies
66) https://mediaarea.net/ko/MediaInfo
67) https://aws.amazon.com/ko/textract
68) https://aws.amazon.com/ko/blogs/machine-learning/extracting-custom-entities-from-documents-with-amazon-textract-and-amazon-comprehend
69) https://learn.microsoft.com/en-us/azure/ai-services/document-intelligence/prebuilt/general-document?view=doc-intel-4.0.0
70) https://blog.yeeflow.com/post/invoice-recognition-yeeflow-azure-document-intelligence
71) https://docs.cloud.google.com/document-ai/docs/extracting-overview
72) https://cloud.google.com/blog/products/ai-machine-learning/reducing-invoice-processing-with-document-ai

❑ 자동 분류·색인(주제 분류, 색인어 추천)

자동 분류・색인 서비스는 디지털자원의 내용을 분석해 이 자료가 어떤 주제에 속하는지, 어떤 분류기호(KDC/LCC/DDC)를 붙여야 하는지, 어떤 색인어로 검색될 수 있어야 하는지를 AI가 먼저 제안해 주는 기술이다. 규모가 큰 디지털 컬렉션에서는 자료마다 일일이 분류기호를 부여하고 색인어를 정하는 일이 막대한 시간과 노력을 요구한다. 자동 분류・색인 모델을 도입하면, AI가 우선 후보를 제시하고 사람은 이를 검토・확정하는 방식으로 업무 부담을 크게 줄이면서도 검색 품질과 재현율(찾고자 하는 자료를 실제로 찾아낼 수 있는 정도)을 높일 수 있다.

자동분류색인서비스를 구현하려면, 분류기호가 부여된 기존자료들을 학습데이터로 준비하는 것이 필요하다. 이 자료들의 원문 텍스트, 제목, 요약, 목차, 메타데이터 등을 AI에게 보여주고 이런 내용의 자료에는 이런 분류기호를 붙인다는 패턴을 학습시킨다. 여기에 주제명, 시소러스 용어, 키워드 목록을 함께 제공하면 AI는 특정 단어와 표현이 어떤 주제 범주와 연결되는지 자연스럽게 습득할 수 있다. 이렇게 축적된 데이터는 새로 들어오는 자료에 대해 가장 비슷한 패턴을 가진 분류기호와 색인어를 자동으로 추천하는 기반이 된다.

알고리즘 측면에서 자동 분류・색인은 주로 Transformer 기반 텍스트 분류 모델을 활용한다. BERT, RoBERTa, KoELECTRA와 같은 모델은 문장을 벡터(숫자 집합)로 변환하여 의미를 분석하고 그 벡터가 어떤 주제 또는 분류기호와 가장 잘 맞는지를 판단한다. 하나의 자료가 여러 개의 주제나 색인어를 동시에 갖는 경우가 많기 때문에, 한 자료에 여러 개의 라벨을 붙이는 멀티레이블 분류 모델이 사용된다. 또한 텍스트 임베딩을 이용해 비슷한 자료끼리 가깝게, 다른 자료는 멀게 배치한 뒤, 최근접 이웃(KNN) 같은 방법으로 이전에 이런 자료에 이런 분류를 붙였으니, 새 자료도 비슷하다는 방식으로 추천을 강화하기도 한다.

서비스 절차를 보면, 먼저 신규 디지털자원이 등록되면 시스템이 해당 자료의 텍스트와 메타데이터(제목, 요약, 목차 등)를 수집한다. 그런 다음 자동 분류 모델이 분석을 수행해 KDC 등 분류기호 후보를 여러 개 제안한다. 시소러스와 연계된 색인어 추천 모델은 관련 키워드와 색인어 목록을 생성해 담당자에게 보여준다. 사람은 제안된 분류기호와 색인어를 검토하여 적절한 것을 선택하거나 일부 수정・보완할 수 있으며, 최종 확정된 결과는 검색엔진 인덱스와 디지털아카이브의 논리적 구조에 반영된다. 이렇게 하면 이용자는 보다 정교한 분류・색인 체계 위에서 디지털자원을 검색하고 관련 자료를 쉽게 찾아볼 수 있게 된다.

물론 이 과정에도 주의해야 할 점이 있다. 우선 분류 체계가 개편되거나 새로운 주제 영역이 등장하는 경우, 기존 모델은 최신 구조를 반영하지 못할 수 있어 주기적인 재학습과 업데이트가 필요하다. 희귀 분야나 자료 수가 적은 특수 주제는 학습 데이터가 부족해 정확도가 떨어질 가능성도 있다. 또한 개인정보나 민감한 단어가 색인어로 그대로 노출되지 않도록 필터링 규칙을 적용해야 한다. 무엇보다도, 이용자나 전문가가 왜 이 분류기호가 선택되었는지를 물을 수 있기 때문에, 완벽한 설명까지는 어렵더라도, 최소한 어떤 단어·문장 때문에 이 분류가 선택되었는지를 대략적으로라도 설명할 수 있는 수준의 설명가능성을 확보해야 한다.

❑ 멀티모달 요약(문서·이미지·영상·음성)

멀티모달 요약은 텍스트 문서뿐만 아니라 이미지, 영상, 음성까지 포함한 다양한 형태의 디지털자원을 AI가 이해하고 그 핵심 내용을 짧은 요약문으로 만들어 주는 기술이다. 이용자 입장에서는 긴 보고서 전체를 일일이 읽지 않거나 긴 강연 영상 전체를 재생하지 않고도 이 자료가 자신에게 필요한 내용인지를 빠르게 판단할 수 있게 해 주는 매우 실용적인 기능이다. 검색 결과 화면에서 각 자료의 요약이 함께 제공되면, 검색 만족도와 이용 효율성이 크게 향상된다.

이를 위해서는 먼저 원문과 요약이 함께 제공된 학습 데이터가 필요하다. 기본적으로 문서 원문 텍스트와 그에 대한 사람이 직접 작성한 요약문을 다량 확보해야 한다. 영상·음성 자료의 경우에는 STT를 통해 전사된 텍스트뿐 아니라 발화자 정보, 시간 구간, 장면 전환 등 타임라인 데이터도 함께 학습에 활용할 수 있다. 이렇게 하면 AI가 어떤 장면에서 어떤 내용이 중요한지를 더 잘 파악할 수 있다. OCR 결과와 도표·캡션 텍스트도 함께 반영하면 다양한 형태의 자료를 균형 있게 요약할 수 있는 기반이 마련된다.

알고리즘은 크게 두 가지 축으로 나뉜다. 하나는 텍스트 중에서 중요한 문장만 골라 붙이는 추출형 요약, 다른 하나는 전체 내용을 압축해 새 문장으로 재구성하는 생성형 요약이다. 실제 서비스에서는 두 방식을 혼합한 모델이 많이 쓰인다. 여기에 텍스트와 이미지를 동시에 처리하는 멀티모달 LLM, 텍스트와 영상 캡션을 함께 이해하는 모델 등이 더해지면 영상·이미지·자막 등을 통합해서 요약하는 것도 가능해진다. 긴 회의록이나 장편 영상처럼 분량이 많은 자료를 다루기 위해서는 긴 문맥을 처리할 수 있는(long-context) 모델이 필요하며, 이 모델들은 수십 페이지에 이르는 문서를 한 번에 요약하도록 설계된다.

서비스 절차를 보면, 먼저 자료가 등록될 때 OCR과 STT를 통해 가능한 많은 텍스트를 확보하고 영상・이미지・표・도표 정보도 함께 정리한다. 그런 다음 요약 모델이 전체 내용을 대상으로 한 전체 요약과, 필요하다면 장면・섹션별 부분 요약을 생성한다. 생성된 요약문은 메타데이터의 요약, 내용 소개 필드에 저장되며, 검색 결과, 전시 설명, 교육 자료 안내 등 다양한 서비스 화면에서 활용될 수 있다. 예를 들어, 긴 강연 영상이라면 “이 영상은 ○○ 주제를 다루며, 주요 내용은 1) △△, 2) □□, 3) ◇◇이다”와 같은 짧은 요약을 먼저 보여줄 수 있다.

다만 요약 과정에서도 여러 가지 위험과 고려사항이 존재한다. AI가 내용을 압축하는 과정에서 중요한 정보가 빠지거나, 일부 표현이 왜곡될 수 있다. 특히 정치・종교・사회적으로 민감한 내용이나 개인정보가 포함된 자료의 경우, 요약문에 민감한 문장만 도드라지게 노출되는 위험이 있어, 별도의 검토・필터링 절차를 마련해야 한다. 또한 이용 대상에 따라 요약문의 길이와 톤을 달리할 필요가 있다. 예를 들어 학술용 요약과 어린이용 요약은 표현 방식과 난이도가 크게 다르므로, 프롬프트 설계를 통해 누가 읽을 요약인지를 미리 지정하고 필요하면 한 자료에 대해 여러 버전의 요약문을 생성하는 체계를 마련하는 것이 바람직하다.

❑ 포맷 분석·장기보존형 변환·손상·위험 탐지

디지털자원이 장기간 안전하게 보존되려면, 파일의 내용만이 아니라 파일 포맷과 저장 상태를 지속적으로 관리해야 한다. 오래되거나 비표준적인 포맷은 시간이 지날수록 열어볼 수 있는 프로그램이 줄어들고, 특정 운영체제나 소프트웨어 버전에 종속되는 위험이 커진다. 또한 스토리지 장비의 노후화나 시스템 오류로 인해 파일이 서서히 손상될 수도 있다. 포맷 분석・장기보존형 변환・손상・위험 탐지 모델은 이러한 위험을 조기에 감지해, 어떤 파일이 보존에 취약한지, 어떤 포맷으로 미리 변환해 두어야 하는지, 어떤 파일이 손상될 가능성이 높은지를 AI가 자동으로 알려주는 기술이다.

이를 위해 AI는 다양한 학습 데이터를 참조한다. HWP, DOC, PDF, AVI, MKV, MP3, WAV 등 여러 포맷의 샘플 파일과, JHOVE・DROID 같은 기존 포맷 검증 도구의 결과 로그를 함께 학습함으로써, 각 포맷이 갖는 구조적 패턴과 특징을 익힌다. 체크섬・무결성 검증 이력, 스토리지 에러 로그, 손상 파일과 정상 파일의 사례 데이터는 어떤 패턴이 나타날 때 손상 위험이 높아지는지를 학습하는 데 사용된다. 이런 데이터를 바탕으로 AI는 단순히 확장자만 보고 판단하는 것이 아니라 헤더와

내부 구조, 바이트 패턴까지 종합적으로 분석해 위험을 진단할 수 있다.

알고리즘 측면에서는 포맷 분류용 CNN/RNN 모델과 시그니처 기반 규칙을 함께 사용하는 하이브리드 방식이 많이 활용된다. 이상 탐지(Autoencoder, Isolation Forest 등) 모델은 평소와 다른 패턴을 보이는 파일이나 로그를 찾아내는 역할을 한다. 시간에 따른 로그 데이터를 분석해 미래의 손상·오류 발생 가능성을 예측하는 시계열 모델(LSTM, Prophet 등)도 사용할 수 있다. 이처럼 규칙과 머신러닝을 결합한 위험도 스코어링 모델을 적용하면, 단순히 정상/비정상이 아니라 위험도가 어느 정도인지까지 숫자로 표현할 수 있어 우선순위 관리를 쉽게 할 수 있다.

실제 서비스에서는, 신규 파일이 들어올 때마다 포맷 자동 분석을 수행해 표준 포맷인지, 비표준 또는 장기보존에 위험한 포맷인지 구분한다. 비표준 포맷으로 판단되면 PDF/A, TIFF, FFV1 같은 장기보존형 포맷으로의 변환 후보를 제안해 준다. 또한, 정기적으로 스토리지 로그와 체크섬 데이터를 분석해 이상 탐지 모델을 돌리고 손상 가능성이 높은 파일을 리스트로 만들어 보존 위험 대시보드에 표시한다. 담당자는 이 대시보드를 보면서 우선적으로 재복제, 재마이그레이션 등의 조치를 취할 수 있다.

다만, 포맷 변환과 위험 탐지에는 여러 가지 주의사항이 따른다. 먼저, 포맷 변환 과정에서는 레이아웃이 미세하게 바뀌거나 숨겨진 메타데이터가 손실되는 등 부작용이 발생할 수 있으므로, 원본 파일과 변환된 보존본을 동시에 보존하는 정책이 필요하다. 변환에 사용하는 소프트웨어가 국제표준을 제대로 준수하는지도 검증해야 한다. 로그와 체크섬 데이터가 제대로 쌓여 있지 않으면, 아무리 좋은 AI 모델이라도 위험을 정확히 감지하기 어렵다. 또한 경보가 너무 많이 발생하면 담당자가 피로를 느끼고 무시해 버릴 수 있고, 반대로 경보 기준을 너무 높게 잡으면 위험한 파일을 놓칠 수 있다. 이 둘 사이의 균형을 잡는 것이 실제 운영에서 매우 중요하다.

❑ OCR/STT 전사 및 자동 정제

디지털자원의 내용이 이미지나 음성 속에만 들어 있다면, 검색을 통해 찾아내거나 다른 AI 서비스에서 활용하기가 어렵다. OCR(문자인식) 과 STT(음성→텍스트 변환) 기술은 이미지·스캔본·음성·영상에서 텍스트를 뽑아내, 내용을 컴퓨터가 이해하고 검색할 수 있는 형태로 바꿔 주는 첫 단계이다. 여기에 생성형 AI를 이용한 자동 정제(오탈자·띄어쓰기·문장 구조 보정)를 더하면, 사람이 일일이

교정하지 않고도 상당히 수준 높은 전사 결과를 얻을 수 있다.

이를 위해서는 다양한 형태의 학습 데이터가 필요하다. 한글·한자·세로쓰기·필사본 등 여러 형식의 스캔 이미지와 그에 대응하는 정답 텍스트, 여러 방언과 억양을 포함한 음성·영상과 정답 전사 데이터가 대표적이다. 해상도, 노이즈, 그림자, 왜곡 등이 서로 다른 이미지와, 주변 소음이 섞인 음성 데이터도 함께 학습해야 실제 환경에 잘 대응할 수 있다. 이렇게 모인 데이터는 AI가 어려운 조건에서 글자와 말을 어떻게 읽어야 하는지를 익히는 데 큰 도움이 된다.

알고리즘으로는 Vision Transformer 기반 OCR 모델이 이미지 속 글자를 인식하며, Whisper 같은 End-to-End ASR 모델이 음성을 텍스트로 변환한다. 그 다음 단계에서 후처리용 언어모델이 등장한다. 이 모델은 맞춤법, 띄어쓰기, 문장 부호, 문장 구조를 자연스럽게 보정해 주어, 사람이 직접 입력한 문장에 가깝게 다듬는다. 예를 들어, “국 립 중 앙 도 서 관 에 서” 처럼 띄어쓰기가 엉망으로 나온 결과를 “국립중앙도서관에서” 로 자동 수정해 주는 식이다.

서비스 절차를 보면, 스캔된 문서나 디지털 이미지, 녹음·영상 자료를 시스템에 올리면 먼저 OCR/ASR을 통해 1차 전사가 이루어진다. 이후 언어모델이 이 전사 결과를 분석해 오탈자·띄어쓰기·문장 구조 등을 자동으로 보정한다. 시스템은 전사 결과의 신뢰도 점수를 계산해, 예를 들어 90% 이하로 판단되는 구간만 사람에게 검토 대상으로 표시할 수 있다. 담당자는 표시된 부분만 확인·수정하면 되므로, 전체 텍스트를 처음부터 끝까지 검토하는 것보다 훨씬 효율적이다. 최종 확정된 전사 텍스트는 메타데이터와 검색 인덱스에 반영되어, 이후 요약·분류·색인·챗봇 등 다양한 AI 서비스의 입력으로 활용된다.

그러나 이 과정이 절대적인 것은 아니다. 필사본, 매우 오래된 고문서, 해상도가 극히 낮은 이미지에서는 인식 정확도가 크게 떨어질 수 있으며, 고어체·한자 혼용·특수기호가 많은 자료의 경우에는 도메인 특화 모델을 추가로 튜닝해야 한다. 녹취·인터뷰처럼 사람 목소리가 그대로 담긴 음성자료는 개인정보가 많이 포함될 수 있기 때문에, 전사 결과에 대해 별도의 보호조치(익명화, 민감표현 마스킹 등)를 마련해야 한다. AI가 전사했다고 해서 곧바로 공개 가능한 것은 아니며, 기술적 정확성과 함께 개인정보 보호 기준을 함께 고려해야 한다.

❑ 중복·유사 디지털자원 식별

대규모로 디지털자원을 수집하다 보면, 같은 자료가 여러 번 들어오거나,

내용은 거의 같은데 버전이나 포맷만 다른 파일이 쌓이는 일이 빈번하게 발생한다. 이러한 중복·유사 자료를 그대로 두면 저장 공간이 낭비되고, 어느 버전을 기준으로 삼을지 혼란이 생길 수 있다. 중복·유사 디지털자원 식별 모델은 텍스트, 이미지, 영상, 음원 등을 모두 숫자 벡터로 바꾸어, 서로 얼마나 비슷한지를 AI가 계산해 주는 기술이다. 이를 통해 사람은 어떤 자료들을 묶어서 관리해야 할지를 빠르게 파악할 수 있다.

이 모델을 학습시키기 위해서는 우선, 중복·유사·비유사 관계를 사람이 라벨링해 둔 사례 데이터가 필요하다. 예를 들어, 같은 보고서의 PDF 버전과 스캔 이미지 버전, 같은 영상을 다른 해상도로 인코딩한 파일, 약간 다른 편집본 영상과의 관계 등을 ‘중복’ 또는 ‘유사’로 표시해 학습에 활용할 수 있다. 반대로, 표면적으로는 비슷하지만 내용은 전혀 다른 자료를 ‘비유사’로 라벨링해 주면, 모델이 단순히 겉모습만 보고 판단하는 것이 아니라 내용의 의미적 유사성을 더 정확하게 구분하는 법을 배운다.

알고리즘 측면에서는 텍스트·이미지·음성을 각각 벡터로 변환하는 임베딩 모델을 활용한다. 텍스트는 BERT계열 모델, 이미지는 CLIP 같은 비전·언어 융합 모델, 음성은 전용 임베딩 모델을 통해 표현할 수 있다. 영상의 경우에는 프레임 단위로 특징을 추출한 뒤, 시퀀스 전체의 유사성을 계산하는 방식을 사용한다. 이렇게 임베딩 된 벡터들 사이의 거리를 코사인 유사도나 K-means 클러스터링으로 비교하면, 이 벡터와 어느 정도 이상 비슷한 것들을 같은 그룹으로 묶는다와 같은 작업이 가능해진다.

실제 서비스에서는, 먼저 전체 컬렉션 또는 신규로 들어오는 자료에 대해 임베딩을 계산해 저장한다. 그런 다음 유사도 기준(예: 0.9 이상)을 정하고 이 기준을 넘는 자료들끼리 중복·유사 후보 세트를 도출한다. 시스템은 이 후보 그룹을 리스트로 보여주고, 담당자는 각 그룹을 검토해 완전 중복이므로 하나만 남기고 나머지는 정리할지, 서로 다른 버전이므로 관계를 메타데이터로 연결해 둘지 등을 결정한다. 최종 결정 내용은 보존 메타데이터에 반영되어, 어느 파일이 원본이고 어느 파일이 파생본인지, 어떤 자료들끼리 서로 관련되어 있는지가 명확하게 기록된다.

주의해야 할 점도 많다. 겉으로 보기에는 비슷하지만, 실제 의미가 다른 자료를 AI가 중복으로 잘못 판단할 수 있다. 예를 들어 같은 템플릿을 사용한 다른 보고서들이 표면상 비슷하게 보이는 경우가 그렇다. 반대로, 포맷이나 인코딩 방식이

달라서 유사도가 낮게 계산되든지, 임계치 설정이 적절치 않아 실제 중복임에도 탐지하지 못하는 경우도 있을 수 있다. 무엇보다 중복 정리는 보존 정책과 직결되는 문제이므로, AI가 자동으로 삭제까지 하는 일은 피해야 하며, 최종 결정은 반드시 사람이 내리는 구조로 설계하는 것이 중요하다.

❑ 포맷 식별 자동화

지금까지 파일 포맷을 식별하는 데에는 JHOVE, DROID 같은 시그니처 기반 정적 도구가 널리 사용되어 왔다. 이 도구들은 파일 헤더에 기록된 고정된 특징(시그니처)을 읽어 어떤 포맷인지 판단한다. 하지만 실제 현장에서는 확장자만 보고 포맷을 판단하는 실무 관행, 오래된 포맷에 대한 인식 부족, 손상된 파일에서 시그니처를 제대로 읽지 못하는 문제, 영상・음원처럼 컨테이너와 코덱 구조가 복잡한 경우의 식별 오류 등 여러 한계가 드러나고 있다. AI 기반 포맷 식별은 이런 한계를 보완하여, 파일의 겉모양이 아니라 속까지 들여다보고 포맷과 버전・코덱・손상 여부까지 함께 판단하려는 시도다.

※ 파일 시그니처 : 소프트웨어 파일들은 각각 고유한 포맷(JPEG, GIF, DOC, PPT 등)을 가지고 있으며 이를 파일 시그니처 (File Signature)라고 함
※ JHOVE(JSTOR/Harvard Object Validation Environment) : 형식 검증을 위한 확장 가능한 프레임워크를 개발하기 위한 JSTOR와 Harvard University Library의 공동 프로젝트에서 Java로 작성된 형식별 디지털 개체 유효성 검사 API 개발
※ DROID(Digital Record Object Identification): PRONOM에서 제공되는 포맷 정보를 이용하여 파일의 포맷을 식별하는 별도의 소프트웨어

AI 기반 포맷 식별은 파일 헤더 정보, 바이트 패턴, 메타데이터, 구조적 특징 등을 종합적으로 학습해 기존 시그니처 기반 정적 도구보다 훨씬 정밀하게 포맷을 판단한다. 예를 들면, 확장자는 PDF지만 실제 내부는 JPEG 이미지 세트일 때 포맷 오류를 제시하고 HWP97 포맷인지 또는 HWP2005인지 여부를 자동 판정하여 적합한 파일 포맷의 버전을 제시하고 영상 AVI 파일인데 내부 코덱은 MPEG-4 인지 여부를 판정할 수 있다. 이처럼 AI는 파일의 겉모양이 아니라 내부 구조의 통계적 패턴을 분석하므로 고정된 규칙보다 정확도가 더 높다.

이를 위해 먼저 다양한 파일 포맷 샘플을 수집한다. 예를 들어 PDF/A와 일반 PDF(1.5~2.0), HWP 97/2002/2005, DOC・DOCX, TIFF・JPEG・PNG・HEIF, MP3・WAV・FLAC, AVI・MKV・MP4 등 광범위한 포맷과 버전을 포함한다. 여기에 JHOVE・DROID가 이미 식별해 둔 결과 로그, 파일 헤더와 바이트 패턴 데이터, 손상 파일과 정상 파일의 시그니처 쌍, 암호화・DRM이 걸린 파일까지

함께 모아 학습 데이터로 사용한다. 이렇게 하면 AI는 어떤 바이트 패턴이 어떤 포맷과 버전에 해당하는지를 통계적으로 파악할 수 있고, 시그니처가 일부 손상된 경우에도 근사값을 제시할 수 있다.

알고리즘 측면에서는 파일의 바이트 스트림을 그대로 입력으로 받아 패턴을 학습하는 Byte-level BERT/Transformer 나, 파일 헤더의 구조를 분석하는 CNN 모델을 활용한다. 기존 시그니처 기반 규칙과 AI 모델의 판단을 결합해, 서로의 약점을 보완하는 혼합 모델을 구성할 수도 있다. 포맷과 코덱의 관계(예: 컨테이너는 AVI인데 코덱은 MPEG-4)를 추론하는 모델은 내부 스트림을 분석하여 세부 정보를 구분해 낸다. 또한 Autoencoder 기반 이상 탐지 모델은 정상적인 포맷 구조에서 벗어나는 패턴을 통해 손상되었거나 비정상적인 파일을 찾아내는 역할을 한다.

서비스 절차는 다음과 같이 이루어진다. 파일이 업로드되거나 대량 이관될 때, AI는 먼저 확장자와 헤더를 읽고, 이어 바이트 스트림의 구조를 분석해 포맷・코덱・버전을 자동으로 판정한다. 이 정보는 장기보존 표준 포맷 목록(PDF/A, TIFF, FFV1 등)과 비교되어, 보존에 적합한지 여부를 평가하는 데 사용된다. 위험도가 높은 포맷(Deprecated・Legacy・Unknown 등)은 별도로 표시되고, 담당자는 이를 검토해 보존 정책에 따라 적절한 변환・보존전략을 세울 수 있다.

그러나 AI 기반 포맷 식별도 몇 가지 제약을 가진다. 첫째, 파일 포맷 생태계는 계속 변화하고 새로운 포맷과 버전이 등장하기 때문에, 모델을 주기적으로 재학습하고 포맷 사전을 업데이트해야 한다. 둘째, DRM이 걸려 있거나 암호화된 파일은 내부 구조를 직접 분석하기 어렵기 때문에, DRM/암호를 해제한 뒤 별도의 처리 절차를 마련해야 한다. 셋째, 포맷 식별 결과는 보존 정책과 직접 연결되므로, AI에 의존하기보다는 기존 도구와의 교차 검증, 관리자 검토를 병행하는 것이 안전하다.

❑ 파일 진단

디지털자원이 제대로 보존되고 있는지 확인하려면, 실제로 파일을 열어 보고 깨진 화면이 나오지는 않는지, 소리가 끊기지는 않는지, 특정 페이지가 열리지 않지는 않는지 하나하나 확인해야 한다. 과거에는 사람이 해당 포맷을 열 수 있는 프로그램으로 직접 실행해 보고 눈과 귀로 확인하는 방식에 의존했지만, 자료가 수십만・수백만 건으로 늘어난 현재에는 사실상 불가능한 작업이 되었다. AI 기반 파일 진단은 이런 문제를 해결하기 위해 도입되는 기술로, 파일 내부의 손상, 인코딩 오류, 프레임 누락, 깨진 이미지, 오디오 노이즈, 비트스트림 오류 등을 자동으로 감지하고 미리 경고를 제공한다.

이 모델을 학습시키기 위해서는 먼저 다양한 형태의 손상 사례 데이터가 필요하다. 깨진 이미지와 정상 이미지, 드롭 프레임이나 코덱 오류가 있는 비디오와 정상 비디오, 노이즈가 심한 음원과 깨끗한 음원, PDF 내부 이미지나 폰트가 손상된 사례, HWP·DOC에서 특정 섹션이 읽히지 않는 사례 등이다. 여기에 파일 시스템 오류 로그까지 결합하면, 어떤 로그 패턴이 실제 파일 손상과 연결되는지를 AI가 배울 수 있다.

알고리즘 측면에서 파일 진단은 파일 구조 자체를 분석하는 바이트 기반 딥러닝 모델과, 영상·이미지·오디오 각각에 특화된 이상 탐지 모델이 함께 사용된다. 영상·이미지는 프레임별로 특징을 뽑아 어디서부터 이상한 패턴이 나타나는지를 탐지하고 오디오의 경우 스펙트로그램을 분석해 특정 주파수 대역이 끊겨 있거나 비정상적인 노이즈가 있는지 확인한다. 문서 포맷의 경우에는 LayoutLM 같은 모델이 문서 구조를 분석해 특정 페이지나 객체가 읽히지 않는 패턴을 찾아낸다. HWP·PDF처럼 구조가 복잡한 포맷은 전용 파서와 AI 모델을 함께 사용해, 규칙 기반 분석과 학습 기반 분석을 결합한다.

서비스 절차는 다음과 같다. 파일을 수집하거나 정기 점검 시점에, 시스템이 전체 파일을 한 번 스캔하면서 각 파일의 건강도를 평가한다. 이때 AI는 손상이 의심되는 구역을 구체적으로 표시해 줄 수 있다. 예를 들어 “PDF 내부 이미지 중 20% 이상이 깨져 있음”, “HWP 문서의 특정 섹션이 읽히지 않아 로딩 실패 가능성 있음”, “MP4의 특정 구간에서 프레임 건너뛰기가 자주 발생”과 같이 구체적 경고를 제공할 수 있다. 각 파일에는 정상/경미한 손상/심각한 손상과 같은 위험 등급이 부여되고, 심각한 손상으로 표시된 파일에 대해서는 복원 모델을 연계해 복원 후보를 제안할 수도 있다. 모든 진단 결과와 조치 내역은 PREMIS(Preservation Metadata: Implementation Strategies) 이벤트로 기록되어, 나중에 감사·점검 시에 활용할 수 있다.

AI 기반 파일 진단을 통한 사전 경고 예시
(문서)PDF 내부 이미지 중 20%가 깨져 있음
(문서)HWP 문서의 특정 섹션이 읽히지 않아 로딩 실패 가능성 있음
(영상)MP4 파일에서 프레임 건너뛰기(skip)가 자주 발생
(이미지)JPEG 파일에서 압축률이 비정상적으로 높아지며 손상 발생
(음원)WAV 파일 주파수 대역이 끊겨 있음

다만, AI가 손상되었다고 판단한 것이 실제로 의미 있는 손상인지, 또는 단지 포맷 특성이나 인코딩 방식 차이를 과민하게 반응한 것인지 구분할 필요가 있다.

따라서 심각한 손상으로 표시된 파일은 반드시 사람이 직접 확인하고 복원 여부를 최종적으로 결정해야 한다. 복원 과정에서 원본성을 훼손할 우려가 있기 때문에, 원본 손상본을 그대로 보존하면서 복원본을 별도로 저장하는 이중 구조가 필요하다. 특히 전자문서처럼 법적 효력이 중요한 자료에 대해서는, 어떤 진단과 복원 조치를 수행했는지에 대한 상세한 로그를 남기는 것이 필수적이다.

3.3.4. 디지털자원 복원을 위한 인공지능 서비스 제안

AI 기반 디지털자원 복원은 훼손되거나 저품질로 남아 있는 디지털자원을 AI 모델이 자동 보정・보완・재구성하여 서비스・보존에 적합한 고품질 상태로 만드는 기술이다. 디지털자원 복원의 목적은 원본을 복제하는 것이 아니라 원본을 훼손하지 않으면서 이용과 보존을 위한 대체 디지털본을 확보하고자 하는 것이다.

❑ 문서 복원

문서 복원 서비스는 낡고 훼손된 문서, 희미해진 글자, 번짐・오염・낙장・잉크 바램 등이 있는 스캔본을 대상으로 AI가 문자와 레이아웃을 복원해 보존용・서비스용으로 활용 가능한 상태로 되살리는 기술이다. 특히 역사적 가치가 있는 고문서의 경우, 원본을 직접 열람하기 어려운 상황에서 디지털 복원본이 이용・연구의 핵심 수단이 되기 때문에, AI 기반 문서 복원은 문화유산 보존에도 중요한 역할을 한다.

이를 위해서는 다양한 유형의 스캔 이미지와 그에 대응하는 정답 데이터가 필요하다. 희미한 글자를 또렷하게 만든 사례, 번짐・낙장・잉크 바램이 있는 문서와 복원본의 쌍, 한자・세로쓰기・필사본 같은 고문서 이미지, 표・문단・이미지・각주・제목 등 문서 구성 요소가 라벨링 된 레이아웃 데이터, 그리고 OCR 정답 텍스트 등이 대표적인 학습 데이터다. 이런 데이터들을 바탕으로 AI는 “어떤 모양의 얼룩이 글자이고, 어떤 부분이 오염인지”, “문서가 어떻게 구성되어 있는지”를 학습하게 된다.

알고리즘 측면에서 문서 복원에는 여러 모델이 단계적으로 활용된다. 먼저 Diffusion/Enhancement 모델이 전체적으로 흐릿한 이미지를 선명하게 만들고 노이즈를 제거한다. Vision Transformer 기반 문자 생성 모델은 잉크가 바랜 글자를 추정・복원하는 데 사용된다. 문서 인페인팅 모델은 문서의 일부가 찢어지거나 떨어져 나간 부분에 대해 주변 문맥과 레이아웃을 참고하여 합리적인 형태를 채워 넣는다. LayoutLMv3, Donut, DiT 같은 레이아웃 인식 모델은 표, 본문, 제목, 각주 등 문서

구조를 복원하는 데 중요한 역할을 한다. 마지막으로 OCR 모델이 복원된 이미지에서 글자를 읽어내고, 후처리 언어모델이 문장 구조와 맞춤법을 다듬는다.

서비스 절차를 보면, 먼저 스캔 이미지를 업로드하면 시스템이 해상도・왜곡・바림 상태를 분석한다. 이후 복원 단계에서 선명화, 잉크 번짐 제거, 결손 부분 인페인팅, 레이아웃 구조 복원 작업이 순차적으로 이루어진다. 복원이 완료되면 OCR을 통해 텍스트를 생성하고 언어모델을 통해 문장과 문맥을 보정한다. 담당자는 원본과 복원본, 전사 텍스트를 비교해 보면서 복원 수준을 검토하고 승인 여부를 결정한다. 보존 측면에서는 원본 이미지, 복원된 이미지, 텍스트 전사본 세 가지를 모두 보존해, 원본성・사용성・가독성을 균형 있게 확보하는 것이 중요하다.

다만, 문서 복원에는 민감한 문제도 있다. AI가 추정한 글자가 실제 역사적 사실과 다를 수 있고, 고문서의 경우 오・복원은 해석과 연구에 큰 영향을 미칠 수 있다. 또한 색 보정과 잉크 복원 과정에서 원본의 질감이나 흔적이 사라질 수 있어, 얼마나 복원할 것인지에 대한 기준을 정해야 한다. OCR 정확도가 낮으면 전체 복원 결과의 신뢰도도 떨어지기 때문에, 고문서 전용 OCR과 후처리 모델을 별도로 준비하는 것이 바람직하다.

❑ 이미지 복원

이미지 복원 서비스는 오래된 사진, 필름 스캔 이미지, 저해상도・저화질 이미지를 대상으로 해상도를 높이고 노이즈를 줄이며 색상을 복원해 보다 선명하고 보기 좋은 상태로 만드는 기술이다. 특히 역사 사진이나 필름 사진처럼 시간이 지나면서 색이 바래고 스크래치가 생긴 자료를 복원하면, 과거의 장면을 오늘날의 이용자에게 더 생생하게 전달할 수 있다.

이를 위해서는 고해상도와 저해상도 이미지의 쌍, 손상본(스크래치・노이즈・빛바램 등)과 정상본의 쌍, 다양한 유형의 역사 사진・신문 삽화・지도・문서 스캔 이미지, 그리고 컬러/흑백 매핑 데이터 등 풍부한 학습 데이터가 필요하다. 이러한 데이터를 통해 AI는 저해상도 이미지를 어떻게 고해상도로 재구성할지, 노이즈와 실제 이미지 내용을 어떻게 구분할지, 흑백 사진을 어떤 색감으로 복원할지를 점점 더 잘 배우게 된다.

알고리즘으로는 Super-resolution 모델(ESRGAN, Real-ESRGAN, SR3 등)이 가장 핵심적인 역할을 한다. 이 모델들은 저해상도 이미지를 입력으로 받아, 디테일이 살아 있는 고해상도 이미지를 생성한다. DnCNN, Noise2Noise, Diffusion 기반

모델은 이미지에서 노이즈를 제거해 보다 깨끗한 이미지를 만들어 준다. U-Net, LaMa, Stable Diffusion 기반 인페인팅 모델은 긁힌 부분이나 일부가 지워진 영역을 주변 정보를 바탕으로 자연스럽게 채워 넣는다. DeOldify나 Colorization GAN 같은 모델은 흑백 사진에 자연스러운 색을 입히는 데 사용된다. 필요한 경우 OCR 전에 이미지 품질을 높이기 위해 Vision Transformer 기반 정제 모델을 함께 사용할 수도 있다.

서비스 절차는, 먼저 복원 대상 이미지를 수집하고 해상도・노이즈・손상 유형을 분석하는 것에서 시작한다. 그 다음, 선명화 → 노이즈 제거 → 색상 복원 → 인페인팅 순으로 복원 파이프라인을 적용한다. 복원 결과는 원본과 나란히 비교할 수 있는 인터페이스에서 확인하며, 담당자가 복원의 수준이 적절한지를 판단해 서비스용 복원본으로 승인한다. 보존 측면에서는 원본 이미지는 원본 그대로의 상태로 AIP(Archival Information Package)에 보존하고 복원본은 이용과 서비스 중심으로 관리하는 이원화 전략이 필요하다. 복원 과정과 사용된 모델・파라미터는 PREMIS(Preservation Metadata: Implementation Strategies) 이벤트로 기록해, 나중에 복원 내역을 추적할 수 있도록 한다.

주의해야 할 점은 색상과 디테일이 어디까지나 AI가 추정한 값이라는 점이다. 실제 역사적 색감과 다를 수 있고, 지나치게 선명하게 만들면 당시 촬영 환경과 질감이 왜곡될 수 있다. 특히 인물 사진에서는 얼굴을 과도하게 재구성하는 경우, 원래 존재하지 않던 디테일이 추가되면서 윤리적 논란이 생길 수 있다. 저작권・초상권과 관련해서도, 복원 과정 자체가 2차 저작물에 해당할 수 있으므로, 법적 측면과 이용자 안내도 함께 검토해야 한다.

❑ 영상 복원

영상 복원 서비스는 오래된 필름, 저화질 동영상, 노이즈나 스크래치가 많이 있는 기록 영상 등을 대상으로 해상도를 높이고, 프레임을 보완하며, 색상을 보정하고 흔들림을 줄여 보다 안정적인 시청 경험을 제공하는 기술이다. 이는 단순한 화질 개선을 넘어, 역사적 영상 자료를 후대에 전달하는 보존전략의 일부로도 중요한 의미를 가진다.

학습 데이터로는 저해상도(LQ)와 고해상도(HQ) 영상의 쌍, 노이즈・스크래치・흔들림이 포함된 필름 영상과 복원본, 프레임 누락이나 끊김이 있는 영상과 정상 영상, 장면별 색보정 정보, 8mm・16mm・35mm 등 다양한 필름 스캔본이 활용

된다. 이런 데이터는 AI가 손상된 영상이 어떤 특징을 보이는지와 좋은 품질의 영상이 어떤 상태인지를 동시에 학습할 수 있게 해준다.

알고리즘 측면에서 EDVR・BasicVSR・Video SR Transformer와 같은 Super-resolution 모델은 영상의 해상도를 높이는 역할을 한다. DAIN・RIFE 같은 프레임 보간 모델은 두 프레임 사이의 움직임을 추정해 중간 프레임을 생성, 끊김 없이 부드러운 화면 전환을 만들어 준다. 노이즈・스크래치 제거는 CNN 기반 디노이징 모델이나 Diffusion 기반 복원 모델이 담당하며, 흔들림 보정은 Optical-flow를 활용해 카메라 흔들림을 완화한다. 색보정 모델은 장면별 색 균형을 맞추어 전체 영상의 색감을 자연스럽게 복원한다. Scene Detection 모델은 영상 속에서 장면 전환 시점을 자동으로 인식해, 복원 단위를 효율적으로 나누는 데 도움이 된다.

서비스 절차는, 먼저 필름이나 영상 파일을 디지털화하거나 수집하는 단계에서 시작된다. 이후 프레임 단위 분석을 통해 노이즈와 흐림, 스크래치, 프레임 누락 정도를 파악한 뒤, 노이즈 제거, 스크래치 제거, 프레임 보간(예: 24fps를 60fps로), Super Resolution, 색 보정 등의 단계를 순차적으로 적용한다. 복원된 영상은 원본과 나란히 비교하여 품질과 자연스러움을 점검하고 담당자가 승인한 후 서비스용 복원본으로 등록한다. 복원 처리 과정과 결과는 PREMIS(Preservation Metadata: Implementation Strategies) 이벤트로 기록해, 어떤 조치가 언제, 어떤 설정으로 이루어졌는지 추적할 수 있게 한다.

영상 복원에서도 주의할 점이 많다. AI가 보정한 결과는 원본의 역사적 정보나 분위기를 바꿔 버릴 수 있다. 프레임 보간 과정에서 존재하지 않던 움직임이 만들어지는 등 왜곡이 발생하기도 한다. 또한 고해상도 영상 처리는 GPU 자원을 많이 소비하기 때문에, 대규모 작업을 위해서는 배치 처리 시스템과 자원 관리 전략이 필수적이다. 영상 속에 등장하는 인물의 초상권・저작권 문제도 고려해야 하며, 복원본과 원본을 명확히 구분해 이용자에게 안내하는 것이 중요하다.

❑ 음원 복원

음원 복원 서비스는 노이즈가 많은 테이프 녹음, 현장 녹음, 구술・증언 음원 등을 대상으로 잡음을 줄이고 음성・음악을 보다 명확하게 들을 수 있도록 만드는 기술이다. 다만 목표는 CD 음반처럼 완벽하게 깨끗한 소리를 만드는 것이 아니라 원본의 중요한 정보와 특성을 유지하면서도 이해 가능한 수준으로 개선된 음원을

제공하는 데 있다.

이를 위해서는 노이즈가 포함된 음원과 노이즈 제거 후의 정답본 쌍, 테이프 잡음(hiss, hum, pop, crackle 등) 사례, 다양한 환경에서 녹음된 음성・음악・현장 녹음 데이터, 다화자・마스킹 음성 등 여러 조건이 섞인 음원이 필요하다. 이런 학습 데이터를 바탕으로 AI는 어떤 주파수대가 잡음이고, 어떤 부분이 유의미한 음성・음악인지 구분하는 법을 배우게 된다.

알고리즘 측면에서 RNNoise, Demucs 같은 노이즈 제거 모델은 스펙트럼을 분석해 불필요한 성분을 줄여준다. Speech Enhancement Transformer와 같은 모델은 특히 사람 목소리를 또렷하게 하는 데 초점을 맞춘다. Audio Inpainting GAN은 특정 구간의 음원이 끊어졌거나 손상되었을 때 주변 정보를 바탕으로 자연스럽게 채워 넣는다. Harmonic Reconstruction 모델은 손실된 고주파 성분을 보완해 음질을 향상시키고, STT 전처리용 향상 모델은 음원을 STT에 넣기 전에 인식하기 좋은 형태로 정제하는 역할을 한다.

서비스 절차는, 먼저 테이프나 녹취 자료를 디지털화한 뒤 스펙트럼과 잡음 유형을 분석하는 것에서 시작한다. 분석 결과를 바탕으로 잡음 유형별로 적합한 복원 모델을 선택해 노이즈 제거 → 음성 강화 → 주파수 보정 순으로 처리한다. 필요할 경우 STT를 적용해 전사(텍스트)를 함께 생성하고 이를 통해 실제 내용 전달력이 얼마나 향상되었는지 평가할 수 있다. 복원된 음원은 이용자 평가나 담당자 검토를 거쳐 서비스용으로 승인하며, 원본 파일과 복원본, 해당 처리 과정과 파라미터는 PREMIS(Preservation Metadata: Implementation Strategies)에 기록해 둔다.

그러나 음원 복원에서도 얼마나 복원할 것인가에 대한 균형점이 중요하다. 노이즈를 너무 강하게 제거하면 사람 목소리나 악기 소리까지 함께 깎여 나가 음질이 부자연스러워질 수 있다. 특히 역사적인 현장 녹음의 경우, 약간의 잡음과 환경 소리 자체가 당시 상황을 보여주는 중요한 맥락이 될 수 있으므로, 이를 모두 제거하는 것이 항상 옳지는 않다. 음성 속에는 개인의 이름, 발언 내용 등 민감한 정보가 포함되어 있을 가능성이 크기 때문에, 복원과 전사 이후에는 반드시 개인정보 보호 관점에서의 검토도 병행해야 한다.

제4장 메타데이터 영역 기술동향 분석 및 설계

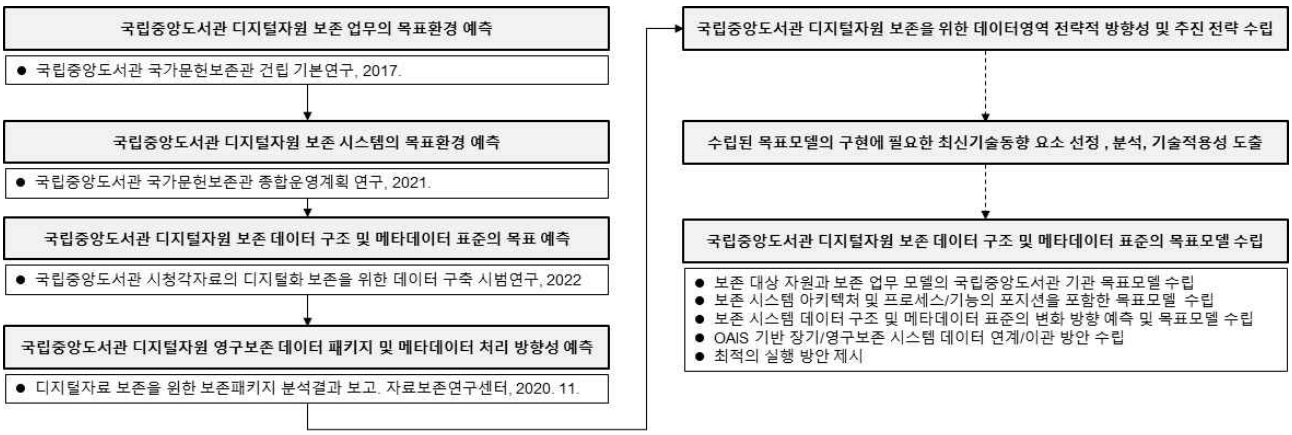
4.1. 현황분석

4.1.1. 선행 연구 조사 분석

선행 연구 조사분석의 프레임워크는 분석 대상인 4개 연구의 수행 시기적 순서에 따르지 않고, 정보화 시스템 분석 절차 관점에서, 먼저 국립중앙도서관의 보존 업무와 시스템 분석 연구 결과를 분석하고, 디지털자원 보존 영역의 데이터 연구 결과를 분석하는 순서로 진행하였다. 그 이유는 데이터 부분의 연구가 디지털자원의 일부 현상에 대한 한정적 연구 및 시범사업 결과를 보고한 내용으로서 업무 및 시스템 부분에 영향을 미치지 않았기 때문이다.

이런 방식을 사용하여, 사용과 보존이 공존하는 국립중앙도서관의 업무와 시스템 목표모델을 예측하고, 여기에 적합한 추진전략과 관련 최신기술동향분석의 방향성과 범위를 설정하여 신뢰성 있는 데이터 목표모델을 수립하고, 최적의 실행방안을 제시하도록 하였다.

다음의 프레임워크 그림에서 각 연구의 분석 목표와 대상 과제를 순서대로 제시하고, 분석 결과가 이후 단계의 업무에 적용되는 과정을 설명하였다.



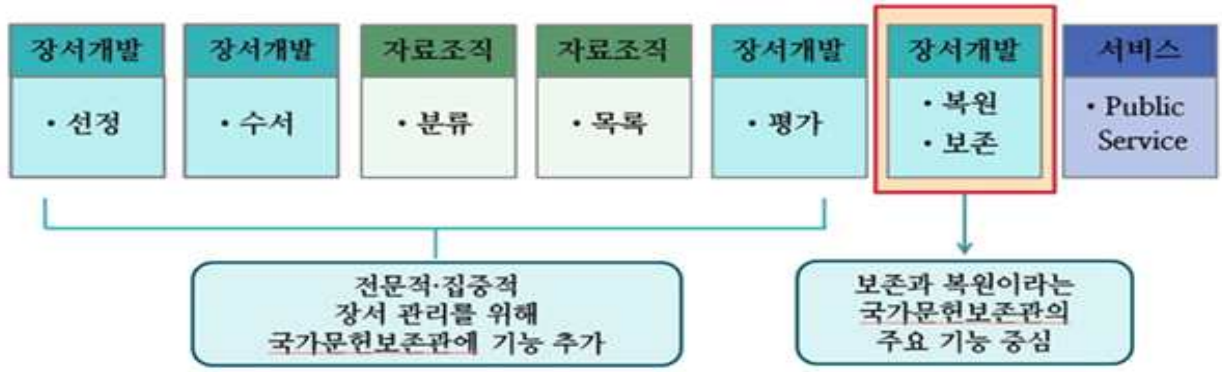
[그림 25] 선행 연구 조사 분석

❑ 국립중앙도서관 국가문헌보존관 건립 기본연구, 2017.

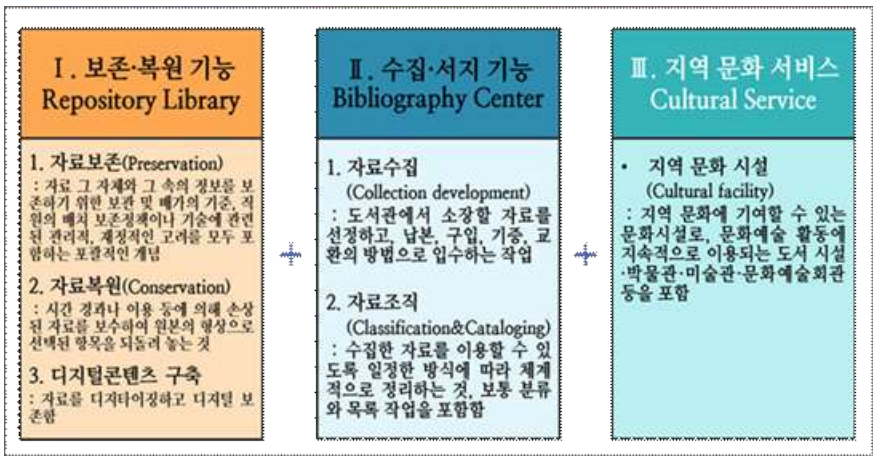
「국립중앙도서관 국가문헌보존관 건립 기본연구(2017)」는 국립중앙도서관이 지향하는 ‘국가 지식정보자원 수집·보존 강화’라는 정책 목표를 실현하기 위한 국가문헌보존관 건립의 기본 계획을 수립하는 것을 목적으로 한다. 이를 위해 본

연구는 국가문헌보존관의 물리적·기능적 구축 방향을 종합적으로 검토하며, 보존관의 역할과 기능, 운영 모델을 체계적으로 제시하는 데 연구 범위를 둔다.

연구의 세부 목표는 첫째, 국가문헌보존관 건립을 위한 수장공간 확보 방안을 제시하는 데 있으며, 둘째, 국가문헌보존관의 건립 방향성과 중장기적 비전을 설정하는 데 있다. 셋째, 보존관이 수행해야 할 핵심 기능을 정의하고 넷째, 이러한 기능을 안정적으로 수행하기 위한 운영의 기본 모델을 제시하는 것을 주요 과제로 설정한다.



[그림 26] 도서관 자료 생명주기 상의 국가문헌보존관의 위치 설정 모델링



[그림 27] 국가문헌보존관의 기능 구성 모델링

관련 연구 결과를 종합하면, 국가문헌보존관은 국립중앙도서관을 비롯한 각급 도서관에서 상대적으로 이용 빈도가 낮으나 장기적 보존 가치가 있는 도서 및 기타 인쇄물을 체계적으로 보존하고 이를 기반으로 한 서비스 기능을 수행하는 기관으로 정의된다. 더 나아가 국가문헌보존관은 Born digital 자원, Original analog 자원, Digitized analog 자원을 모두 포괄하는 디지털 콘텐츠를 대상으로 생산, 관리, 저장, 보존 및 복원 업무를 수행함으로써, 디지털 아카이브로서의 역할

또한 담당하는 것으로 설정된다.

국가문헌보존관의 관리 대상 자원은 국립중앙도서관과 지역 도서관이 소장하고 있는 자원 전반을 포괄하며, 이들 소장 자원 전체가 국가문헌보존관의 입수 자원 유형과 범위에 포함될 수 있는 것으로 제시된다. 이는 국가문헌보존관이 단일 기관의 보존 기능을 넘어 국가 차원의 문헌 보존 인프라로 기능함을 전제로 한다.

기능적 측면에서 국가문헌보존관의 최우선 목적은 문헌의 보존과 복원에 있으며, 국가장서의 안정적인 유지와 보존자료의 효율적 관리를 위해 자료의 재등록 및 재분류, 디지털화 자료에 대한 메타데이터 구축을 포함한 자료조직 업무 수행 기능을 갖는 것으로 설정된다. 아울러 국가적 차원의 자료 보존 및 복원 처리 업무를 통해 자료 훼손을 지연시키고, 자료 생애주기를 확장하며, 도서 및 비도서 자료의 원형 보존을 위한 보존·복원 처리, 매체 변환, 디지털 보존, 디지털 아카이빙을 보존복원 기능의 핵심 요소로 규정한다.

한편, 국가문헌보존관의 자료 이관 및 폐기 기준의 표준화, 자료보존정책과 지침의 수립 등 법·제도적 장치에 관한 사항은 본 연구의 범위에 포함되지 않는 것으로 명시된다. 이는 본 연구가 제도 설계보다는 건립 기본계획과 기능·운영 모델 제시에 중점을 두고 있음을 의미한다.

❑ 국립중앙도서관 국가문헌보존관 종합운영계획 연구, 2021.

「국립중앙도서관 국가문헌보존관 종합운영계획 연구(2021)」에서는 국가문헌보존관의 합리적이고 체계적인 운영계획을 수립하는 것을 주요 목표로 한다. 이를 위해 본 연구는 국가문헌보존관의 운영목표와 운영원칙, 운영방향을 검증하고 이를 토대로 세부 운영계획, 조직 및 인력 구성, 공간 배치, 시스템 구성 등을 설계·제시하는 데 연구의 범위를 둔다.

연구 결과에 따르면, 국가문헌보존관의 목표 시스템은 세 가지 핵심 요소로 구성되는 것으로 제시된다. 첫째, 보존관 자료의 통합적 보존을 수행하기 위한 통합보존관리시스템, 둘째, 축적된 국가자료를 분석·활용하기 위한 데이터 분석 시스템, 셋째, 국가문헌 및 디지털자료의 보존과 데이터 활용을 지원하는 온라인 플랫폼이다. 이 가운데 통합보존관리시스템은 다시 도서 및 비도서 자료(광매체, 전자매체)를 대상으로 한 자료보존 업무 프로세스 기반의 자료 보존 시스템과, 온라인자료 및 디지털화 자료를 대상으로 한 디지털 보존 업무 프로세스 기반의 디지털 보존 시스템으로 구성되는 것으로 설정된다.



[그림 28] 국가문헌보존관 목표 시스템 구성도



[그림 29] 국가문헌보존관 목표 통합보존관리 시스템 구성도



[그림 30] 국가문헌보존관 목표 온라인 플랫폼 개념도

자료 생명주기 기반의 보존업무 포지셔닝 요건 분석 결과, 목표 시스템 구성의 구체성 측면에서 여러 한계가 확인된다. 먼저 목표 시스템 구성도 상 통합보존관리시스템 영역에서 자료보존연구센터가 차지하는 위치와 담당 업무 및 기능을 명확히 식별할 수 있는 수준의 구체화가 필요하다는 점이 지적된다. 또한 목표 통합보존관리시스템 내에서 자료보존연구센터가 담당해야 할 프로세스와 기능 역시 보다 명확하게 정의될 필요가 있는 것으로 분석된다.

특히 디지털 보존 프로세스, 즉 온라인자료와 디지털화 자료 전반을 자료보존연구센터의 기능 영역으로 설정할 수 있는지에 대한 검토가 필요함에도 불구하고

디지털 콘텐츠 구축, 디지털 복원, 디지털 장기보존 등 전체 업무 중에서 자료보존연구센터가 실제로 운영・관리・서비스할 수 있는 자원과 서비스의 범위를 식별할 수 있는 정보는 제시되지 않은 것으로 나타난다. 아울러 온라인 플랫폼의 자료보존 중심 구조에서도 민원 서비스 영역에 포함된 자료보존・복원 서비스 외에, 정보 서비스 영역에서 자료보존연구센터가 수행할 수 있는 데이터 기반 서비스 영역은 확인되지 않는 한계가 존재한다.

목표 데이터 아키텍처 측면에서도 한계가 도출된다. 목표 시스템 구성을 위해 요구되는 데이터 구조와 메타데이터 표준에 관한 구체적인 사항이 제시되지 않았으며, 현업 시스템과 보존 시스템 간의 데이터 전환 방식이나 데이터 구조 간의 상관관계에 대한 설계 역시 확인되지 않는다. 이는 통합보존관리시스템의 실질적 구현과 운영을 위한 데이터 기반 설계가 충분히 제시되지 않았음을 의미한다.

목표 시스템 아키텍처와 관련하여서는, 보존과 복원 활동이 시스템 기능 영역에 포함되어 있기는 하나, 기본적으로 OAIS(Open Archival Information System) 참조 모델의 구조를 채택하고 있지 않은 것으로 분석된다. 또한 운영 데이터 구조 역시 AIP(Archival Information Package) 정보 패키지 단위의 운영이 어려운 전형적인 도서관 정보관리 시스템의 데이터 구조를 채택하고 있을 가능성이 높은 것으로 판단된다. 이러한 구조는 장기보존 관점에서의 체계적 관리와 패키지 단위 보존 운영에 제약 요인으로 작용할 수 있음을 시사한다.

❑ 국립중앙도서관 시청각자료의 디지털화 보존을 위한 데이터 구축 시범연구, 2022.

「국립중앙도서관 시청각자료의 디지털화 보존을 위한 데이터 구축 시범연구(2022)」에서는 국립중앙도서관 자료보존연구센터에서 수행하는 시청각자료 디지털화 파일을 체계적이고 효율적으로 보존하기 위한 보존 메타데이터 패키지 운영 방안을 제시하는 것을 목적으로 한다. 연구의 범위는 VHS, LP, CT에 수록된 시청각자료를 대상으로 하며, 해당 자료에 대해 MARC 데이터와 MODS(Metadata Object Description Schema) 데이터를 구축하고 이를 KOLIS 시스템에 업로드한 후, 보존관리에 필요한 요소 값을 추출・분석함으로써 디지털화 시청각자료의 보존 메타데이터 패키지 운영 방안을 도출하는 데에 한정된다. 또한 통합서지용 KORMARC의 비도서 자료유형에 속하는 자원의 메타데이터를 MODS(Metadata

Object Description Schema)로 전환하여 디지털화 시청각자료의 보존 메타데이터 패키지 운영에 활용하는 것을 연구의 전제로 한다.

연구 내용 및 결과를 종합하면, 먼저 디지털화 자료의 탐색 및 식별 영역에서 데이터 구조상의 문제점을 확인하고 이에 대한 개선 방향을 제시하고 있다. 시청각자료의 경우 표지에 기재된 타이틀보다 곡명 등 실제 내용상의 타이틀이 중요하므로, 내용주기의 적절한 활용을 통해 검색 가능성이 확보되어야 함이 강조된다. 또한 시청각자료를 원본 자원으로 디지털화된 자료를 복제본으로 간주하여 MODS(Metadata Object Description Schema) 주기정보를 확장 적용하는 방안이 제시된다. LP와 CT 자료의 경우 수록 최소 단위인 개별 노래 단위의 검색을 가능하게 하기 위해 목차정보를 활용하고 TABLE OF CONTENTS와 각 노래의 물리적 수록 위치를 식별할 수 있어야 한다는 점이 지적된다.

아울러 원시 자원에서 디지털화 자원으로 다시 그 이후의 다양한 파생 기록물로 이어지는 기록 생성 과정을 현재의 구조로는 충분히 해결하지 못하는 상황을 메타데이터를 통해 보완하고자 하는 시도가 제시된다. 단일 기록매체에 수록된 다수의 기록물과 관련하여, MODS(Metadata Object Description Schema)로 전환하더라도 기록매체 내부에 포함된 각각의 독립된 기록물을 명확히 식별하고 기록물별 속성을 분리하여 표현하는 데에는 한계가 존재하며, 이러한 문제를 메타데이터 차원에서 해결하고자 한다는 점이 연구 결과로 나타난다. 그러나 PREMIS(Preservation Metadata: Implementation Strategies), PBCORE 등 특정 자원의 속성 처리에 유리한 다양한 메타데이터를 선택적으로 활용하는 데에는, MODS(Metadata Object Description Schema)와 기술규칙, 평면적 데이터 구조가 유지되는 상황에서는 해결이 쉽지 않을 것이라는 한계 역시 제시된다.

시청각자료의 보존 및 실행을 위해 요구되는 기술 메타데이터 운영과 관련해서는, 시청각자료 보존을 위한 형태 정보를 코드로 제시하는 007 형태기술 필드를 정확히 이해하고 활용할 필요성이 강조된다. 또한 시청각자료의 유형을 목록 이용자에게 사전에 알리는 자료유형표시가 부정확하게 사용되는 오류가 발생하지 않도록 주의해야 함이 지적된다. 책임표시, 즉 저자에 대한 기술은 자료의 유형에 따라 상이하므로, 이에 대한 명확한 규정과 일관된 적용의 필요성이 제기된다. 더불어 형태사항의 첫 번째 요소인 측정자료 종별과 수량표시 역시 정확한 사용이 요구된다.

형태기술필드 007, GMD, 006, 008 간의 논리적 유형 구분이 일치하지 않는 문제에 대해서도 연구에서는 문제를 제기하고 있으나, 이에 대한 해결책으로는

사용자의 주의를 통한 대응 방식을 제시하는 데 그치고 있다. 그러므로 관련 표준과 기술지침 자체에 내재한 불일치나 모호성에 대한 구조적 문제 제기와 근본적 해결 방안은 제시되지 않았다는 한계가 확인된다.

마지막으로 보존 메타데이터 패키지 적용 시의 고려사항으로서 자료보존에 최적화된 preservation metadata, technical metadata, descriptive metadata 체계를 구축하고 디지털화 시청각자료의 보존・관리・운영 전반과 메타데이터 레코드 수준에서 원본 매체와의 연계가 가능하도록 하는 보존 메타데이터 패키지 구성의 필요성이 제시된다. 이는 디지털화 시청각자료의 장기적 관리와 보존을 위한 핵심 전제로 정리된다.

❑ 디지털자료 보존을 위한 보존패키지 분석결과 보고, 2020.11.

「디지털자료 보존을 위한 보존패키지 분석결과 보고(자료보존연구센터, 2020.11)」에서는 디지털자료 장기보존을 위한 보존패키지의 유형과 특성을 분석하고 그 적용 가능성을 검토하는 데 목적을 둔다. 본 보고가 추진된 배경에는 다양한 유형의 디지털 객체와 이와 연계된 메타데이터를 장기적으로 보존하기 위한 통합적 보존패키지가 부재한 현실이 자리하고 있다. 그러므로 문서와 동영상 등 디지털 객체와 메타데이터를 통합하여 저장・관리할 수 있는 보존 포맷이 마련되지 않은 상황에서 장기보존 대상과 메타데이터를 선별하고 이를 기반으로 한 보존 방안 도출의 필요성이 제기되었다. 나아가 디지털자료 보존전략 수립을 위해 보존 패키지의 특성을 체계적으로 분석할 필요성이 본 분석의 출발점이 되었다.

현황분석 결과, 도서자료의 경우 MARC 데이터를 기반으로 MODS(Metadata Object Description Schema) 메타데이터를 생성하여 활용하고 있으며, 디지털자료에 대해서는 KOLIS 시스템 내부에 PREMIS(Preservation Metadata: Implementation Strategies) 메타데이터 80개 입력 항목을 설계하였으나 실제 운영 단계에서는 아직 활용되지 않고 있는 것으로 나타난다. 또한 메타데이터와 디지털자료를 함께 보존할 수 있는 보존패키지가 부재하여, 파일형 또는 폴더형 보존패키지 전략을 검토해야 하는 상황임이 확인된다.

이에 따라 본 보고에서는 보존패키지 유형을 파일형과 폴더형으로 구분하여 비교분석을 수행하였다. 파일형 보존패키지는 데이터 비트스트림을 BASE64 인코딩을 통해 문자화한 후, 메타데이터와 디지털 객체를 XML(Extensible Markup Language) 기반의 문자 데이터로 서술하여 보존하는 유형으로 정의된다. 반면 폴더형 보존

패키지는 디지털 객체와 메타데이터를 분리하여 저장하되, 폴더 구조를 통해 각 자원을 원형 그대로 관리하는 유형으로 설명된다.

자료유형별 분석 결과를 살펴보면, 먼저 문서 및 도서 유형의 경우 KOLIS 시스템에서 생성한 MODS(Metadata Object Description Schema) 데이터와 PREMIS(Preservation Metadata: Implementation Strategies) 필수 보존 데이터를 결합하고 보존 포맷인 PDF/A-1 파일과 함께 자동으로 폴더형 보존패키지를 구축하는 방식이 제시된다. 이때 보존패키지는 MODS(Metadata Object Description Schema) 데이터 58개 요소, PREMIS(Preservation Metadata: Implementation Strategies) 데이터 14개 요소, 그리고 PDF/A-1 보존 포맷 파일로 구성된다. 시청각자료유형의 경우에는 KOLIS 시스템에서 생성한 MODS(Metadata Object Description Schema) 데이터와 PREMIS(Preservation Metadata: Implementation Strategies) 필수 보존 데이터에 더하여, 시청각자료의 기술 메타데이터를 자동으로 추출함으로써 폴더형 보존패키지를 구축하는 방식이 제시된다. 이 경우 MODS(Metadata Object Description Schema) 데이터 58개, PREMIS(Preservation Metadata: Implementation Strategies) 데이터 14개, 기술 메타데이터 10~12개, 그리고 영상은 MXF, 음향은 WAV 형식의 보존 포맷 파일로 구성된다. 디지털자료유형에 대해서는 KOLIS 시스템의 MODS(Metadata Object Description Schema) 데이터와 PREMIS(Preservation Metadata: Implementation Strategies) 필수 보존 데이터를 생성하고 디지털 원본의 메타데이터 특성을 분석하여 기술 메타데이터를 추출한 후, 디지털 원본 포맷을 그대로 보존 포맷으로 활용하는 폴더형 보존패키지 구축 방안이 제시된다.

보고서 결과에 대한 종합 분석에 따르면, 자료보존연구센터가 보존패키지 개발 방향을 폴더형으로 설정한 것은 전반적으로 적절한 선택으로 평가된다. 특히 BagIT를 컨테이너 포맷으로 활용하고 PREMIS(Preservation Metadata: Implementation Strategies) 메타데이터를 사용하는 방식은 자료보존연구센터의 장기보존 포맷 구성 방식으로 타당한 것으로 분석된다. 다만 이러한 기본 방향을 전제로 하여, 몇 가지 추가적인 고려 사항이 제시된다.

첫째, 자료유형을 포괄하는 통합 메타데이터 요소 표준 관리 체계의 필요성이 지적된다. 현재와 같이 자료유형별로 개별 메타데이터 요소 표준을 지정하는 방식보다는, 최상위 개체 수준에서 정의된 속성을 각 자료유형이 상속받으면서도 유형별 세부화가 가능한 통합 구조가 보다 적절한 것으로 판단된다. 이러한 통합 체계를 통해 모든 자료유형이 공통 속성을 공유하되, TYPE OF RESOURCE나 GENRE와 같이 빈번하게 적용되는 규칙과 절차는 지침이나 편람의 형태로 분리하여 관리할

필요성이 제기된다.

둘째, 보존데이터의 서지적·기술적·구조적 정보를 외부 스키마를 활용하여 구현하는 방식의 필요성이 강조된다. PREMIS(Preservation Metadata: Implementation Strategies)에서 요구하는 포함 자원에 대한 서지 정보, 관리 정보, 보존 및 관련 이벤트 정보 등은 KOLIS Ⅲ나 OAK와 같은 외부 스키마로부터 가져와 활용하는 것이 적절한 것으로 제시된다. 또한 폴더 내부에 존재하는 자원 간 관계 정보는 METS(Metadata Encoding and Transmission Standard)나 DC 컬렉션 정보 등, 폴더 구조를 설명하는 메타데이터로부터 확보하는 방식이 제안된다. 이렇게 수집된 정보들은 SIP(Submission Information Package) 및 AIP(Archival Information Package) 단계에서 PREMIS(Preservation Metadata: Implementation Strategies)가 요구하는 OBJECT, EVENTS, AGENTS, RIGHTS 영역으로 매핑·전환하여 활용할 수 있도록 전환 계획을 수립할 필요가 있다고 분석된다.

셋째, 현재 KOLIS 장서와 디지털 콘텐츠의 연결 구조를 활용하는 방식에 대한 개선 필요성이 제시된다. 디지털자원의 구조 측면에서는 하나의 기록매체에 수록된 다양한 독립적 저작물과 이를 구성하는 요소들 간의 계층 관계와 요소 관계를 자유롭게 표현할 수 있는 다면적 관계 표현 능력이 요구된다. 또한 디지털자원의 유형 측면에서는 기존에 제시된 원문 콘텐츠, 표지, 썸네일, 보존 유형 외에도 다양한 파생 기록물 유형을 수용할 수 있어야 하며, 각 개체 간의 연결 관계를 정확하게 표현할 수 있는 구조적 표현 능력이 필요하다는 점이 강조된다.

❑ 선행연구 결과 조사 분석

연구에서는 자원의 개체 단위에 대한 유일한 식별과 구조적 관계 문제를 평면적인 메타데이터 요소와 기존 목록기술규칙을 적용하여 해결하려는 접근 방식이 근본적인 한계를 가질 수 있음을 지적한다. 하나의 자료 또는 하나의 매체가 다수의 독립적인 하위 자료로 구성되는 경우, KORMARC나 KCR 4.0과 같은 메타데이터 요소 및 목록기술규칙만으로 이를 해결하려는 방식은 완전한 구조적 표현에 제약이 따르는 것으로 분석된다. 이에 따라 자원을 구성하는 각 요소를 분리하고 이들 간의 관계를 구조적으로 연결하여 표현할 수 있는 다면적 데이터 구조의 적용을 검토할 필요성이 제기된다.

또한 형식 및 형태 정보를 중심으로 한 기술 메타데이터 운영 방식의 한계를 극복할 필요가 있음이 강조된다. 특히 KORMARC의 006, 007, 008 필드와 리더 06에 포함된 Type of Resource, Genre 등의 요소 간 관계가 모호하게 설정되어 있어, 기술 메타데이터에서 기본적인 자원 유형을 명확하게 결정하는 데 어려움이

존재하는 것으로 지적된다. 이러한 관계의 불명확성은 기술 메타데이터의 일관성과 신뢰성을 저해하는 요인으로 작용할 수 있으므로, 유형 결정의 명확성을 제고하기 위한 개선이 요구된다.

아울러 추상체인 기록, 구현체로서의 기록물, 그리고 기록물을 수록한 매체 간의 다중적 관계를 표현할 수 있는 ERD의 개발 필요성이 제시된다. 이는 추상체 수준의 기록, 구현체인 기록물, 기록물의 저장 및 운반 매체, 매체의 구동과 데이터 표현에 필요한 기술 메타데이터, 기록물의 의미적 내용과 데이터 이해를 위한 기술 메타데이터, 기록 보존에 필요한 보존 메타데이터, 그리고 다양한 파생 기록물 집합체의 구성을 모두 포괄할 수 있는 구조를 의미한다. 더 나아가 독립적 단위로의 분해가 가능하고 계층적 기록물 집합체의 구성과 해석이 가능한 구조적 메타데이터 체계로의 구성이 필요하다는 점이 강조된다.

국립중앙도서관의 메타데이터 운영 체계와 관련해서는, 책자형 자료에는 KORMARC를, 온라인 디지털자료에는 MODS(Metadata Object Description Schema)를 적용하는 현행 방식에 대해 부분적인 개선의 필요성이 제기된다. 자원 수집과 생산, 운영 환경의 다양성을 유지하는 동시에 자원 유형 구분과 메타데이터 표준 변화에 대한 대응 능력을 제고하기 위해서는, 현재 MODS(Metadata Object Description Schema)를 중심(Pivot) 표준으로 운영하는 메타데이터 관리 방식을 다중 메타데이터 표준 운영을 전제로 한 크로스워크(Crosswalk) 체계로 전환하는 방안을 검토할 필요가 있는 것으로 제시된다.

마지막으로 현행 업무 시스템과 기록물 장기보존 시스템이 공존할 수 있는 데이터 모델의 개발 필요성이 강조된다. 일반적으로 도서관 정보관리 시스템은 아이템 중심의 서지 레코드와 소장정보를 기반으로 한 평면적 구조를 채택하는 반면, 기록관리 시스템은 기록물 메타데이터, 디지털 파일, 이벤트 정보를 하나의 패키지로 구성하는 구조를 사용하는 것으로 비교된다. 이러한 두 구조의 차이를 극복하기 위해, 양자의 융합 모델을 적용하여 시스템 간 연계・연동・협업이 가능한 구조를 구현하고 현용 단계에서 보존 단계로 데이터와 기능이 자연스럽게 이행될 수 있는 방안을 모색할 필요성이 제시된다.

4.1.2. 자료보존연구센터 현황 개선방향

국립중앙도서관 자료 입수 담당자 및 자료보존연구센터 담당자 인터뷰와 현황 자료 조사 결과를 종합・분석한 결과, 메타데이터 표준 영역, 데이터 구조 영역, 그리고 보존센터 서비스 목표 설정 측면에서 제도적・구조적 개선이 요구되는 것으로 판단된다.

❑ 메타데이터 표준 영역 개선 방향

메타데이터 표준 영역과 관련하여, 현재 온라인자료의 메타데이터 표준으로 MODS(Metadata Object Description Schema)가 사용되고 있으나, 원본인 Original-analog 자료가 보유한 KORMARC 메타데이터와의 관계에 대한 검토가 필요함을 지적한다. 또한 국가서지과를 중심으로 2030년을 목표로 BIBFRAME 적용이 예정되어 있는 상황에서 현행 메타데이터 체계와의 연속성 및 전환 문제에 대한 사전 검토가 요구된다. 특히 책자형 자료에는 MARC를, 디지털 파일에는 MODS(Metadata Object Description Schema)를 적용하는 이원적 운영 과정에서 발생하는 문제점, 예를 들어 역문헌가중치 계산 결과를 활용한 랭킹 서비스 알고리즘상의 문제를 해소할 필요성이 제기된다.

아울러 MARC와 MODS(Metadata Object Description Schema) 체계에서 표현이 어려운 메타데이터 요소에 대해서는 더블린 코어(Dublin Core, DC), schema.org 등 다른 메타데이터 표준에서 해당 요소를 식별하고 이를 적용할 수 있어야 한다는 점이 강조된다. 기술 메타데이터 및 보존 포맷과 관련해서는, 메타데이터 요소의 과도한 확장이 운영상의 부담으로 작용할 수 있다는 요구를 고려하여, FORMAT을 중심으로 자원 타입에 따른 요소 확장, 나아가 포맷 레지스트리로 이어지는 단계적 구조를 구성하는 방안을 검토할 필요성이 제시된다. 더불어 파생 기록물 및 컬렉션 집합체 구조를 표현하기 위해 필요한 메타데이터 요소 정의의 필요성이 제기되며, 이를 위해 DCMI에서 제시하는 관련 타입의 메타데이터 표준 어휘를 활용하는 방안에 대한 검토가 요구된다. 자료 간 관계 표현 방식과 관련해서는, MARC 계열의 76X-78X 연관저록, 5XX 주기사항, 6XX 주제명 부출표목, 9XX 로컬 필드를 활용한 관계 표현과 비교하여, MODS(Metadata Object Description Schema)나 DC의 Relation 요소를 활용하는 방식의 효율성을 검토할 필요성이 제시된다. 또한 발간자나 생산자가 사용하는 메타데이터 표준과 파일 형식을 국립중앙도서관이 강제할 수 없다는 점에서 다중 메타데이터 표준을 전제로 한 운영 방안을 마련하는 것이 필요하다는 인식이 제시된다.

❑ 데이터 구조 영역 개선 방향

데이터 구조 영역의 개선 방향으로는, 자원의 개체 분할 기능과 하나의 기록매체에 수록된 다수의 기록자원 개체를 분할하여 관리하고 서비스할 수 있는 구조의 필요성이 강조된다. 특히 대용량 파일에 대한 스트리밍 서비스를 위해 분할 파일을 생성하거나, 하나의 기록매체에 함께 수록되어 있으나 의미적으로

서로 다른 기록데이터를 분리한 파생 기록물을 생성해야 하는 요구가 증가하고 있음이 지적된다. 온라인자료의 경우 원본 파일의 용량 크기에 따라 저용량 변환이나 파일 형식 변환 작업이 빈번하게 발생하는 현실 또한 반영된다.

이와 함께 다양한 분할 파일들을 하나의 그룹으로 체계화하여 관리할 필요성이 제시된다. 보존용 파일과 서비스용 파일, 기타 파생용 파일을 보존센터가 생성·수집·관리하는 구조가 요구되며, 이는 아날로그 자료의 보존·복원 처리 과정에서 생성되는 파생 기록물까지 포함하는 개념으로 제시된다. 현재 원본 파일이 각 부서별로 분산 관리되고 있는 상황을 개선하여, 보존센터가 이를 체계적으로 관리할 필요성이 제기되며, 활용과 서비스의 대상은 보존센터에서 각 해당 부서로 연계되는 구조가 적절한 것으로 판단된다. 이러한 분산 관리 체계하에서도, 하나의 기록물에 대해 생성된 다양한 파생 기록물이 여러 부서와 시스템에 분산되어 있더라도, 전체를 하나의 그룹으로 상세히 파악하고 관리할 수 있는 기능이 요구되며, 이 역할을 보존센터가 수행하는 것이 적절하다는 인식이 제시된다.

보존센터 서비스의 목표 설정과 관련해서는, 보존용 파일은 보존센터가 직접 관리하고 서비스용 파일은 각 서비스 부서가 보존센터에 생산을 요청하여 제공받아 활용하는 구조가 적절한 모델로 제안된다. 이 경우 보존센터는 생성된 모든 파일을 보존하며, 서비스 부서는 이를 활용하는 역할을 담당하는 구조와 기능을 갖추는 것이 필요하다고 제시된다. 다만 이러한 시스템은 KOLIS의 단일 모듈이나 서브시스템으로 보기 어렵고, OAIS(Open Archival Information System) 참조모델과도 직접적으로 일치하지 않는 구조로 인식된다. 그러므로 보존 및 복원 업무를 수행하는 기능적 단면과, 저장된 디지털자원 그룹을 각 부서에 서비스하는 단면으로 구성되는 모델이 적절하다는 판단이 제시된다.

저작권 관리 관점에서는 관내 열람과 대외 서비스 등 서비스 유형의 구분, Audience 및 Availability 관점에서 발생하는 파생 기록물의 생산, 그리고 하나의 자원 아이템을 구성하는 컴포넌트 요소들 간 저작권의 상이함을 반영한 요소 객체화 문제를 관리할 수 있는 데이터 구조와 기능의 필요성이 제기된다. 이에 따라 보존센터의 서비스 대상은 대외 서비스보다는 대내 서비스를 중심으로 설정하는 것이 적절하다는 방향성이 제시된다. 보존센터의 핵심 업무는 보존과 복원에 있으며, 디지털화 작업의 결과 역시 각 부서가 원본을 개별적으로 소장하기보다는 보존센터로 이관함으로써, 이후 생산되는 파생 기록물의 체계적 관리와 서비스가 가능해질 것으로 판단된다. 더 나아가 보존센터는 향후 충분한 저장 용량을 갖춘 인프라를 구축한 이후, 현업 부서로부터 보존용 파일을 이관받아 관리하는 역할을 수행할 필요가 있다는 결론에 이른다.

4.2. 데이터 영역 전략

4.2.1. 전략 수립 프레임워크

현황분석 전체에서 도출된 이슈 및 시사점을 추진 영역별, 관점별로 종합 분석하여 전략적 방향성과 추진전략을 수립하였다. 추진 영역별 추진 전략 수립 방향은 1) 업무, 서비스, 자원, 시스템 영역은 자료보존연구센터의 시스템이 지향하는 전략적 방향성의 설정을 위해서 필요한 사항이며, 2) 데이터 구조와 메타데이터 표준은 목표모델 수립에 필요한 사항들을 세분화 하였다.

NO	추진 영역	추진 전략 수립 방향
1	업무 모델	자료보존연구센터가 속하는 국가문헌보존관의 업무모델을 분석하여 자료보존연구센터의 데이터 아키텍처가 지향해야 할 전략적 방향성을 식별한다.
2	서비스 모델	업무모델 중 서비스 모델을 분석하여 자료보존연구센터의 데이터 아키텍처 설계에 미치는 영향을 식별한다.
3	대상 자원 범위	자료보존연구센터의 수집 혹은 이관 대상 자원의 유형, 장르, 소스를 식별하여 데이터 구조, 메타데이터 표준의 정의에 필요한 사항을 식별한다.
4	시스템 모델	국가문헌보존관의 시스템 운영체계로 OAIS 모델 및 PREMIS 모델 적용여부를 분석하여 자료보존연구센터 데이터 아키텍처가 지향해야 할 전략적 방향성을 도출한다.
5	데이터 구조	자료보존연구센터 데이터 아키텍처 설계의 전략적 방향성을 다음의 영역별로 식별한다. – 보존과 사용의 동시 발생 요구 수용성 – 데이터 구조의 평면성과 다면성 – 다양한 파생 자원의 체계적 관리 – 사용시스템과 보존시스템의 데이터 연계
6	메타데이터 표준	자료보존연구센터 메타데이터 설계의 전략적 방향성을 다음의 영역별로 식별한다. – ERD 및 PREMIS 모델의 적용성 – 자원의 구조화 및 관계 설정 메타데이터 – 다중 메타데이터 표준의 수용 – 테크니컬 메타데이터 표준의 확장

[표 35] 데이터 영역 전략 수립 프레임워크

4.2.2. 데이터 영역별 전략

제시된 내용은 국가문헌보존관과 자료보존연구센터를 중심으로 업무・서비스・대상 자원・시스템・데이터・메타데이터의 각 층위에서 데이터 영역의 전략적 추진 방향을 설정한다.

❑ 업무 모델

업무 모델의 측면에서 국가문헌보존관은 국가 도서관의 디지털 아카이브로서 ‘사용’과 ‘보존’이 통합된 업무 모델을 지향하며, 이용률이 상대적으로 낮으나 보존 가치가 있는 도서 및 기타 인쇄물을 보존하고 서비스하는 역할을 수행하는 것으로 규정된다. 동시에 국가적 차원의 자료보존 및 복원처리 수행을 통해 자료 훼손을 지연시키고 생애주기를 확장하며, 도서・비도서자료의 원형 보존을 위한 보존・복원 처리, 매체변환, 디지털 보존, 디지털 아카이빙을 수행하는 기능을 핵심으로 둔다. 이러한 기능 구성에서 최우선 목적은 문헌 보존과 복원으로 설정되며, 국가장서의 보존과 보존자료의 효율적 관리를 위해 자료의 재등록・재분류, 디지털화 자료의 메타데이터 구축을 위한 자료조직 업무를 수행할 수 있는 것으로 제시된다. 이때 자료보존연구센터는 보존과 복원 업무의 수행 주체이자 국가문헌보존관의 디지털 아카이브 업무와 보존 업무를 통합・연결하는 중심으로서 이를 뒷받침하는 업무 모델과 시스템 모델을 구현해야 한다는 방향성이 제시된다.

❑ 서비스 모델

서비스 모델의 측면에서 국가문헌보존관의 온라인 서비스 플랫폼은 공공도서관의 일반 서비스 모델, 자료 보존 및 복원 서비스, 데이터 분석 및 활용 서비스로 구성되는 것으로 설정된다. 이 구조에서 자료보존연구센터는 최종 이용자에게 서비스를 직접 전달하는 주체라기보다, 온라인 서비스 플랫폼이 제공하는 서비스의 원천 데이터 공급 소스로서 기능하는 것이 적절하다는 방향이 제시된다. 나아가 자료보존연구센터는 서비스 유형과 개별 자료 아이템의 audience 및 availability 정보에 따라 서비스 운영이 가능하도록 데이터 구조와 메타데이터 표준을 갖추어야 하며, 보존・복원 과정에서 수집・생성된 자료와 데이터의 메타데이터 및 파일 전체를 체계적으로 구성・관리・가공・보존・공급하는 데이터 공급자의 역할과 데이터 기반 서비스 기능을 수행하는 것으로 설정된다.

❑ 대상 자원 범위

대상 자원의 전략적 추진 방향에서는 국가문헌보존관을 Born-digital, Original-analog, Digitized-analog 자원을 모두 포함하는 디지털 아카이브로 규정하고 국립중앙도서관 및 지역 도서관의 소장 자원 전체를 입수 자원 유형 및 범위에 포함하는 것으로 설정해야 한다는 방향을 제시한다. 또한 국립중앙도서관의 현행 대상 자원 범위에 더하여, 보존과 복원 업무 활동에서 발생하는 파생 자원과 데이터를 포함하고 이들 전체 소장자원의 관리와 서비스에 최적화될 수 있도록 대상 자원 범위를 설정한다는 원칙이 제시된다.

❑ 시스템 모델

시스템 모델에 관한 추진 방향에서는 국가문헌보존관 정보시스템이 기본적으로 OAIS(Open Archival Information System) 구조를 채택하지 않으며, 운영 데이터 구조 또한 정보패키지 단위 운영이 불가능한 전형적인 도서관 정보관리 시스템 아키텍처를 사용한다는 전제가 제시된다. 업무 모델 분석 결과 국가문헌보존관이 통상적인 도서관 자원 생명주기 관리 체계를 갖는다는 점이 확인되며, 기능 구성상 주요 기능은 보존과 복원이지만, 전체 기능은 수집・서지 기능과 지역문화 서비스 등을 포함하여 국립중앙도서관의 서고 공간을 확장하고 이용률이 낮은 자료의 밀집 소장을 통해 장서개발・자료조직・서비스를 포괄 운영하는 분관 개념으로도 해석될 수 있다는 관점이 병기된다. 이러한 맥락에서 자료보존연구센터의 정보시스템 또한 OAIS(Open Archival Information System) 기반 채택이나 PREMIS(Preservation Metadata: Implementation Strategies) 메타데이터 표준의 즉시 적용이 적절하지 않다는 방향이 제시되며, 자료보존연구센터의 목표 시스템은 ‘사용’ 관점에서 전형적인 도서관 정보관리 시스템과 데이터 아키텍처를 지향하는 것으로 설정된다. 다만 장기보존 시스템의 운영 시점을 고려할 때, 사용 시스템과 보존 시스템 간 자원과 데이터의 이관・이동을 반영한 데이터 구조와 메타데이터 표준을 사용하여 상호운용성과 일관성을 보장할 필요가 있으며, AIP(Archival Information Package) 패키지 구성과 PREMIS(Preservation Metadata: Implementation Strategies) 전환을 고려한 데이터 구조 및 메타데이터 표준을 사용하는 것이 적절하다는 원칙이 제시된다.

❑ 데이터 구조

추진 영역	추진 방향	추진 내역
OAIS 모델 적용 범위	자료보존연구센터 시스템의 데이터 아키텍처의 기본은 OAIS AIP 모델이 아님	• 자료보존연구센터가 사용과 보존을 동시 만족하는 시스템에서 운영되어야 할 데이터 아키텍처는 장기보존을 위해서 Encapsulation 되지 않고, AIP (Archival Information Package) 패키지 방식을 최소한도 적용한 상태에서 운영상태로 전개된 데이터 구조와 메타데이터 표준을 사용한다.
데이터 구조의 다면성	자료보존연구센터 시스템의 운영 대상 자료는 원본, 서비스본 및 일부 디지털 오브젝트 개체의 제한적 집합 구조를 탈피하는 유연하고 확장 가능한 다면적 데이터 구조 적용	• 하나의 자료, 혹은 매체가 다수의 독립된 하위 자료들로 구성된 경우, 자원 구성 요소들을 의미적으로 분리하고, 이들 간의 관계를 구조적으로 연결, 표현하는 다면적 데이터 구조를 적용하는 방향으로 검토가 필요하다. • 자원의 개체 분할 기능, 하나의 기록매체에 수록된 다수의 기록자원 개체의 분할 관리 및 서비스가 가능해야 한다.
파생 기록물 연계 체계	자료보존연구센터 시스템의 보존 복원 활동 및 기타 디지털화 활동에 의해서 발생하는 다양한 파생기록물의 생성, 등록, 파생관계의 연결, 추적 역추적이 가능한 관리체계 사용	• 온라인 자료의 경우, 원본의 용량 크기에 따라서 저용량 변환, 파일 변환 등의 작업이 빈번하게 발생한다. • 특히 대용량 파일들의 스트리밍 서비스를 위한 분할 파일 생성, 하나의 기록매체에 같이 수록되어 있지만, 의미적으로 서로 다른 기록데이터들의 분할 파일 생성 등의 파생기록물 생성 활동을 지원하는 데이터 구조가 필요하다. • 분산된 관리 체계에서, 물리적으로는 한 기록물에 대한 다양한 파생기록물들이 다수의 부서, 시스템에서 분산 발생하는 상황에서도 전체를 하나의 그룹으로 상세하게 파악, 관리하는 기능이 자료보존연구센터에 필요하다.
장기보존체계 이관 연계성	사용 중심의 자료보존연구센터 시스템과 장기보존 중심의 보존시스템과의 데이터 연계 및 상호운용성을 고려한 데이터 구조 적용	• 보존시스템에서 구현될 것으로 예상되는 BagIT 방식의 보존데이터 패키징 방식을 고려해서 ETL 및 SIP 방식을 구현할 필요가 있다. • BagIT 폴더 안에 있는 보존정보 패키지는 최대한 자료보존연구센터의 사용 시스템에서 기계적으로 이관 받을 수 있도록 할 필요가 있다. • 서지적 정보와 테크니컬 정보는 주로 KOLIS III를 포함한 외부 스키마에서 가져오는 방식을 고려한다. • 자원과 자원과의 관계는 사용 시스템의 Structure information 영역에서 METS, 혹은 DC 컬렉션 정보 등에서 가져온다. • 폴더의 구조에 대한 정보를 저장하고 있는 자료보존연구센터의 사용 시스템에서 기계적으로 가져온다. • 통상적인 전자기록물의 장기보존관리에서 적용하는, 국가기록원의 “전자기록물 장기보존 포맷 기술규격 (NAK31:2017)에 해당하는 데이터 구조 및 메타데이터 표준은 본 연구에서는 제외한다. • 평면적 구조의 도서관시스템과 계층적 구조의 기록관리시스템의 자연스러운 데이터 흐름을 목표로 두 시스템 간의 연계/연동/협업이 가능한 데이터 모델을 구현한다.

[표 36] 데이터 구조 영역 전략

❑ 메타데이터 표준

추진 영역	추진 방향	추진 내역
엔터티 관계 모델	자료보존연구센터 시스템의 엔터티 모델은 PREMIS 3.0을 참조하여 개발	• 추상체인 기록, 구현체로서 의 기록물과 기록물을 수록한 매체 간의 다중 관계가 가능한 ERD를 개발한다. • 다만 국가문헌보존관 및 자료보존연구센터가 속한 자원생명주기 단계가 활용 단계에 해당함에 따라서, 데이터 구조 및 메타데이터 표준은 PREMIS 모델을 채택하지 않는다.
메타데이터 표준 계열	파생기록물 및 다양한 계층/요소의 자원 집합체를 다면적으로 표현하는데 적합한 RDF 계열에 유리한 메타데이터 표준을 기본으로 채택	• MARC 계열의 76X-78X 연관저록, 5XX 주기사항, 6XX 주제명 부출표목, 9XX 로컬필드 등을 이용한 자료와 자료 간의 관계 표현 능력보다 METS, MODS, DC의 Relation 요소를 사용하는 방안의 효율성 검토가 필요하다.
서지 메타데이터의 외부 스키마 연계 원칙 적용	다중 메타데이터 표준의 사용 체계를 구축하고, 사용 시스템의 메타데이터 중 서지, 기술, 구조 등의 해당 부분은 보존 시스템이 외부 스키마를 연계 이용하는 방식으로 운영할 필요가 있음	• 발간자, 생산자가 Intellectual entities의 메타데이터 표준, 파일 형식에 사용하는 표준을 국립중앙도서관이 강요할 수 없다. 따라서 다중 메타데이터 표준을 사용하는 방안을 강구하는 것이 필요하다. • 특히 SIP에 해당하는 각 부서와 외부 지역 도서관 들에서 입수되는 메타데이터를 최대한 그대로 운영하는 상황, 즉 보존용 서지 메타데이터는 외부 스키마를 이용하는 지침을 고려할 필요가 있다. • MARC, MODS에서 정의하지 않은 메타데이터 요소들은 DC, SCHEMA.ORG 등에서 발견하고, 적용할 수 있어야 한다. • 장기적으로는 국립중앙도서관이 계획하고 있는 BIBFRAME, RDA 표준도 포함되어야 할 것이다. • 자원 수집, 생산, 운영 환경의 다양성을 유지하고, 자원 유형 구분과 메타데이터 표준의 변화 대응능력을 높일 수 있도록, MODS를 Pivot 표준으로 사용하는 현재의 메타데이터 관리 방식을 다중 메타데이터 표준의 운영을 전제로 하는 Crosswalk 체계로 개선하는 방안을 고려할 필요가 있다.
테크니컬 메타데이터 확장 지원	테크니컬 메타데이터 표준의 확장 방안을 수립하여 시스템이 자동적으로 자원의 타입을 식별하고, 자원의 이용과 재현에 필요한 기술적 사항 제공	• KORMARC 006, 007, 008, 리더 06에서 가지고 있는 Type of Resource, Genre 등의 관계 모호성을 해결하여 기술 메타데이터의 기본적인 유형 결정의 명확성을 높일 필요가 있다. • 자원의 유형을 Text, Image, Sound, Video, 3D 혹은 MODS의 11개 Type of Resource 등으로 세분하거나, 더 나아가서 Format Registry를 이용하여 이용과 재현에 필요한 어플리케이션의 버전과 운영환경에 관한 정보를 제공하는 방안이 필요하다.

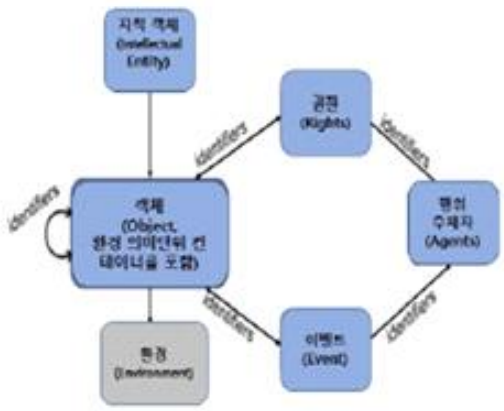
[표 37] 메타데이터 표준 영역 전략

4.3. 관련 기술동향 조사 분석

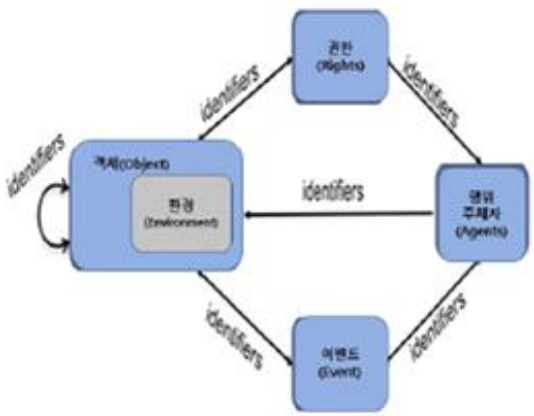
❑ PREMIS(Preservation Metadata: Implementation Strategies) 3.0

PREMIS(Preservation Metadata: Implementation Strategies) 3.0에 대한 분석은 OAIS(Open Archival Information System) 모델에 상응하는 디지털 보존 도메인의 엔터티 모델과 메타데이터 표준을 정의하기 위한 기반 기술을 조사·정리하는 필요성에 의해 제기된다. 그러므로 디지털자료 장기보존을 위한 정보의 기록과 관리가 어떤 엔터티 구조와 메타데이터 표준에 의해 가능해지는지, 그리고 보존 업무의 실무적 구현을 위해 어떤 데이터 모델이 요구되는지를 파악하는 것이 본 분석의 핵심 포인트로 설정된다.

PREMIS(Preservation Metadata: Implementation Strategies)는 디지털 보존 메타데이터 구현 전략에 관한 표준으로서 디지털자료의 장기보존을 위해 필요한 정보를 기록·관리하기 위해 디지털자료의 원본 속성, 특성, 변화 등을 추적하고 문서화하는 것을 목적으로 한다. 또한 PREMIS(Preservation Metadata: Implementation Strategies) 표준은 Library of Congress와 Online Computer Library Center, Inc.(OCLC) 등이 개발하였으며, PREMIS(Preservation Metadata: Implementation Strategies) 엔터티 데이터 모델은 지적 객체(Intellectual Entity), 객체(Object), 에이전트(Agent), 활동(Activity), 권한(Rights) 등을 포함하는 것으로 정리된다. 이들 엔터티는 디지털자료의 식별 정보, 관리자 또는 사용자 정보, 수행 작업 정보, 접근 권한 정보 등을 포함하여 보존관리의 핵심 정보를 구성하는 역할을 수행한다.



[그림 31] PREMIS 데이터모델 ver.2



[그림 32] PREMIS 데이터모델 ver.3

이 가운데 지적 객체(Intellectual Entity)는 디지털 자산과 관련된 지적인 콘텐츠 또는 개념적 엔터티로서 지적 재산권 정보, 라이선스 및 사용 권한 정보, 인용 및 참조 정보, 저작자 및 창작물 정보, 문화적·사회적 맥락 정보 등을 포함하여 관리하는 데 사용되는 것으로 제시된다. 객체(Object)는 File, Bitstream, Representation의 서브 타입을 가지며, PREMIS(Preservation Metadata: Implementation Strategies) 스키마는 XML(Extensible Markup Language) 형식으로 제공되어 데이터 모델에 따라 보존 관련 정보를 저장하는 방식으로 설명된다. 이에 따라 객체의 식별자, 생성 일시, 형식, 크기 등 다양한 항목이 포함되며, 객체의 형식이나 변화에 대한 정보, 관리자 정보, 사용 권한 정보 등도 포함될 수 있는 것으로 정리된다.

시청각기록물의 메타데이터 모델 관점에서 PREMIS(Preservation Metadata: Implementation Strategies) 3.0 모델의 주요 특징은 세 가지로 제시된다. 첫째, 지적 개체(Intellectual Entity)를 '객체' 개체 범주에 포함시켜, 지적 개체가 복잡한 디지털 객체의 속성과 구조를 표현할 수 있도록 '객체' 개체의 일부로 수정하였다는 점이다. 이는 실행성 관점에서 가상의 논리적 객체로 이해되던 지적 개체를 객체 개념으로 포괄함으로써 디지털 객체의 복잡한 구조와 관계를 표현할 수 있도록 한 것으로 서술된다. 그 결과 보존 업무에서 빈번히 발생하는 상황, 즉 하나의 지적 개체가 여러 지적 개체로 구성되는 관계를 표현할 수 있게 되었으며, 예시로는 하나의 지적 개체인 웹페이지에 또 다른 지적 개체인 사진이 포함되는 경우 각각을 독립 객체로 식별하고 '관계' 개체를 통해 연결하여 표현할 수 있게 된 점이 제시된다. 또한 하나의 지적 개체가 여러 버전의 표현물로 생성·관리되는 현실을 반영하여, 하나의 디지털 이미지가 JPEG이나 TIFF 등 다양한 포맷으로 저장·관리될 수 있는 상황에서 지적 개체와 관련된 표현물들을 식별하고 보존 특성을 기술하며 상호 연계할 수 있도록 수정되었다는 점이 포함된다.

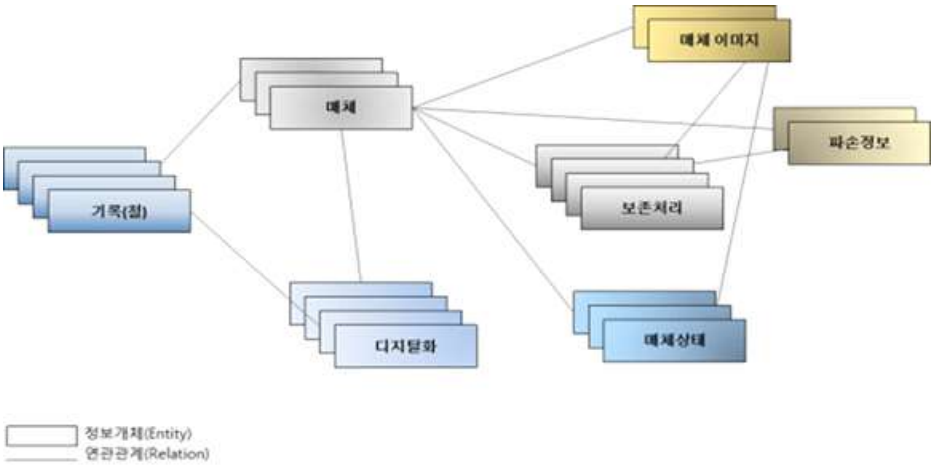
둘째, '환경(Environment)' 개체를 '객체' 개체에서 분리하여 독립된 개체로 추가하였다는 점이 제시된다. 환경 개체는 디지털 객체 재현과 이용에 필요한 운영체제, 응용 소프트웨어, 컴퓨팅 자원 등에 대한 기술 정보로 정의되며, 이를 별도의 데이터 모델로 추가함으로써 기술 변화에 따른 파일과 소프트웨어의 상세 정보를 지속적으로 확보할 수 있도록 한 것으로 서술된다. 또한 PREMIS(Preservation Metadata: Implementation Strategies) 3.0의 환경 개체는 모듈화되어 다른 객체들 간 재사용과 공유가 가능하도록 보존 기술 정보를 제공할 수 있으며, 관계 표현의 중요성을 반영하여 related Environment Purpose, related Environment Characteristic

등의 요소가 추가되었다는 점이 포함된다.

셋째, ‘객체’ 개체에 물리적 객체(Physical Object)의 개념을 추가하였다는 점이 제시된다. 물리적 객체는 필사본이나 인쇄 문서의 내용 정보에 한정되지 않으며, 관련 디지털 정보를 구현하는 데 필요한 물리적 하드웨어 기기 등의 환경 요소나 디지털 객체의 인쇄형 표현물 등을 포함하는 것으로 설명된다. 실제 보존 업무에서 물리적 객체가 여러 디지털 객체와 밀접하게 연관되어 있음을 고려하여 데이터 모델에 물리적 객체를 추가하였고, 이로써 PREMIS(Preservation Metadata: Implementation Strategies) 3.0은 객체의 형태가 디지털이든 아날로그이든 관계없이 기술할 수 있으며, 물리적 객체는 관련 디지털 객체와 ‘관계’를 통해 연계하여 기술하도록 수정되었다는 점이 정리된다.

한편 PREMIS(Preservation Metadata: Implementation Strategies) 모델은 다양한 디지털 리포지터리에서 수행되는 보존 프로세스에 대한 이해를 바탕으로 디지털 자원의 보존 생애주기를 반영하고 있다는 점이 주목되는 것으로 제시된다. 더블린 코어(Dublin Core) 등 다른 메타데이터 스킴이 요소세트를 표준으로 제시하는 방식과 비교하여, PREMIS(Preservation Metadata: Implementation Strategies)는 다른 메타데이터 표준과 결합하여 기관의 실무 환경과 사용 목적에 맞게 요소를 설계할 수 있는 확장성을 갖는 것으로 서술된다. 또한 PREMIS(Preservation Metadata: Implementation Strategies) 의미단위는 보존환경에 따라 1:1, 1:多, 多:1 관계로 다양하게 구현 가능하며, 데이터 사전은 디지털 리포지터리 간 보존 정보패키지 교환에 포괄적으로 활용될 수 있는 표준으로 제시된다. 나아가 OAIS(Open Archival Information System)의 SIP(Submission Information Package), AIP(Archival Information Package), DIP(Dissemination Information Package)에 적용하여 실행하도록 제시하고 있다는 점이 포함된다.

이와 연계하여 PREMIS(Preservation Metadata: Implementation Strategies) ERD 적용 사례로 국가기록원 연구사례가 제시되며, 시청각기록물 보존 업무에서 기록물과 기록매체를 구분하고 업무에 적합한 ERD를 응용 개발한 사례로 분석의 대상으로 설정된다. 해당 사례는 「시청각 유형별 관리 프로세스 재정립 및 메타데이터 체계 재구축 연구(2023년 국가기록관리 활용기술 연구개발, 국가기록원, 2023)」이며, 국립중앙도서관의 2022년 디지털자원 보존 시범 연구와 동일한 필요성과 배경을 가진 사례로서 목록기술규칙이 아닌 엔터티 분석을 통해 데이터 구조와 메타데이터 표준의 개선 방안을 제시했다는 점이 분석 포인트로 제시된다. 따라서 이 연구의 분석 결과와 목표 모델을 참고할 필요가 있다는 방향성이 제시된다.



[그림 33] 기록과 매체의 관계

해당 연구의 배경과 필요성은 생산 기록물의 유형 다양화와 더불어, 시청각기록물에서 내용과 매체의 분리 관리라는 특성으로 인해 기록과 매체의 관계가 1:1로 고정되지 않는 사례가 다수 발생하고 있다는 문제의식으로 서술된다. 연구 초기 단계에서 제시된 엔터티 관계 문제는 첫째, 하나의 기록물이 다수의 기록매체에 분할 저장되어 매체별 분리 식별·처리 체계가 효과적이지 않았다는 점, 둘째, 하나의 기록매체에 다수 기록물이 저장되어 기록물별 분리 식별·처리 체계가 효과적이지 않았다는 점으로 정리된다. 이러한 사례를 소장관리, 상태점검, 보존복원, 디지털화 업무에서 처리하는 과정에서 기록과 매체 관계를 의미적 맥락으로 표현하는 표준이 정의되지 않았고, 각 개체를 탐색·식별하며 관련 개체를 정확하고 신속하게 추적하는 것이 불가능하였다는 문제가 제시된다.

연구의 최종 목표는 기록물 내용과 매체 간 다중 관계를 가지는 시청각기록물의 보존복원 수행 과정을 표현·관리하기 위한 메타데이터 모델을 설계하고 적용 가능성을 시험하는 것으로 설정된다. 이를 위해 기록(철)과 기록매체의 연관관계가 복잡한 오디오·비디오 기록을 대상으로 보존처리 업무별 메타데이터를 분석하고 과정 메타데이터 관리 모델을 설계하며, 업무 프로세스별 활동과 생산 정보의 특성과 구조를 분석하고 관리개체(Entity)를 정의한 후 처리 단위를 설계하고 각 개체에 포함될 메타데이터를 설정하는 접근이 제시된다. 또한 개체 간 메타데이터 연관관계(Cardinality)를 설계하고 획득 메타데이터와 과정 메타데이터를 반영하여 기록 내용과 매체를 관리하는 다중 연관 관계 모델을 정의하는 것으로 목표가 구체화된다.

연구 결과에서는 시청각기록물을 ‘기록’과 ‘기록매체’라는 두 엔터티로

구분하고 기존의 기록:기록매체=1:1의 고정 관계를 n:n의 cardinality를 갖는 다중 상관관계로 개선하였다는 점이 핵심으로 제시된다. 또한 이 개념을 분할 기록매체나 파생기록물뿐 아니라 기록물 전 생명주기에서 발생하는 다양한 파생 기록물 관계의 모델링에도 동일한 개념으로 적용할 수 있는 개념 모델로 설계하였다고 정리된다. 아울러 분할 기록매체와 파생 기록물 등이 시청각기록물에 한정되지 않고 구현체를 가지는 모든 기록물에서 발생할 수 있는 현상임을 전제로, 제시된 엔터티 데이터 관계 모델이 국가기록원 전체 기록물 관점에서의 적용 가능성을 제시하였다는 점이 포함된다.

마지막으로 연구결과의 자료보존연구센터 기술 적용성과 관련하여, ERD 측면에서는 추상체인 기록과 구현체인 기록물을 서로 다른 엔터티로 구분함으로써 구현체 영역에서 다양하게 변화하는 파생 기록물들의 맥락과 정확한 상관관계를 표현할 수 있음을 증명하였고, 이를 토대로 기록물과 기록매체를 서로 다른 엔터티로 추가 구분하는 엔터티 관계 모델 설계가 가능해졌다는 점이 제시된다. 또한 파생 기록물 관리 측면에서는 파생 기록물의 의미적 상관관계 표현에 DCMI의 컬렉션 구조 및 관계 타입 어휘를 사용할 수 있으며, 이를 통해 기록물 관리 데이터 구조상에서 파생 기록물의 관리, 추적, 탐색 및 서비스가 가능하다는 방향이 제시된다.

❑ BIBFRAME 3.0

국립중앙도서관의 미래 메타데이터 표준 및 데이터 구조를 모델링하기 위해서는, 현재 기관 내부에서 연구・검토 중인 BIBFRAME의 적용 가능성을 선행적으로 평가할 필요가 있다. 특히 BIBFRAME은 단순한 서지 메타데이터 포맷의 변경이 아니라, 향후 데이터 구조의 중심축(도메인 온톨로지)으로 기능할 수 있는 잠재력을 갖는다는 점에서, 현행 KORMARC/MARC 계열 체계와의 관계, 전환 비용과 운영 부담, 그리고 다중 표준 환경에서의 상호운용성 확보 전략을 함께 고려한 분석이 요구된다.

BIBFRAME(Bibliographic Framework)은 서지 기술을 위한 데이터 모델로서, MARC 표준을 대체하기 위한 목적하에 미의회도서관이 2011년 프로젝트를 착수하고 2012년 설계 초안을 발표하였다. 이후 2016년 BIBFRAME 2.0에서 WORK/INSTANCE/ITEM 데이터 모델을 기반으로 한 핵심 구조가 정립되면서 MARC의 레코드 중심 기술 방식에서 개체(엔터티) 중심 구조로의 전환 방향이 보다 명확해졌다. 이 흐름의 연장선에서 2025년 12월 1일 BIBFRAME 3.0이 발표

되었으며, 이는 단순 버전업이 아니라 변환 도구·온톨로지의 정합성 강화, 그리고 BIBFRAME의 지향점이 MARC을 RDF로 변환하는 기술적 목표를 넘어 도서관 영역의 상위 도메인 온톨로지로서 의미 모델을 구성하는 중심으로 이동할 수 있음을 시사한다.

BIBFRAME 3.0에서 2.0 대비 강조되는 변화는, MARC-to-BIBFRAME 및 BIBFRAME-to-MARC 변환 프로그램의 오류 수정과 최신 버전 반영, 그리고 BIBFRAME 및 BFLC 온톨로지 v3.0 공개로 요약된다. 아울러 변환 프로그램 v3.0의 주요 변경 사항으로는 (1) MARC 필드 048, 382에 포함된 연주 매체(Medium of Performance) 정보를 변환하는 과정에서 PMO의 속성과 클래스를 사용하는 방식의 도입, (2) MARC 필드 383(음악 식별자)을 단순 문자열(literal)로 처리하지 않고 식별자의 하위 클래스(subclass)로 다루는 구조적 개선, (3) MARC 필드 024 $2의 식별자 출처 코드에 대해 Standard Identifiers 스킴 목록의 URI를 사용하는 방식의 적용이 제시된다. 이러한 변화는 변환 정확도와 표현력의 개선을 목표로 하면서도 결과적으로 BIBFRAME이 단순히 MARC를 RDF로 옮기는 매핑 규칙의 집합을 넘어 도메인 의미 모델로서의 역할을 강화하는 방향과 연결되어 해석될 수 있다.

다만 기술 적용성 관점에서 BIBFRAME 채택은 몇 가지 구조적 고려를 전제한다. 본 내용에서 지적하듯이 BIBFRAME은 RDF 문법과 형식을 채택했으나 MARC 필드 의미론을 상당 부분 유지하고 있으며, WORK와 INSTANCE가 분리되는 구조를 갖추었더라도 책임표시사항(statement of responsibility)이 여전히 문자열로 남아 있거나 편차·주기·기술 주석 등 다수의 기술 요소가 엔터티로 충분히 분해되지 않는 한계가 존재한다. 이러한 특성은 BIBFRAME이 운영 가능한 RDF 모델이지만 불완전하다는 평가와 함께 대규모 서지 메타데이터 구축·소장 기관에서 MARC으로부터 탈출하기 위한 출구 전략으로도 해석되는 배경을 이룬다. 즉 BIBFRAME은 전환의 방향성을 제공하면서도, 즉시 완전한 의미 모델로 기능하기에는 분해 수준과 표현 범위의 제약이 남아 있음을 전제로 해야 한다.

이와 관련하여 상호운용성과 전환 전략 측면에서는 ISO 11179 메타데이터 레지스트리 표준 맥락에서의 크로스워크 방안 도입 가능성이 함께 제시된다. BIBFRAME을 채택할 경우, RDF 단계로의 진화와 동시에 MARC 계열 메타데이터 표준과의 상호운용성을 확보하기 위해 다수의 변환 명세서와 온톨로지 참조 수단을 활용하는 접근이 가능하다. 그러나 BIBFRAME 온톨로지를 도서관의 상위 온톨로지로 채택하고, 메타데이터 단일 표준 혹은 중심 표준으로 위치시키는 전략은 운영비용, 전문 인력, 검증 부담, 전환 이후 지속적 관리 요소의 증가,

그리고 고비용 구조로의 고착 가능성을 동반할 수 있다는 점을 유의해야 한다.

따라서 국립중앙도서관의 관점에서는 충분한 검토가 완료되기 전까지 BIBFRAME을 미래의 단일 표준으로 단정하기보다, 여러 메타데이터 표준 중 하나로서의 채택 가능성을 열어 두고 다중 표준 환경에서의 운용 전략을 병행 검토하는 것이 상대적으로 안전한 접근으로 정리된다. 특히 다중 메타데이터 표준의 크로스워크 테이블을 구축·운영하는 과정에서 BIBFRAME을 핵심 메타 모델로 활용하는 정도의 역할 설정은 RDF 기반 구조로의 확장 가능성을 확보하면서도 전환 리스크와 운영 부담을 단계적으로 관리할 수 있는 현실적 대안으로 제시될 수 있다.

❑ IFLA-LRM

IFLA-LRM에 대한 분석은 자료보존연구센터가 취급해야 하는 원본과 서비스본뿐만 아니라 다양한 목적과 업무 과정에서 생성되는 파생기록물을 데이터 구조 차원에서 체계적으로 처리하기 위한 기준을 모색하는 데 그 필요성이 있다. 특히 동일 저작성에 대한 판단 기준을 어떻게 설정하고 다양한 파생기록물을 집합적으로 관리할 수 있는 체계를 어떻게 구성할 것인지를 검토하기 위해 IFLA-LRM의 개념과 구조를 참조하는 것이 본 분석의 핵심 포인트로 설정된다.

IFLA-LRM의 이론적 배경은 IFLA가 개발한 FRBR 계열 개념모델에 있다. FRBR은 지적 창작물 데이터베이스에 대한 온라인 정보검색 능력의 최적화를 목적으로 설계된 개념적 ERD 엔터티 관계 모델이며, 이후 FRBR, FRAD, FRSAD로 구성된 모델군 전체를 통합한 최신 개념모델로 IFLA-LRM(Library Reference Model)이 제시되었다. IFLA-LRM은 도서관뿐만 아니라 박물관, 기록관, 미술관 등 GLAM 기관 이용자가 자원을 탐색, 식별, 선택, 확보, 탐사(Find, Identify, Select, Obtain, Explore)하는 일련의 태스크 수행을 최적화하는 것을 목표로 하며, 서지적 기능 중심의 목록 구축을 넘어 이용자의 발견과 탐색을 용이하게 하는 방향으로 진화한 개념모델로 설명된다. 이러한 특성으로 인해 IFLA-LRM은 GLAM 기관의 데이터 저장 구조와 서비스 최적화를 위한 개념적 방향성을 설정하는 데 활용 가능한 토대 자료로 평가된다.

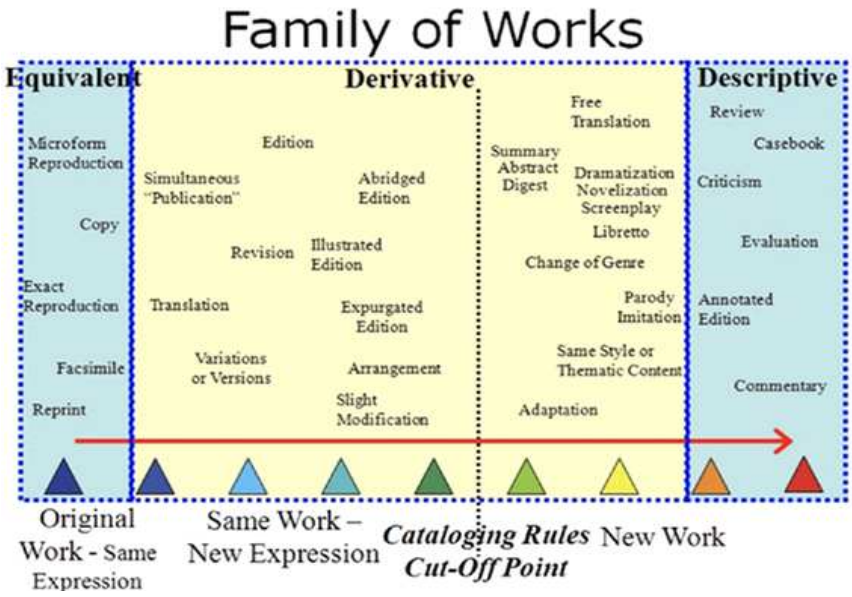
적용성 검토 측면에서 IFLA-LRM은 FRBR을 구성하는 FRBR, FRAD, FRSAD 전체를 통합한 주요 배경 중 하나가, 현재 국립중앙도서관 시청각기록물 보존 업무에서 직면하고 있는 문제, 즉 기록과 기록매체의 분리 관리, 기록과 기록매체의

물리적・의미적 변화로 인해 발생하는 원본・사본・파생본의 다층적 관계 문제와 본질적으로 유사하다는 인식에 근거하고 있다. IFLA-LRM은 강화된 ERM(Entity-Relationship-Modelling) 프레임워크를 토대로 개발된 상위 계층의 개념적 참조 모델을 지향하며, 실세계에서 발생하는 자원과 자원을 둘러싼 이용 행태의 복잡성을 고려하여 WEMI(Work, Expression, Manifestation, Item) 계층관계를 포함하되, 특정 관점이나 특정 형식・형태에 종속되지 않는 포괄적 개념 수준에서 참조 모델을 제시한다.

또한 IFLA-LRM은 FRBR 그룹을 하나의 시스템으로 구현하기 위해 필요한 원칙을 명시적으로 제시함으로써, 실제 애플리케이션 구현 단계에서는 이 모델의 맥락 내에서 확장 또는 생략을 통해 적절한 수준으로 구현할 수 있는 보편성과 지속가능성을 확보한 것으로 평가된다. 이와 함께 RES(개체)와 NOMEN(명칭)을 구분하는 개념을 도입함으로써 보다 발전된 KOS(Knowledge Organization System) 개념모델 구현에 대한 가능성도 제시되고 있다.[73)]

원본으로부터 파생되는 다양한 파생저작물 또는 파생기록물에 대한 고려 필요성은 FRBR 초기 단계에서도 이미 문제로 지적된 바 있다. IFLA-LRM으로 발전하기 이전의 FRBR 초기 모델에서는 Work, Expression, Manifestation, Item의 4계층 구분이 실세계에서 발생하는 복잡한 엔터티 간 관계를 충분히 설명하지 못하는 사례가 빈번히 발생하였다. 특히 Work와 Expression 계층은 논리적 모호성과 비일관성으로 인해 실제 자료조직 업무 적용 과정에서 많은 문제를 야기하였으며, 이 두 계층 사이의 cut-off 포인트를 결정하는 것이 쉽지 않은 상황이었다. 또한 이용자가 자원을 확보(Obtain)한 이후, 확보된 자원에 대해 수행하는 내용적・물리적・형태적 가공으로 인해 Expression과 Manifestation 영역에서 복잡도가 급격히 증가하면서, 기존의 계층적 구조 적용에 한계가 드러났다. 더 나아가 FRBR, FRAD, FRSAD 간 연계와 융합을 전제로 한 통합적 개념모델에 대한 고려가 부족하여, 공통 이슈를 다루면서도 상호 간의 Silo 현상이 심화되고, 하나의 시스템으로 구현하기 위한 가이드라인이 부재하였다는 점이 한계로 제시된다. 아울러 Linked Data 환경에서 요구되는 엔터티 기반 개체 식별과 전거통제를 통한 서지・전거 통합, 그리고 RDA, BIBFRAME 등 기술규칙의 진화에 대응하는 측면에서도 개념과 어휘 정의의 모호성으로 인해 지식・의미적 관계 표현이 충분하지 못하다는 문제가 제기되었다.

73) IFLA, IFLA Library Reference Model: a conceptual model for bibliographic information, 2017



[그림 34] IFLA-LRM(Library Reference Model)

기술적용성 측면에서 IFLA-LRM은 원본과 서비스본 간의 관계 설정뿐만 아니라 다양한 목적과 업무 현장에서 발생하는 파생 기록물 개체들의 파생 유형을 식별하고 개체 간 정확한 관계를 설정하는 데 적용 가능한 참조 모델로 평가된다. 특히 시청각기록물의 기록 개체와 기록매체 개체 간 상관관계, 그리고 개체 내부의 분할·분리에 따른 변화를 수용할 수 있는 엔터티 관계 모델을 설계하는 데 있어 IFLA-LRM을 참조할 필요성이 제시된다.

또한 IFLA-LRM은 사용자 중심의 USE/CASE 모델로 적용될 수 있으며, GLAM 기관의 보편적 자원에 대한 기능적 요구사항과 이용자 태스크 모델(탐색, 식별, 선택, 확보, 탐사)을 설계 맥락으로 설정함으로써 데이터 모델과 서비스 모델에 공통적으로 적용될 수 있는 기본 틀을 제공한다. 이 과정에서 서지와 전거를 분리된 체계로 다루기보다는, 엔터티와 관계 중심의 단일 개념적 모델로 구현함으로써 하나의 정보 공간을 구성하는 단일 체계 구현이 가능하다는 점이 강조된다.

마지막으로 IFLA-LRM은 시청각기록물 데이터베이스 구축 시 정보검색과 활용을 중시하는 발견·탐색 중심의 미래 지향적 데이터 모델로 적용 가능성이 제시된다. 이는 전통적인 목록 메타데이터 표준과 기술규칙 중심의 접근에서 벗어나, 이용자의 발견과 탐색 관점을 반영한 데이터 모델 설계를 지향하며, RDA, BIBFRAME, KCR 등의 개선 방향과의 정합성을 유지하고 CIDOC-CRM, EDM 등 GLAM 참조 데이터 모델과의 완전한 호환성과 연계성을 확보할 수 있는 방향으로 정리된다.

❑ DC Collection Application Profile

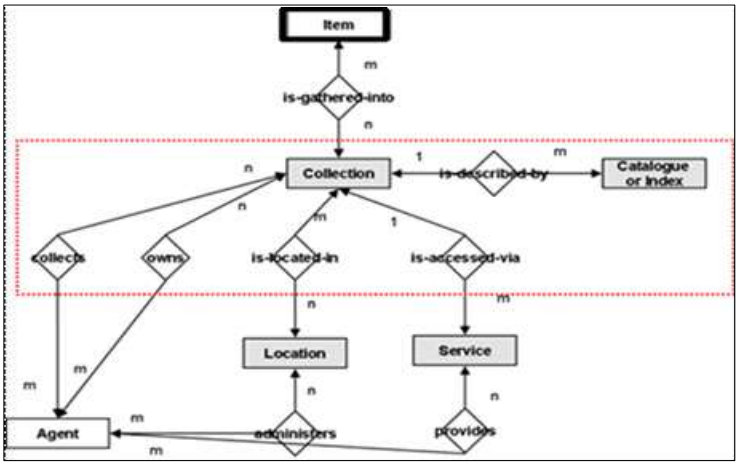
더블린 코어(Dublin Core) Collection Application Profile에 대한 분석은 시청각 기록물과 같이 다수의 개체가 분할·파생되어 집합체를 형성하는 자원을 데이터 구조 차원에서 어떻게 표현하고 관리할 것인가라는 문제의식에서 출발한다. 특히 자료보존연구센터가 기록매체의 분할, 파생기록물의 생성 등으로 인해 형성되는 다양한 집합체를 체계적으로 기술하고 관리하기 위해, 컬렉션 단위 메타데이터 표준의 적용 가능성을 검토하는 데 분석의 필요성이 있다.

더블린 코어(Dublin Core) Collections Description Application Profile(DCCDAP)은 저작물 또는 전자·비전자 자원들의 집합체를 체계적으로 표현하기 위해 필요한 데이터 구조와 메타데이터 요소를 권고하는 표준으로 설명된다. 여기서 '컬렉션'이라는 개념은 물리적 자원과 디지털자원을 모두 포괄하는 자원(resource)의 집합체에 적용될 수 있으며, 집합체를 구성하는 자원의 유형에는 자연적 개체, 창작된 개체, Born-digital 아이템, 물리적 아이템의 디지털 대체물, 그리고 이러한 컬렉션에 대한 메타데이터 레코드들의 집합체 등이 포함될 수 있는 것으로 제시된다. 또한 컬렉션을 구성하는 기준은 아이템의 유형이나 형식, 소유 또는 기증과 같은 프로비넌스, 소유권이나 출처 등 다양한 관점에서 설정될 수 있음을 전제로 한다.

컬렉션의 기술적 특징과 관련하여, 컬렉션은 포함할 수 있는 아이템의 수나 영속성 수준에 제한이 없으며, 컬렉션 수준의 기술은 개별 아이템이 아니라 컬렉션 자체를 하나의 단위로 설명하는 데 목적이 있음을 명확히 한다. 이러한 컬렉션 수준 기술은 Michael Heaney의 「An Analytical Model of Collections and their Catalogues」에서 제시된 'unitary finding-aids' 개념을 참조할 필요가 있으며, DCCDAP은 이러한 컬렉션 수준 기술을 제공하기 위한 더블린 코어(Dublin Core) 메타데이터 기술 세트의 구성 방식을 명시하고 있는 것으로 설명된다.

DCCDAP의 데이터 모델은 여러 엔터티 타입으로 구성된다. 우선 컬렉션(collection)은 아이템들의 집합체를 의미하며, 아이템(item)은 물리적 또는 디지털 자원(resource)을 지칭한다. 장소(LCation)는 컬렉션을 소장하고 있는 물리적 또는 관리적 위치를 의미하고 서비스(service)는 컬렉션 내 아이템들에 대한 접근을 가능하게 하는 시스템을 의미한다. 또한 목록 및 색인(catalogue or index)은 컬렉션을 설명하는 아이템들의 집합으로 정의된다. 이와 같은 엔터티 구조를 통해 컬렉션과 개별 자원, 접근 및 관리 환경을 포괄적으로 기술할 수 있는 데이터 모델이 제시된다.[74]

74) http://dublincore.org/groups/collections/collection-application-profile/



[그림 35] DCCDAP의 데이터 모델

기술적용성 측면에서 DC 컬렉션 메타데이터 표준은 개체의 의미적·물리적 분할에 따라 형성되는 상하 관계를 표현하고 다수의 개체들을 다양한 관점에서 집합체로 구성할 수 있는 기능을 제공함으로써, 시청각기록물의 기록매체 분할에 따라 생성되는 분할 개체 간의 관계를 표현하는 데 유용한 참조 모델로 평가된다. 또한 DC 컬렉션 메타데이터 표준은 메타데이터 속성을 어휘로 정의하고 속성 간의 의미적 관계를 RDF 구조로 표현할 수 있는 메타데이터 어휘 표준을 채택하고 있어, 집합체 표현을 위한 메타데이터 표준으로 채택할 필요성이 제시된다.

아울러 DC 컬렉션 메타데이터 표준은 상대적인 계층관계를 표현하는 방식을 채택함으로써, 시청각기록물에서 기록과 기록매체의 분리 및 분할로 인해 발생하는 계층 수의 증가와 무관하게 계층 구조를 표현할 수 있는 장점을 갖는 것으로 설명된다. 다중 메타데이터 간의 매핑 완전성이 높고, 서로 다른 메타데이터 표준의 속성들을 하나의 레코드 내에서 조합하여 사용할 수 있는 높은 상호운용성을 보유하고 있다는 점에서 타 메타데이터 표준과의 공존 가능성 또한 강조된다. 마지막으로 DC 컬렉션 메타데이터 표준은 아이템의 계층적 컬렉션 구조를 기술규칙 중심이 아니라 데이터 모델 관점에서 명확하게 지원하며, DCMI에서 제공하는 풍부한 사용 설명서와 가이드라인을 활용할 수 있어, 시청각기록물 메타데이터 관리 업무의 효율성을 제고하는데 기여할 수 있는 표준으로 정리된다.

❑ DCAT

DCAT(Data Catalog Vocabulary)에 대한 분석은 시청각기록물과 같은 디지털 자원을 데이터세트 단위로 조직하고 이를 외부 부서・기관・시스템과 유통・공유하기 위한 표준적 데이터 목록 모델의 적용 가능성을 검토하는 데 목적이 있다. 특히 자료보존연구센터가 관리・생성하는 데이터와 파생 데이터들을 공개 데이터 목록 체계와 연계하거나, 데이터세트 기반 관리 방식으로 확장할 경우 참조할 수 있는 메타데이터 표준으로서 DCAT의 기술적 특성과 적용성을 검토할 필요성이 분석의 포인트로 설정된다.

DCAT은 DERI가 개발하고 eGOV Interest Group이 개선하였으며, Government Linked Data(GLD) Working Group을 통해 최종적으로 표준화된 W3C 권고 RDF 어휘로서, 웹상에서 발행되는 데이터 목록들 간의 상호운용성을 제고하기 위해 설계된 표준이다. DCAT을 활용하여 데이터 목록 내 데이터세트를 기술함으로써, 데이터 발행자는 다양한 데이터 목록의 메타데이터를 보다 효율적으로 운영할 수 있으며, 이용자 역시 데이터 검색과 접근을 용이하게 수행할 수 있는 환경을 제공받게 된다. 이러한 특성으로 인해 CKAN 역시 DCAT을 기본 데이터 환경으로 채택하고 있는 것으로 제시된다.

CKAN에서의 메타데이터 표준 응용 사례를 살펴보면, Project Open Data Metadata Schema를 따르며, 어휘 표준은 W3C의 DCAT을 기반으로 하고 있음이 제시된다. 이 표준은 주로 더블린 코어(Dublin Core) 계열 표준과 기타 네임스페이스를 참조하는 RDF 어휘 기반 구조를 채택하고 있다. 또한 메타데이터 매핑이 가능한 타 표준으로는 CKAN API(Application Programming Interface), DCAT, Schema.org 간의 매핑 구조가 제시되며, 지리공간 메타데이터와의 크로스워크로는 CSDGM과 ISO 19115가 언급된다. 메타데이터 교환 포맷으로는 JSON(JavaScript Object Notation)이 주로 사용되는 것으로 정리된다.

데이터 모델 측면에서 DCAT은 메인 클래스 구조를 중심으로 설계되어 있다. dcat:Catalog는 데이터세트의 개요를 파악할 수 있는 메타데이터 요소들로 구성된 목록 메타데이터를 의미하며, dcat:Dataset은 목록이 지시하는 개별 데이터세트에 대한 상세 메타데이터를 나타낸다. dcat:Distribution은 데이터세트의 보급을 위한 파일 다운로드, RSS 피딩, 웹 서비스 등 다양한 서비스 형태에 대한 메타데이터를 기술하는 데 사용된다. 또한 dcat:CatalogRecord는 데이터세트의 기본적인 기입 사항에 대한 메타데이터로서 선택적으로 사용될 수 있는 요소로 제시된다. 목록에 포함된 데이터세트들은 해당 영역의 주제에 따라 분류되며, 이러한 분류 체계의

기술에는 SKOS 표준이 활용되는 것으로 설명된다.[75] 더 나아가 DCAT 메타데이터, 데이터 파일, readme 문서를 결합한 데이터세트 구성 방안을 적용할 필요성이 제시된다.

기술적용성 측면에서 DCAT의 데이터 구조는 일반화되고 대중화된 오픈 데이터 목록 방식으로서 시청각기록물을 데이터세트 방식으로 생성하여 타 부서·기관·시스템으로 유통하는 경우 데이터세트 생성 표준의 기초 자료로 활용할 수 있는 가능성이 제시된다. 또한 시청각기록물 메타데이터 표준을 개발하는 과정에서도 참조 모델로 활용할 필요성이 제기된다. 이 과정에서 Catalog fields, Dataset fields, Dataset distribution fields로 구성된 CKAN API(Application Programming Interface), DCAT, Project Open Data Metadata Schema, Schema.org를 데이터 스키마로 사용하는 접근이 제시된다.

아울러 DCAT 메타데이터 표준은 더블린 코어(Dublin Core)를 기반으로 한 RDF 구조를 채택하고 있어 상호운용성이 높고, 타 메타데이터 표준과의 공존이 가능하다는 점이 강조된다. 구체적으로는 DC, DC Terms, DCAT, FOAF, RDF, RDFS, SKOS, VCARD, XSD 등의 어휘 표준을 사용하며, 지리공간정보의 경우 CSDGM과 ISO 19115를 활용하는 구조가 제시된다. 메타데이터 교환 포맷으로는 JSON(JavaScript Object Notation)이 주요하게 사용되며, 데이터 제공 형식으로는 CSV, XML(Extensible Markup Language), KML, SHP, PDF 등이 활용되는 것으로 정리된다.

마지막으로 CKAN에서 사용하는 공공데이터 분류 및 조직 체계는 정부 업무·기능 시소러스와 대표적인 주제 분류 체계를 함께 사용할 수 있는 구조를 갖는 것으로 제시된다. 이와 관련하여 DCAT의 theme taxonomy는 기능 시소러스에 해당하며, SKOS의 Concept Scheme을 활용하여 어휘와 의미 통제를 수행할 수 있다는 점이 기술적용성의 중요한 특징으로 정리된다.

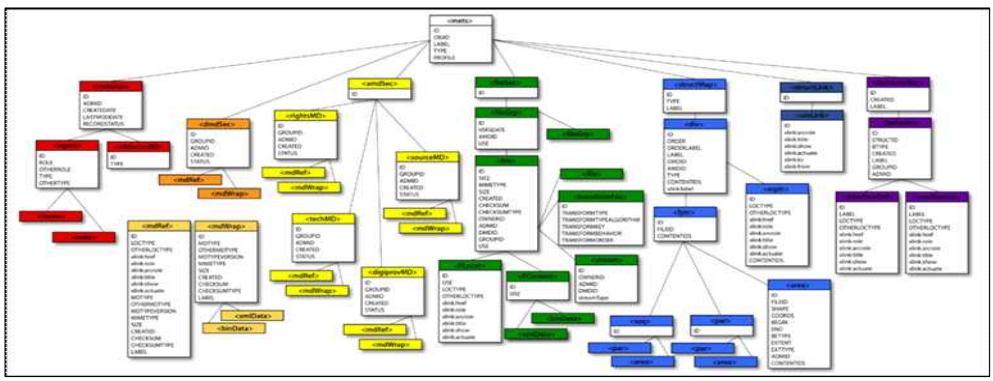
❑ METS(Metadata Encoding and Transmission Standard)

METS(Metadata Encoding and Transmission Standard) 표준에 대한 분석은 복합미디어 매체를 포함한 디지털자원을 구조적으로 표현하고 보존·교환·해석까지 포괄하는 메타데이터 표준의 적용 가능성을 검토하는 데 목적이 있다. 특히 시청각기록물과 같이 이질적인 형식과 속성을 가진 다수의 개체로 구성된 자원을 장기보존 및 서비스 관점에서 어떻게 관리할 것인가라는 문제의식 속에서 METS(Metadata Encoding and Transmission Standard)의 기술적 특성과 실질적 적용성을 검토할 필요성이 제기된다.

75) http://www.w3.org/TR/vocab-dcat/

METS(Metadata Encoding and Transmission Standard)는 복합미디어 매체의 표현, 인코딩, 상호교환 및 해석에 적합한 메타데이터 표준으로 일반적으로 평가되며, 전자도서관 소장 개체에 대한 서지적·관리적·구조적 메타데이터를 XML(Extensible Markup Language) 스키마 기반으로 인코딩하는 표준이다. METS(Metadata Encoding and Transmission Standard) 문서는 OAIS(Open Archival Information System) 참조모델의 SIP(Submission Information Package), AIP(Archival Information Package), DIP(Dissemination Information Package) 단계에서 디지털자료와 그 구조, 관계, 기술 정보를 하나의 묶음으로 구성하여 전달·교환하는 데 유효한 것으로 평가되며, 실제로 다수의 기록관 환경에서 활용되고 있는 것으로 알려져 있다.[76] METS(Metadata Encoding and Transmission Standard) 표준은 미의회도서관(Network Development and MARC Standards Office)에서 공식 웹사이트를 통해 관리·제공되고 있다.[77]

스키마 관점에서 METS(Metadata Encoding and Transmission Standard) 문서는 일곱 개의 주요 섹션으로 구성된다.[78] 구체적으로는 METS(Metadata Encoding and Transmission Standard) header, 기술 메타데이터를 포함하는 Descriptive Metadata Section, 관리 메타데이터를 포함하는 Administrative Metadata Section, 파일 정보를 다루는 File Section, 자원의 구조를 표현하는 Structural Map, 구조적 연계를 정의하는 Structural Link, 그리고 동작 및 행위를 정의하는 Behavior Section으로 구성된다. 이러한 구성은 디지털 객체의 내용과 구조, 그리고 관리·보존에 필요한 정보를 하나의 문서 구조 안에서 통합적으로 기술할 수 있도록 설계되어 있다.



[그림 36] METS 문서 전체 스키마 구조

76) https://en.wikipedia.org/wiki/Metadata_Encoding_and_Transmission_Standard
77) https://www.LC.gov/standards/mets/
78) https://www.LC.gov/standards/mets/METSPrimer.pdf

기술적 특성 측면에서 METS(Metadata Encoding and Transmission Standard)는 단독 메타데이터 표준으로 운영되기보다는 구조 메타데이터를 중심으로 다른 메타데이터 표준과 결합하여 사용하는 것이 일반적이다. 서지 및 기술 메타데이터 영역에서는 MODS(Metadata Object Description Schema)와 더블린 코어(Dublin Core)가 주로 활용되며, MARC을 직접 사용하는 사례는 드문 것으로 제시된다. 운영 메타데이터 영역에서는 디지털 보존을 위한 PREMIS(Preservation Metadata: Implementation Strategies), 이미지 기술을 위한 MIX, OCR 결과 처리를 위한 ALTO 등 다양한 표준이 함께 적용되는 것이 일반적이다.

METS(Metadata Encoding and Transmission Standard) 운영을 위해서는 XML(Extensible Markup Language) 기반 메타데이터 저장과 처리가 가능한 기술 환경이 전제된다. 데이터베이스 측면에서는 PostgreSQL이나 XML(Extensible Markup Language)을 지원하는 MongoDB 등의 검토가 가능하며, 파일 저장 환경으로는 클라우드 스토리지(Amazon S3, MinIO 등)나 로컬 파일 시스템이 활용될 수 있다. 데이터 접근과 검색을 위해서는 Solr나 Elasticsearch와 같은 전체 텍스트 검색 엔진, OAI-PMH나 REST API(Application Programming Interface) 기반의 메타데이터 접근 방식이 고려된다. 인프라 구성 측면에서는 Linux 기반의 클라우드 서버 환경, 대용량 디지털 콘텐츠 저장이 가능한 확장형 스토리지, 고속 네트워크와 백업 및 재해 복구 시스템의 구축이 요구된다.

기능 구조 관점에서 METS(Metadata Encoding and Transmission Standard) 기반 시스템은 콘텐츠 파일의 도큐멘테이션과 디지털 자산 관리 기능을 포함한다. 복합 미디어 콘텐츠 그룹을 METS(Metadata Encoding and Transmission Standard) 워크벤치에 탑재하고 디지털자원의 편집과 가공 파일 생성을 수행하며, 개체 파일의 메타데이터를 입력하고 개체 간 관계와 구조를 정의하는 기능이 포함된다. 이후 METS(Metadata Encoding and Transmission Standard) 문서를 생성하여 XML(Extensible Markup Language) 형태로 인코딩하고 파일과 메타데이터를 각각 데이터베이스와 스토리지에 저장하며, 문서와 파일의 무결성을 검증하는 단계가 수행된다. 저장된 METS(Metadata Encoding and Transmission Standard) 문서는 다시 파싱되어 업무 규칙에 따라 개체 파일과 메타데이터로 분해되고, 디지털 리포지터리로 이관되며, 도서관 표준 메타데이터 형식으로 변환되어 데이터베이스에 저장된다. 이러한 과정을 거쳐 색인과 가공을 통해 최종 서비스 단계로 이어진다.

METS(Metadata Encoding and Transmission Standard) 운영을 위해서는 기술적 환경 구축 이전에 업무적 선행 조치가 요구된다. 대상 자원의 선정 기준, 문서

구조화 방식, 저작권 관리에 관한 절차와 지침 수립이 필요하며, 개체 파일 가공과 등록 표준, 메타데이터 작성 지침과 입력 절차, 유일식별자 및 전거통제 체계 수립이 요구된다. 또한 METS(Metadata Encoding and Transmission Standard) XML(Extensible Markup Language) 문서 프로파일 작성, 파일 경로와 관계의 정합성 검증 프로세스 수립, 운영 및 유지보수, 자원 백업과 장기보존체계 구축이 선행되어야 한다.

METS(Metadata Encoding and Transmission Standard)를 지원하는 시스템은 상용과 오픈소스로 구분된다. 상용 시스템으로는 Ex Libris Rosetta, OCLC CONTENTdm, TIND DA 등이 제시되며, 오픈소스 시스템으로는 Fedora Commons, DSpace, Archivematica, Islandora, OpenBiblioMETS 등이 기관 리포지터리 관점에서 METS(Extensible Markup Language) 처리를 지원하는 것으로 언급된다. 이 가운데 가장 널리 사용되는 조합으로는 Fedora Commons와 Islandora, DSpace, Archivematica 중 하나를 결합하는 구조가 제시된다.

주요 적용 대상 측면에서 METS(Metadata Encoding and Transmission Standard)는 이질적인 표준과 형식, 기계적 속성을 가진 다수의 개체로 구성된 복합 미디어 오브젝트에 적용성이 높은 것으로 알려져 있다. 보다 구체적으로는 첫째, 하나의 저작물이 텍스트, 이미지, 음성, 비디오 등 다양한 파일로 구성된 경우, 이들 파일을 하나의 저작물로 묶어 관리하고 파일 간의 관계와 순서를 구조적으로 식별해야 하는 상황에서 적용성이 높다. 둘째, 다수의 독립된 저작물 파일을 모아 하나의 이차 저작물이나 임의의 컬렉션을 구성하는 경우에도 METS(Metadata Encoding and Transmission Standard)는 파일 간 관계와 구조, 출처, 저작권, 보존 및 활용 정보를 제공할 수 있는 표준으로 제시된다.

국내 적용 사례를 살펴보면, 기록관 분야에서는 PREMIS(Preservation Metadata: Implementation Strategies) 적용과 함께 복합미디어 기록물 보존 방안으로 METS (Extensible Markup Language)가 다수의 논문에서 언급되고 있으나, 국가기록원 등 영구기록물관리기관이나 특수기록관 수준에서 실제 현장에 적용된 사례는 발견되지 않는다. 도서관 분야에서도 국립중앙도서관과 한국과학기술정보연구원 등을 중심으로 METS(Metadata Encoding and Transmission Standard)와 PREMIS (Preservation Metadata: Implementation Strategies)를 활용한 연구 사례는 존재하나, 국가대표 도서관 수준에서 실질적으로 현장 적용된 사례는 확인되지 않는다. 박물관이나 미술관 분야에서도 유의미한 적용 사례는 발견되지 않는 것으로 제시된다.

기술적용성 평가 측면에서는 METS(Metadata Encoding and Transmission Standard) 도입이 투자 대비 효과가 낮을 가능성이 지적된다. METS(Metadata Encoding and Transmission Standard)를 채택할 경우 별도의 기술 환경 구축과 함께, 단순한 메타데이터 표준 추가를 넘어서는 업무적·시스템적 투자가 요구되기 때문에, 대상 자원, 기술 환경, 서비스 목표를 면밀히 검토한 후 도입 필요성과 경제성을 평가해야 한다는 점이 강조된다. 또한 복합미디어 매체 처리를 위해 METS(Extensible Markup Language) 외의 대안도 존재하는 것으로 제시된다. 예를 들어 더블린 코어(Dublin Core) 계열의 CDAP(Collection Description Application Profile)과 Relation 어휘 확장을 통해 개별 자원 간의 계층적 관계와 구성 순서를 표현할 수 있으며, 서지·기술 메타데이터는 DCMI와 MODS(Metadata Object Description Schema)를 활용함으로써 METS(Metadata Encoding and Transmission Standard)만큼 정교하지는 않지만 대안적 구조를 구성할 수 있다는 점이 제시된다. 다만 이러한 대안적 접근의 경우, 출처와 형식이 다양한 다수의 저작물이나 클립을 하나의 완전한 문서처럼 서비스하는 수준에는 한계가 있을 것으로 예측된다.

4.4. 데이터 목표모델 추진 방안

❑ 파생기록물 데이터 구조 추진 방안

파생기록물 데이터 구조 추진 방안은 기록 엔터티와 기록매체 엔터티 간의 개념적 관계를 재해석하고 시간의 흐름에 따라 생성되는 다양한 파생기록물을 구조적으로 관리·보존하기 위한 데이터 모델을 마련하는 데 목적이 있다. 기록물은 생성 이후 보존, 서비스, 복원, 재처리 등의 다양한 과정을 거치면서 다수의 파생기록물을 생산하게 되며, 이러한 파생기록물은 단순한 부수 산출물이 아니라 원본 기록과 의미적·기능적 연속성을 갖는 중요한 관리 대상 자원으로 간주되어야 한다.

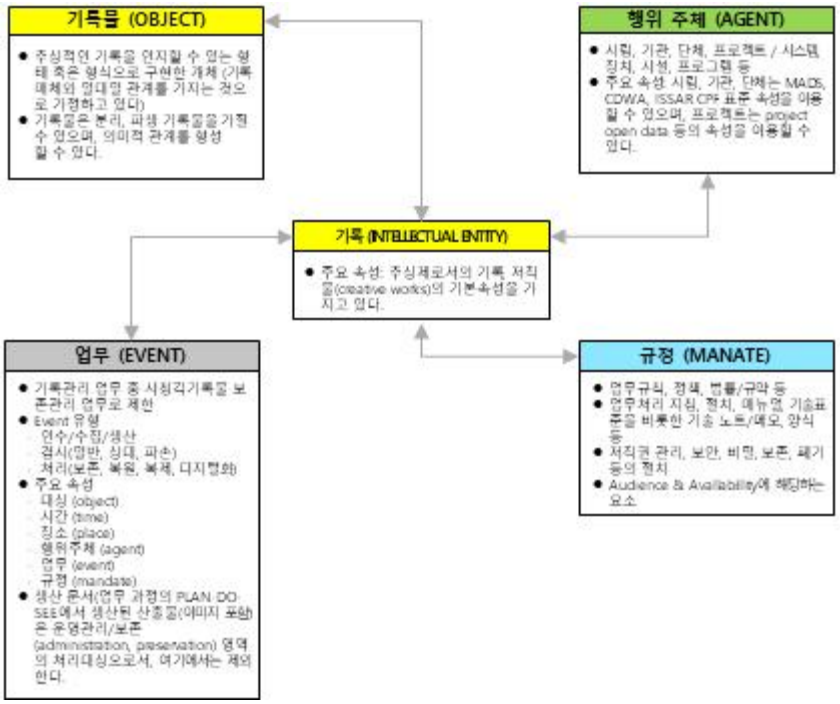
타임라인에 따라 발생하는 파생기록물의 유형은 여러 범주로 구분될 수 있다. 첫째, 보존 목적의 복제 생산은 아날로그 자원의 디지털화, 마이크로필름화, 저장매체의 리프레시(refresh) 또는 마이그레이션과 같은 활동을 통해 발생한다. 둘째, 서비스 목적의 복제 생산은 웹 서비스 제공을 위한 자원 생성, 다양한 사양의 파일 변환, 구간 분할에 따른 부분 자원 생성 등에서 발생한다. 셋째, 훼손된 기록의 복원 과정에서 발생하는 복원 기록물은 훼손 구간의 재생, 노이즈 제거, 색상·음향·형상 복원, 고품질화 작업 등을 통해 생성된다. 넷째, 의미적 또는 내용적 재처리에 의해 새로운 버전의 기록물이 생성되는 경우도 포함되며,

이는 계층 구조의 변경, 개체의 분할 또는 병합, 컬렉션의 분할 또는 병합, 기존에 아이템 단위로 처리되던 복합 개체를 컴포넌트 단위로 세분화하여 재구성하는 경우 등을 포괄한다.

이와 같은 다양한 파생기록물의 발생을 고려할 때, 자료보존연구센터는 원본과 서비스본에 국한되지 않고 위와 같은 모든 파생기록물을 각각 독립된 개체로 식별할 수 있어야 한다. 더 나아가 생성의 선후 관계를 갖는 개체들을 유일식별체계를 통해 상호 연결함으로써, 관리와 서비스 단계에서의 탐색・식별・선택 기능뿐 아니라 장기보존을 위한 정보 패키지 구성 시에도 모든 파생기록물이 하나의 의미적 맥락을 유지한 집합 구조로 조직될 수 있도록 해야 한다.

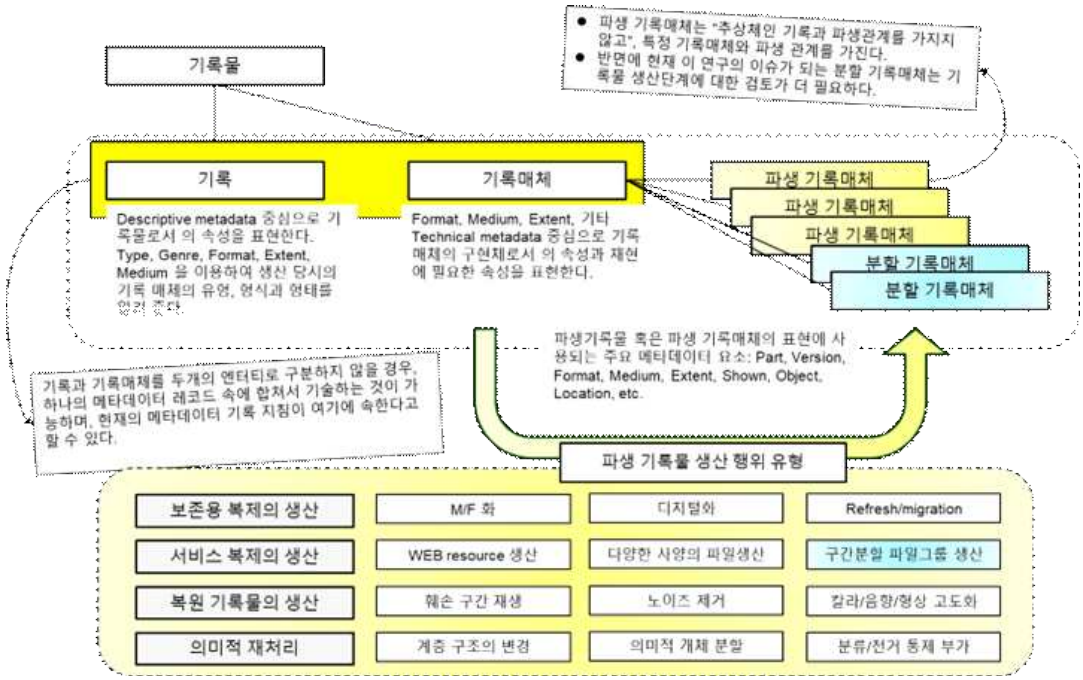
이를 위한 적용방안으로는 우선 추상체인 기록과 구현체로서의 기록물 및 기록매체를 서로 다른 엔터티로 명확히 구분하는 접근이 제시된다. 자료보존연구센터가 관리하는 모든 소장 자원은 아날로그 형태, 디지털 형태, 저장 장치 등 물리적・전자적 형태를 불문하고 자원 개체 단위별로 유일식별체계를 부여받아야 한다. 여기서 자원 개체 단위는 독립된 구현체로 취급 가능한 최소 단위를 의미하며, 자원의 본질적 특성에 따라 정의되거나 관리자의 판단에 의해 설정될 수 있다. 이러한 개체 간 관계는 1:1 구조에 한정되지 않고 n:n의 카디널리티를 가지도록 설계되어야 한다.

특히 하나의 기록이 물리적으로 둘 이상의 기록매체에 분할 수록된 경우에는, 단순히 n:n 관계를 설정하는 데 그치지 않고 관계 유형 정보를 함께 기술함으로써 각 파생 기록물 개체의 성격, 형식, 매체 유형, 물리적・전자적 소장 위치, 파일 구동에 필요한 기술적 정보 등을 정확하게 제공할 필요가 있다. 이러한 관계 표현 방식은 분할 기록매체뿐만 아니라 디지털화・복원・변환 등으로 생성된 파생 기록매체에도 동일하게 적용되어야 한다.



[그림 37] 엔티티 관계 모델

이와 같은 적용방안을 반영하여 개발된 엔터티 관계 모델(ERD)은 파생기록물의 구조적 연계와 관리의 기반이 된다. 파생기록물의 구조적 연계 및 기본 메타데이터 요소 적용에 있어서는, 별도의 특수 표준을 도입하기보다는 DCMI와 같은 이미 공개된 보편적 메타데이터 표준을 활용하는 것으로 충분하다는 점이 제시된다.



[그림 38] 파생기록물의 구조적 연계 및 기본 메타데이터 요소

또한 주요 파생 관계를 표현하기 위해 사용하는 관계 요소는 MARC21, MODS(Metadata Object Description Schema), 더블린 코어(Dublin Core)에서 제공하는 관계 요소를 활용하되, 필요에 따라 확장 가능하도록 설계하는 것이 적절한 것으로 평가된다. 이러한 접근은 실제 디지털 리포지터리 환경에서 파생기록물의 관계 표현에 활용되고 있는 사례, 예를 들어 Digital Repository of Ireland(DRI)가 사용하는 관계 요소 적용 방식과도 맥락을 같이한다.[79]

79) Building the Digital Repository of Ireland Infrastructure, DOI: http://dx.doi.org/10.3318/DRI.2015.5

no	표준	관계표현어휘	메타데이터 요소	설명
1	DC	Relation	dcterms:relation	Link to internal DRI objects, or external digital materials
2		Is Part Of	dcterms:isPartOf	Collection structure/hierarchy representation
3		Is Referenced By	dcterms:isReferencedBy	Link to internal DRI objects, or external digital materials
4		References	dcterms:references	Link to internal DRI objects, or external digital materials
5		Is Version Of	dcterms:isVersionOf	Link to internal DRI objects, or external digital materials
6		Has Version	dcterms:hasVersion	Link to internal DRI objects, or external digital materials
7		Is Format Of	dcterms:isFormatOf	Link to internal DRI objects, or external digital materials
8	MODS	Preceding	relatedPreceding	Expression of sequencing
9		Succeeding	relatedSucceeding	Expression of sequencing
10		Host	relatedHost	Collection structure/hierarchy representation
11		Constituent	relatedConstituent	Collection structure/hierarchy representation
12		Original	relatedOriginal	Link to internal DRI objects, or external digital materials
13		Other Version	relatedVersion	Link to internal DRI objects, or external digital materials
14		Review Of	relatedReview	Link to internal DRI objects, or external digital materials
15		Other Format	relatedFormat	Link to internal DRI objects, or external digital materials
16		Is Referenced By	relatedReferencedBy	Link to internal DRI objects, or external digital materials
17		References	relatedReferences	Link to internal DRI objects, or external digital materials
18		Series	relatedSeries	Link to internal DRI objects, or external digital materials
19	MARC	Other Edition Entry	MARC Tag: 775	Link to internal DRI objects, or external digital materials
20		Additional Physical Form Entry	MARC Tag: 776	Link to internal DRI objects, or external digital materials
21		Preceding Entry	MARC Tag: 780	Expression of sequencing
22		Succeeding Entry	MARC Tag: 785	Expression of sequencing
23		Other Relationship Entry	MARC Tag: 787	Link to internal DRI objects, or external digital materials

[표 38] DRI 관계 요소 사례

종합하면, 파생기록물 데이터 구조 추진 방안은 기록과 기록매체를 단일한 관리 단위로 취급하던 기존의 평면적 구조에서 벗어나, 추상체-구현체-매체 간의 다중 관계를 전제로 한 엔터티 중심 데이터 모델을 구축함으로써 파생기록물의 생성 맥락을 보존하고 장기적인 관리・서비스・보존을 동시에 지원할 수 있는 구조를 마련하는 데 그 의의가 있다.

❑ 컬렉션 및 계층기반 데이터 구조 추진 방안

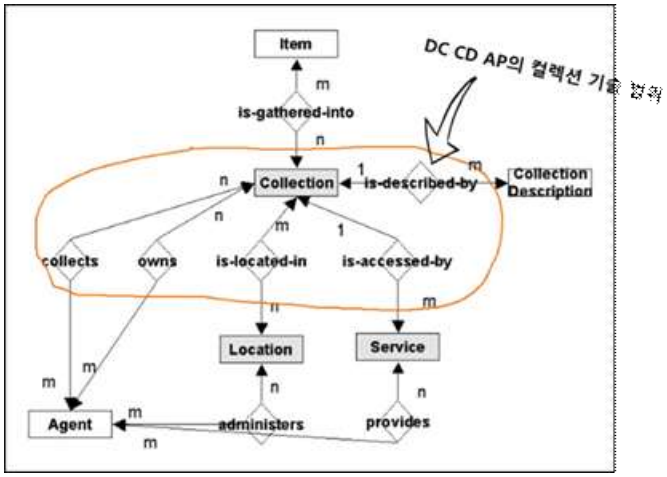
컬렉션 및 계층 기반 데이터 구조 추진 방안은 자원이 가지는 데이터 구조적 속성 가운데 계층적 속성과 요소 간 관계를 체계적으로 표현할 수 있는 목표 모델을 정의하는 데 목적이 있다. 이는 개별 자원을 단위로 하는 평면적 데이터 구조를 넘어, 자원 간의 상하위 계층 관계와 구성 요소 간의 연관성을 구조적으로 표현함으로써, 복합적인 자원 관리와 향후 확장성을 동시에 확보하기 위한 접근으로 설정된다.

계층 관계의 표현을 위해서는 컬렉션과 아이템 간의 계층적 구조를 고정된 위계가 아닌 상대적 상하 관계로 표현하는 더블린코어 컬렉션 기술 응용 모델인 DC-CDAP(Dublin Core Collection Description Application Profile)을 채택하는 것이 목표로 제시된다. 이 모델은 철/건/컴포넌트 구조나 Fond/Series/File/Item 구조(ISAD(G) 지침)와 같이 사전에 규정된 고정적 계층 구조를 전제로 하지 않음으로써, 컬렉션 구축 과정에서 관리 목적과 자원 특성에 따른 유연성과 창의성을 충분히 지원할 수 있는 장점을 가진다. 이를 통해 자원 유형이나 관리 관점의 변화에 따라 다양한 형태의 컬렉션 계층 구성이 가능해진다.

요소 관계와 관련해서는, 하나의 아이템이 서로 다른 형식 표준이나 저작권 조건을 가진 다수의 구성 요소들로 이루어진 경우, 이들 구성 요소 간의 관계를 데이터 구조상에서 표현할 수 있어야 한다는 점이 강조된다. 일반적으로 이러한 복잡한 관계 표현을 위해서는 METS(Metadata Encoding and Transmission Standard)나 OAI-ORE(Open Archives Initiative Object Reuse and Exchange)와 같은 표준의 적용이 원칙으로 제시되지만, 국내 환경에서는 경제성 문제로 인해 실제 적용 사례가 확인되지 않는 상황이다. 이에 따라 경제성이 확보된 대체 표준이 제시되기 전까지는 DC-CDAP의 계층 관계 모델을 활용하는 것이 현실적인 방안으로 판단된다.

DC-CDAP 기반 접근은 문서 구조를 하나의 완결된 책자 형태로 표현하는 데에는 한계가 있으나, 문서를 구성하는 각 요소 개체별로 메타데이터를 식별하고 관리하는 것은 가능하다는 점에서 실질적인 활용성을 가진다. 또한 향후 METS(Metadata Encoding and Transmission Standard)나 OAI-ORE와 같은 고도화된 표준 구조로 전환할 경우에도, DC-CDAP를 통해 구축된 계층 관계는 상호운용성이 높은 기반 구조로 기능할 수 있다는 점에서 단계적 발전을 고려한 적절한 선택으로 평가된다.

계층 관계 표현을 위해서는 컬렉션과 아이템 간의 계층적 구조를 상하위 계층 간의 상대적 관계로 표현하는 더블린코어 컬렉션 기술 응용 모델인 DC-CDAP (Dublin Core Collection Description Application Profile)을 적용하는 것이 목표로 설정된다.



[그림 39] 계층구조 개념도

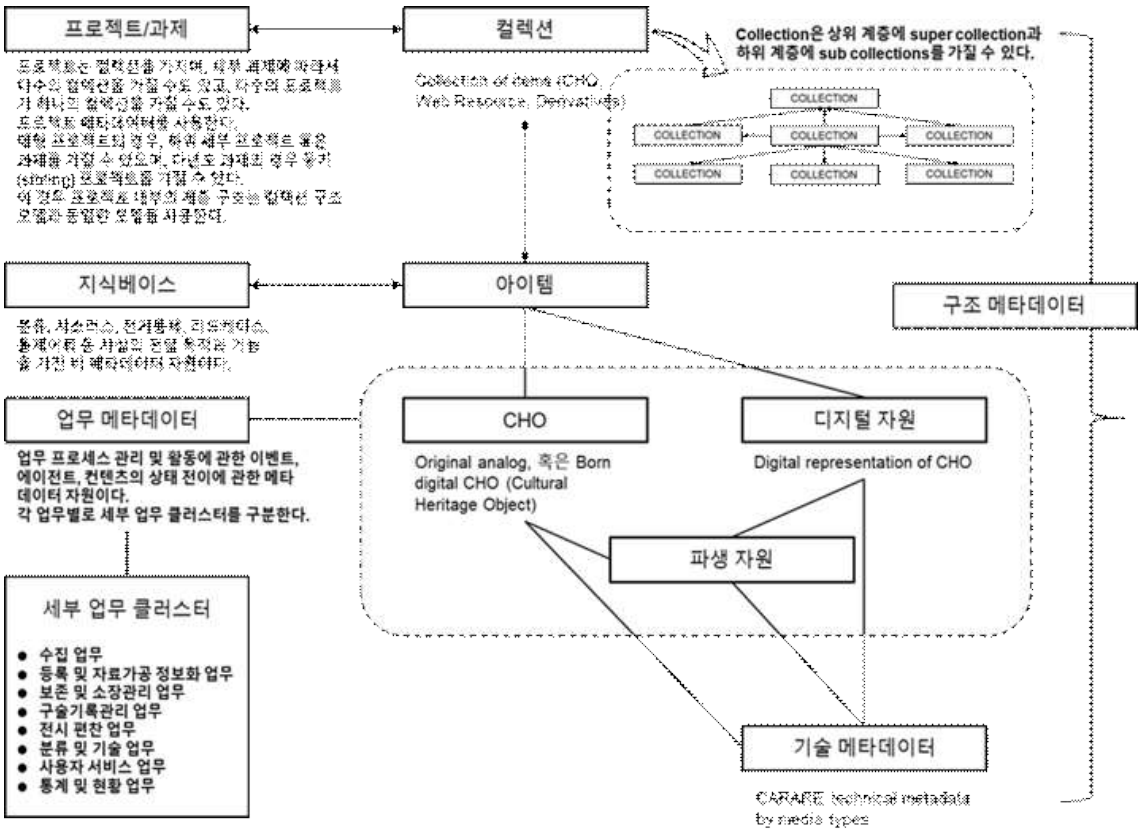
엔터티 유형	정의
컬렉션 (Collection)	다수의 아이템으로 구성된 집합체로서, 물리적 · 디지털자원들을 관리 · 조직하기 위한 상위 단위이다.
아이템 (Item)	물리적이거나 디지털 형태를 가지는 개별 자원(resource)으로서 컬렉션을 구성하는 기본 단위이다.
장소 (LCation)	컬렉션을 물리적 또는 관리적으로 소장하고 있는 장소를 의미한다.
서비스 (Service)	컬렉션에 포함된 아이템들에 대해 접근을 가능하게 하는 시스템을 의미한다.
목록 및 색인 (Catalogue or Index)	컬렉션을 설명하기 위해 구성된 아이템들의 집합으로서 컬렉션의 이해와 탐색을 지원하는 기술 단위이다.

[표 39] 계층구조 설명 용어

관계 유형	주체 엔터티	대상 엔터티	관계 설명
Is-Gathered-Into / Is-Gathering-Of	아이템 (Item)	컬렉션 (Collection)	아이템은 하나 이상의 컬렉션에 모일 수 있으며, 컬렉션은 하나 이상의 아이템으로 구성된 집합체이다.
Is-LCated-In / Is-LCation-Of	컬렉션 (Collection)	장소 (LCation)	컬렉션은 하나 이상의 장소에 위치할 수 있으며, 장소는 하나 이상의 컬렉션의 위치가 된다.
Is-Administered-By / Administers	장소 (LCation)	에이전트 (Agent)	장소는 물리적 · 디지털 환경에 책임을 지는 하나 이상의 에이전트에 의해 관리되며, 에이전트는 0개 이상의 장소를 관리할 수 있다.
Is-Accessed-Via / Provides-Access-To	컬렉션 (Collection)	서비스 (Service)	컬렉션은 하나 이상의 정보적 서비스를 통해 접근되며, 각 정보적 서비스는 정확히 하나의 컬렉션에 대한 접근을 제공한다.
Is-Provided-By / Provides	서비스 (Service)	에이전트 (Agent)	서비스는 하나 이상의 에이전트에 의해 제공되며, 에이전트는 0개 이상의 서비스를 제공할 수 있다.
Is-Described-By / Describes	컬렉션 (Collection)	컬렉션 설명 (Collection Description)	컬렉션은 0개 이상의 컬렉션 설명에 의해 기술되며, 각 컬렉션 설명은 정확히 하나의 컬렉션을 설명한다.

[표 40] 계층구조 관계 유형 (Relationship Types)

목표 데이터 구조 모델은 컬렉션 구조를 중심으로 구성된다. 컬렉션은 전체 아이템, 즉 원본과 모든 파생기록물을 포괄하는 집합체로 정의되며, 관리 목적이나 자원 유형에 따라 계층적 관계를 형성할 수 있다. 아이템은 하나의 자료를 구성하는 모든 개체를 의미하며, 원시 자료, 디지털화 자료, 기타 파생 자료가 하나의 컬렉션 안에 포함된다. 각 아이템은 자체적인 업무 메타데이터와 기술 메타데이터를 보유한다. 업무 메타데이터는 자료보존연구센터 또는 적용 대상 영역의 업무 프로세스 구분에 따라 클러스터 형태로 구성되며, 각 클러스터는 관련된 에이전트, 이벤트, 권한 정보 등을 포함할 수 있으나, 이는 개념도에서는 생략된 것으로 제시된다. 또한 지식베이스는 SKOS 표준을 적용한 그래프 데이터 구조로 구성할 수 있으며, 코드나 스킴 데이터 역시 지식베이스에서 관리하는 방식이 제안된다. 아울러 보존·복원 과제 수행 시에는 해당 과제를 하나의 컬렉션으로 구성할 수 있으며, 기타 보존용 자원의 수집·생성·서비스 업무나 업무 수행의 편의성을 위해 다수의 자원을 집합적으로 관리할 필요가 있는 경우에도 컬렉션 구조를 활용할 수 있는 것으로 제시된다.



[그림 40] 목표 데이터 구조 개념도

목표 메타데이터 요소 참조 모델 사례에서는 컬렉션 메타데이터의 타입을 “Dcmitype:Collection” 으로 선언하여, 해당 메타데이터 레코드가 다수의 아이템이 모인 컬렉션임을 명시하도록 한다. 이 선언은 시스템에서 자동으로 처리되는 것을 전제로 한다. 컬렉션 메타데이터 요소의 정의는 아이템 메타데이터 요소 정의와 동일하게 적용되며, 즉 아이템과 컬렉션이 동일한 더블린 코어(Dublin Core) 메타데이터 요소와 사용 지침을 공유하는 것을 원칙으로 한다. 다만 컬렉션 수준에서 보다 정확한 설명이 요구되는 경우에는 요소 정의에 제시된 설명을 참고하도록 한다. 메타데이터 요소의 관리 측면에서는 색상 구분을 통해 필수 입력 요소와 그룹 처리 요소를 구분하는 방식이 제시되며, 유일식별자와 같이 반드시 입력되어야 하는 요소와, 데이터 존재 시 필수 입력이 요구되는 요소를 명확히 구분하여 관리하는 체계를 지향한다.

Vocabulary Title	Namespace Name	Prefix
The Dublin Core™ Metadata Element Set, v1.1	http://purl.org/dc/elements/1.1/	dc
Dublin Core™ Terms	http://purl.org/dc/terms/	dcterms
Dublin Core™ Type Vocabulary	http://purl.org/dc/dcmitype/	dcmitype
MARC Relator Code Properties	http://www.LC.gov/LC.terms/relators/	marcrel
Collection Description Terms	http://purl.org/cld/terms/	cld
Collection Description Type Vocabulary Terms	http://purl.org/cld/cdtype/	cdtype

[표 41] Dcmitype:Collection에서 사용하는 Namespace

고유번호	요소	요소명	요소 정의 및 사용 설명
col-01	컬렉션 유일식별 ID(*)	collection_identifier	구술기록 컬렉션의 경우도 "컬렉션" 시트를 사용한다. 이 경우 구술기록의 아이템 각각은 "아이템' 시트를 사용한다.
col-02	컬렉션 청구기호(•)	collection_callnumber	아날로그 자원으로 구성된 컬렉션을 위한 요소들이다.
col-03	컬렉션 복본기호(•)	collection_duplicate_identifier	
col-04	컬렉션 인식 ID(•)	collection_mr_identifier	MR = Machine Readable (QR, Barcoded, RFID 등)
col-05	컬렉션 소장 위치 기호(•)	collection_LCation_code	
col-06	컬렉션 온라인 서비스	collection_access_via	온라인 서비스의 URL을 제공한다.
col-07	컬렉션 제목(*)	collection_title_prefer	
col-08	컬렉션 다른 제목	collection_title_alternative	
col-09	컬렉션 생성자(*)	collection_creator	아날로그 컬렉션의 생성 혹은 구축자를 기술한다.
			구술기록 컬렉션의 경우, 구술총서의 면담자, 기여자, 구술자를 기술하는 creator, contributor 요소를 참조할 것
col-10	컬렉션 수집자	collection_collector	수집자, 관리자, 소유자를 contributor에 포함시키고, role로 구분하는 것도 고려할 수 있다.
			구술기록의 경우, 구술총서의 면담자, 기여자, 구술자 요소를 참조할 것
col-11	컬렉션 관리자	collection_manager	아날로그 컬렉션의 관리자이다.
col-12	컬렉션 소유자	collection_owner	아날로그 컬렉션의 소유자이다.
col-13	컬렉션 생성일자(*)	collection_date	컬렉션의 구축을 시작한 날짜이다. 컬렉션 정보의 생성과 구분해야 한다.

고유번호	요소	요소명	요소 정의 및 사용 설명
col-14	컬렉션 언어	collection_language	
col-15	컬렉션 유형	collection_type	
col-16	컬렉션 장르	collection_genre	
col-17	컬렉션 크기(*)	collection_extent	
col-18	컬렉션 재질	collection_medium	
col-19	컬렉션 포맷	collection_format	
col-20	컬렉션 개요	collection_description	
col-21	컬렉션 주제(*)	collection_subject	
col-22	컬렉션 범위(*)	collection_coverage	temporal, spatial을 모두 포함한다. 필요하면 분리한다.
col-23	컬렉션 저작권 사항(*)	collection_rights	
col-24	컬렉션 사용자 등급	collection_audience	컬렉션이 서비스 대상으로 하는 그룹의 등급, 연령, 내외부, 직책 등의 제한이 있을 경우, 여기에 기술한다.
col-25	컬렉션 사용자 열람 제어사항(*)	collection_access_control	
col-26	컬렉션 사용자 대출/반출 제어사항(*)	collection_circulation_control	
col-27	컬렉션 출처	collection_provenance	
col-28	컬렉션에 아이템이 추가되는 방법	collection_accural_method	
col-29	컬렉션에 아이템이 추가되는 주기	collection_accural_periodicity	
col-30	컬렉션에 아이템이 추가되는 정책	collection_accural_policy	
col-31	컬렉션 아이템 연결 정보(*)	collection_item_relation_identifier	
col-32	컬렉션의 하위 계층 연결 정보	collection_haspart_relation_identifier	서브 컬렉션, 시리즈 등 자신의 하위에 있는 집합체 연결 정보를 기술한다.
col-33	컬렉션의 상위 계층 연결 정보	collection_ispartof_realtion_identifier	수퍼 컬렉션 등 자신의 상위에 있는 집합체 연결정보를 기술한다.
col-34	컬렉션 보존상태	collection_presevation_condition	
col-35	컬렉션 디지털 이미지 썸네일 등	collection_media_associated	컬렉션 섹션 정보를 포함한다.
col-36	컬렉션 디지털 파일 경로 URL	collection_file_LCation	컬렉션 섹션 정보를 포함한다.
col-37	컬렉션 생성일자(*)	collection_date_created	컬렉션 정보를 생성한 날짜이다.
col-38	컬렉션 수정일자	collection_date_modified	컬렉션 정보를 수정한 날짜이다.
col-39	컬렉션 정보관리자(*)	collection_datamanager_identifier	관리자의 상세정보가 필요할 경우 리드관리자 메타데이터 요소를 참고한다.
col-40	컬렉션 기타 식별기호	collection_identifier_other	컬렉션에 부여된 기타 식별기호, 내부의 업무용 식별기호와 외부의 자원 식별기호 등을 의미한다.

[표 42] Dcmitype:Collection 입력 항목

※ (*) 가 표시된 요소는 필수 입력 (유일식별 ID), 혹은 데이터가 있을 경우에 필수 입력에 해당하는 항목이다.
※ (•)가 표시된 요소는 “요소 정의 및 사용 설명”에서 하나의 그룹으로 처리하는 요소들임을 표시한다.

❑ 아이템 메타데이터 추진 방안

아이템 메타데이터 추진 방안은 현행 업무 환경에서 이미 생성・운영되고 있는 메타데이터를 최대한 유지하면서, 자료보존연구센터의 목표 데이터 구조와 장기적 보존・관리 체계로 자연스럽게 이관하는 것을 기본 원칙으로 설정한다. 이는 OAIS(Open Archival Information System) 모델에서 SIP(Submission Information Package)의 생산자(Producer)에 해당하는 현행 업무 부서가 작성한 메타데이터를 그대로 활용하는 방향으로 장기보존 단계에서 PREMIS(Preservation Metadata: Implementation Strategies)가 기술 메타데이터(descriptive metadata)를 외부 스키마, 즉 기존 서지정보 관리체계에서 그대로 가져오는 것이 적절하다고 권고하는 원칙과도 부합한다.

아이템 데이터의 전체적인 Ingestion 절차는 단계적・체계적으로 구성된다. 먼저 파일명, 포맷, 식별체계 등 물리적 영역을 중심으로 한 레거시 데이터의 정규화가 수행된다. 이후 레거시 데이터에 대해 오브젝트 단위의 관리가 가능하도록 오브젝트 모델링과 패키징이 이루어지며, 이 과정에서 구조, 기술, 권한, 이벤트, 에이전트 메타데이터와 관계 정보가 결합되어 컬렉션 구조로 래핑된다. 다음 단계에서는 물리적 표준, 의미와 맥락, 내용, 기술적 형식과 형태를 기준으로 레거시 데이터의 클렌징을 수행하며, 이후 목표 데이터 구조에 부합하는 메타데이터 요소의 생성과 검증, 등록을 거쳐 최종적으로 목표 시스템에 데이터가 탑재된다.

파생기록물 아이템 관리 방안에서는 아날로그 자료와 디지털자료, 원본과 사본을 포함하여 보존・활용 과정에서 생성되는 모든 파생기록물이 하나의 컬렉션 안에서 통합적으로 관리되도록 설정한다. 이때 컬렉션의 분리는 국립중앙도서관의 저작 분리 경계선에 따라 이루어지며, 이는 IFLA-LRM에서 새로운 저작으로 분리되는 cut-off 기준에 해당한다. 각 아이템은 서지 메타데이터, 기술 메타데이터, 파일, 소장정보를 기본 구성 요소로 가지며, 자료와 자료 간의 관계는 RELATION 정보를 활용하여 파생기록물 간의 연결고리를 형성한다.

다만 구조가 비교적 단순한 경우에는 컬렉션을 별도로 구축하지 않고 하나의 서지 메타데이터에 대해 다수의 FORMAT과 소장정보를 연결하는 방식, 즉 HAS

FORMAT 관계를 다중으로 사용하는 방법도 가능하다. 이 방식은 하나의 서지 기술 아래 서로 다른 포맷의 존재를 명시하고 참조하게 하는 단순한 구조를 제공한다. 그러나 향후 복잡한 구조를 요구하는 자원이 증가하거나, 하나의 자료에 다수의 저작이 포함되거나 하나의 매체에 다수의 저작이 수록되는 등 계층적 구조가 의미적·물리적으로 복잡해질 가능성을 고려할 때, 이러한 단순 구조는 바람직하지 않을 것으로 예상된다. 특히 기록관리 관점에서 매체 단위의 정확한 관리가 요구되는 상황에서는 '완전한 구조'와 '단순한 구조' 간의 적용성을 비교할 수 있는 충분한 현장 사례가 존재하지 않는다는 한계도 함께 지적된다.

메타데이터 표준 영역에서는 외부 스키마에 존재하는 아이템 메타데이터를 그대로 사용하는 것을 원칙으로 설정함에 따라, 국립중앙도서관이 운영 중인 다양한 자료유형과 업무 유형별 메타데이터 표준을 포괄적으로 수용할 필요성이 제기된다. 이 과정에서 MODS(Metadata Object Description Schema)를 단일 Pivot 표준으로 사용하는 데이터 전환 방식보다는, 정보 손실을 최소화하고 목록기술규칙 중심의 제약을 배제할 수 있는 다중 메타데이터 표준 운영 방식이 요구된다.

이를 위해 국립중앙도서관은 ISO 11179 메타모델을 기반으로 한 메타데이터 Crosswalk 테이블을 참조하여, 자체적인 Crosswalk 테이블을 개발할 필요가 있다. BIBFRAME 표준의 경우, 국립중앙도서관에서 정식으로 채택되기 전까지는 로컬 메타데이터 표준의 하나로서 단일 BLADE 형태로 제한적으로 활용할 수 있으며, 향후 정식 채택이 이루어질 경우에는 메타데이터 Crosswalking 테이블의 기준 표준으로 적용하는 방향을 고려할 필요가 있다. 이러한 접근은 아이템 메타데이터의 지속가능성과 상호운용성을 동시에 확보하기 위한 단계적 추진 방안으로 해석될 수 있다.

❑ 테크니컬 메타데이터 추진 방안

테크니컬 메타데이터는 시스템이 자원의 유형을 자동적으로 식별하고 해당 자원의 이용과 재현에 필요한 기술적 사항을 체계적으로 제공하는 핵심 요소로 설정된다. 특히 업무의 복잡도 단계에 따라 필요한 정보 수준이 상이하다는 점을 고려하여, 단순한 자원 식별 단계에서부터 이용·재현을 지원하는 고도화된 기술 정보 단계까지 점진적으로 활용될 수 있도록 설계하는 것이 요구된다. 이러한 접근은 가장 기초적인 수준에서는 Type of Resource와 Format 정보를 활용하는 방식에서 출발하여, 자원 유형별로 상세화 된 테크니컬 메타데이터를 적용하고

궁극적으로는 자원의 이용과 재현에 필요한 정보를 집약적으로 관리하는 Format Registry와의 연계를 목표로 한다.

자원 유형별 테크니컬 메타데이터의 상세화 방안으로는, 디지털자원 메타데이터의 일반적인 타입 속성에서 출발하여 다섯 가지 유형으로 전개한 사례를 참조한다. 이는 CARARE 기술 메타데이터 표준에서 제시한 방식으로 자원의 기본적인 타입 속성을 기준으로 세부적인 기술 메타데이터 속성을 설계한 접근이다.

DC Type	DC Element	DC Element Refinement	DC Element Refinement (2)	설명
Text	dc:format	File format		파일 형식. PDF, txt, RTF, XML, ePub
		MIME type		MIME 타입. IANA MIME type 통제어휘 사용
		File size		파일 크기. byte
		Resolution		해상도. dpi
		Full text		풀 텍스트 여부. true/false
Image	dc:format	File format		파일 형식. jpg, JPEG, TIFF, PDF
		MIME type		MIME 타입. IANA MIME type 통제어휘 사용
		File size		파일 크기. byte
		Width		이미지의 너비. pixel
		Height		이미지의 높이. pixel
		Color space		이미지의 색깔 표현 유형. sRGB, Greyscale, Black & White
		Color palette		이미지의 색상. CSS3 hexa decimal binary code, up to 6 per image
		Orientation		이미지 대상의 유형. landscape, portrait 등 대상유형에 대한 어휘통제
Sound	dc:format	File format		파일 형식
		MIME type		MIME 타입. IANA MIME type 통제어휘 사용
		File size		파일 크기. byte
		Duration		데이터 재생 시간. Milliseconds
		Channels		오디오 채널 수
		Sample rate		오디오 샘플의 초당 주파수
		Bit depth		오디오 샘플의 비트 수
		Bit rate		미디어 자원의 초당 재생 bit 수
Video	dc:format	CODEC		CODEC 장치 혹은 프로그램 명칭
		MIME type		IANA에서 정의한 MIME 타입
		File size		미디어 자원의 크기. Byte

DC Type	DC Element	DC Element Refinement	DC Element Refinement (2)	설명
		Duration		트랙이나 신호의 재생 시간. Ms
		Width		비디오 프레임이나 이미지의 너비. Pixel
		Height		비디오 프레임이나 이미지의 높이. Pixel
		Resolution		미디어 자원의 공간 해상도. Dpi
		Bit rate		미디어 자원의 초당 재생 bit 수
		Frame rate		비디오 신호 초당 프레임 수
3D	dc:format	File format		파일의 형식 명칭. 통제어휘 사용
			Format name	AutoCAD Drawing 등 어플리케이션에서 주는 파일 명칭
			File extension	dwg, dxf, dwf 등의 파일 확장자
			File version	AutoCAD Drawing 2008, 2018 등의 어플리케이션 버전에 따른 파일 버전
		MIME type		IANA에서 정의한 MIME 타입
		File size		파일의 크기 및 기하학적 구조를 위한 하위 속성 전개
			Size	byte로 표시한 파일 크기
			Scale	파일의 1 unit를 미터법으로 표시한 축척
			Compression	zip, gz, tar, rar 등의 파일 압축에 사용한 표준
			Number of vertices	꼭짓점 숫자
			Geometry type	polygonal, NURBS, CSG 등의 구조 형식
			Number of triangles	삼각형 그리드의 숫자
		File content		파일 내용 표현 방식을 위한 로컬 속성
			Textures	표면의 질감 표현 여부
			Shades	그림자 표현 여부
			Layers	정보 계층 사용 여부 및 숫자
			Blocks/Instances/Components	Are instances being used?
			Lights	광원에 관한 정보
		Application		파일 재현 및 구동 정보를 위한 로컬 속성
			Application name	ArchiCAD, Allplan, Revit Architecture 등의 3D 모델을 생산한 소프트웨어 명칭
			Application version	3D 모델을 생산한 소프트웨어의 버전
			External files	파일의 로딩에 필요한 외부 파일들
			Video data included	동영상 데이터 포함 여부
			Animation data included	애니메이션 데이터 포함 여부
			Sound data included	음향 데이터 포함 여부

[표 43] 테크니컬 메타데이터 CARARE 사례

이 방식에 따르면 기술 메타데이터의 복잡도가 높지 않은 경우에는 더블린 코어(Dublin Core)의 ‘Format’ 타입을 그대로 사용하되, 보다 상세한 기술 정보가 요구되는 경우에는 Format 타입을 Text, Image, Sound, Video, 3D의 다섯 가지 상위 유형으로 세분화하고 이를 다시 하위 전개(refinement)하는 방식을 채택한다. 이러한 단계적 확장은 자원의 특성과 관리 요구 수준에 따라 유연하게 적용될 수 있다는 점에서 실무적 효율성을 갖는다.

기본적으로 설정된 다섯 가지 Type 유형 이외에 추가적인 세분화가 필요한 경우에는, 각 Type별로 Genre 속성을 활용한 하위 전개도 가능하다. 이때 참조가 필요한 코드 목록은 MODS(Metadata Object Description Schema)에서 정의한 Type 및 Genre 항목을 활용하는 것이 적절한 것으로 제시된다. 한편 하나의 자원, 특히 웹자원을 포함한 디지털자원이 복수의 Format을 동시에 가질 경우에는 dcterms:isFormatOf 및 dcterms:hasFormat 관계 요소를 사용하여 이를 표현할 수 있다. 다만 이러한 방식은 향후 하나의 자료에 다수의 저작이 포함되거나, 하나의 매체에 다수의 저작이 수록되는 등 의미적·물리적으로 복잡한 계층 구조가 요구되는 상황에서는 한계를 가질 가능성이 있으며, 장기적으로는 바람직하지 않을 것으로 판단된다.

Format Registry 메타데이터 요소의 구현 방안에서는, 포맷 레지스트리를 파일 포맷의 문법적 특성(바이트 구조, 시그니처), 의미적 특성(정보 모델), 관련 소프트웨어 및 플랫폼, 식별자 등을 체계적으로 관리하는 핵심 참조 체계로 정의한다. 포맷 레지스트리는 이러한 정보를 바탕으로 자원의 보존, 재현, 식별, 검증을 지원하는 역할을 수행한다. 다만 포맷 레지스트리를 지시하는 개별 오브젝트의 서술적(descriptive), 관리적(administration), 보존(preservation) 메타데이터를 레지스트리 내부에 직접 저장하지는 않으며, 대신 외부 연계 포인트나 식별자를 제공하는 방식으로 연계성을 유지하는 것이 적절하다고 판단된다. 파일의 무결성이나 진본성에 관한 정보는 개별 파일 또는 디지털 오브젝트 단위로 저장하는 것을 원칙으로 하되, 특정 포맷에서 반복적으로 발생하는 손상 패턴과 같이 포맷 자체의 특성에 기인한 정보는 포맷 레지스트리에 저장하는 것이 합리적인 것으로 제시된다.

세부 항목		설명
Format Record (포맷 기본 정보)		
	format_id	포맷을 유일하게 식별하기 위한 지속적 식별자
	label	포맷을 지칭하는 기본 명칭
	preferred_name / aliases	권장 포맷명 및 대체 명칭
	version	포맷 버전 정보
	status	포맷 상태 정보(예: active, deprecated, historic, draft)
	EOS / EOP / 대체 애플리케이션 정보	해당 포맷을 사용하는 어플리케이션의 지원 종료 시점 및 대체 가능 어플리케이션 정보
Identification / Signature (식별 · 인증 정보)		
	file_extensions	파일 확장자 정보
	mime_types	MIME 타입(IANA 기준 권장)
	internal_signatures	파일 내부 시그니처 정보
	external_signatures	파일 외부 시그니처 정보
	signature_confidence	시그니처 판별 신뢰도
Information Model (구조 · 구문 정보)		
	structure_description	파일 구조 및 구문에 대한 설명
	byte_ordering	바이트 순서 정보
	character_encoding	문자 인코딩 방식
	compression	압축 방식
	typical_media_types / content_models	대표적인 미디어 유형 또는 콘텐츠 모델(예: text, image, audio)
Rendering / Execution (어플리케이션 환경 정보)		
	required_software	렌더링 · 생성 · 편집에 필요한 소프트웨어(제품명, 버전, 벤더)
	required_os	요구 운영체제
	runtime_dependencies	필요한 라이브러리, 플러그인 등 실행 의존성
Preservation-relevant properties (보존 관련 속성)		
	lossiness	손실 여부(lossless, lossy, N/A)
	potential_for_migration	포맷 마이그레이션 난이도(예: low, medium, high)
	known_issues / known_corruption_signatures	알려진 문제점, 손상 패턴 및 복구 관련 정보
Validation / Integrity (검증 정보)		
	validation_tools	사용 가능한 검증 도구(도구명, 버전, URL)
	validation_rules	구조적 검증 규칙(예: 필수 청크 존재 여부 등)
	sample_files	대표 샘플 파일 정보(미리보기 링크 또는 checksum, size)
Provenance / Maintenance (유지 · 관리 정보)		
	authoritative_source	공식 규격 문서 URL 또는 표준명
	registry_creation_date	레지스트리 최초 생성일

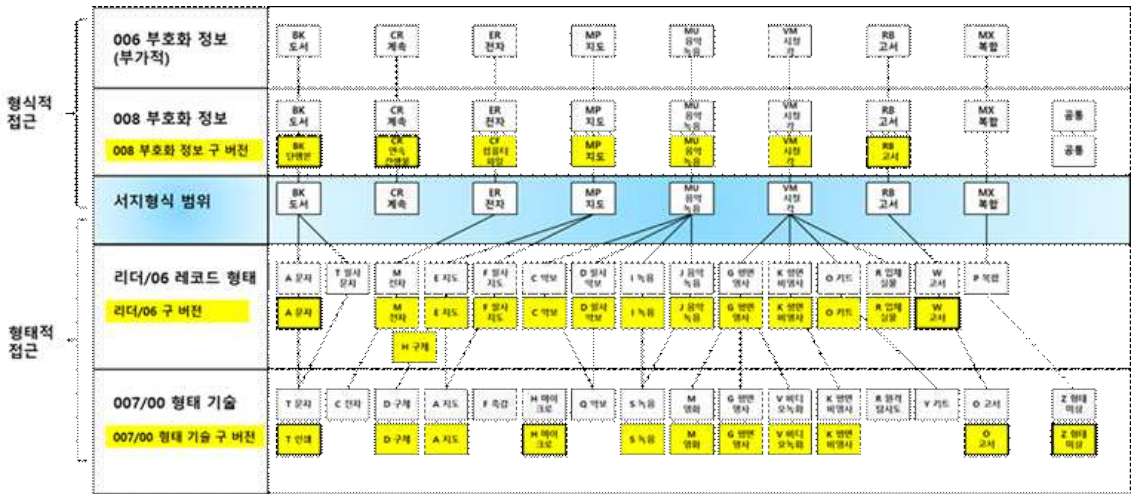
세부 항목		설명
	last_modified	최종 수정일
	curation_contact	관리 책임자 또는 연락처
	change_log	변경 이력(날짜, 에이전트, 변경 내용)
Inter-format relationships (연관 포맷 정보)		
	is_supertype_of	상위 포맷 관계
	is_subtype_of	하위 포맷 관계
	derived_from	파생 관계
	container_of	컨테이너 관계
Identifiers and crosswalks (식별체계 연관 정보)		
	external_identifiers	외부 식별자(IANA MIME, PRONOM PUID 등)
	cross_registry_links	타 포맷 레지스트리 또는 표준과의 연계 링크
Risk / Obsolescence (서비스 상태 정보)		
	obsolescence_risk	포맷 노후화 위험 등급 및 근거
	support_lifecycle	벤더 지원 시작 · 종료 시점

[표 44] 포맷 레지스트리 정보 영역 구성 방안

서지형식, 자원의 타입, 장르 코드와 관련해서는 개선이 필요한 사항도 함께 제기된다. 일반적으로 Type of Resource와 Genre는 자원의 물리적 형태와 콘텐츠의 일반적인 유형을 정의하는 역할을 수행하며, Format과 테크니컬 메타데이터는 자원의 형태적 특성과 디지털자원의 경우 이용과 재현에 필요한 기술적 정보를 제공하는 기능을 담당한다. MODS(Metadata Object Description Schema)에서 정의한 Type of Resource의 유형은 텍스트, 지도 자료, 악보, 음반, 정지 영상, 동영상, 3차원 객체, 소프트웨어 및 멀티미디어, 혼합 자료 등으로 구성되어 있으며, 일부 유형을 제외하면 DCMI에서 사용하는 개념과도 상당 부분 유사성을 가진다.[80] 특히 mixed material 유형은 보존박스, 봉투, 기타 미디어 매체와 같이 컨테이너에 담긴 다수의 자원을 포괄적으로 지칭할 때 사용된다.

반면 통합서지용 KORMARC 표준에서 해당 영역의 코드 매핑 구조는 상대적으로 복잡한 양상을 보이고 있으며, 이는 자료보존연구센터의 실무적 운영 측면에서 부담 요인으로 작용할 가능성이 있다. 따라서 자료보존연구센터 차원에서는 자원의 타입과 장르를 보다 간결하고 명확하게 처리할 수 있는 개선 방안을 모색할 필요가 있으며, 이는 테크니컬 메타데이터와 포맷 레지스트리의 효율적 운영을 위한 전제 조건으로 해석될 수 있다.

80) https://www.LC.gov/standards/mods/userguide/typeofresource.html



[그림 41] 코드 전환을 위한 매트릭스 사례

❑ OAIS(Open Archival Information System) 보존용 메타데이터 요소 분석

OAIS(Open Archival Information System) 보존용 메타데이터 요소 분석에서는 사용 단계가 종료된 이후 디지털자원이 보존 단계로 이관되는 과정에서 현행 시스템에서 필수적으로 생성되고 전달되어야 할 메타데이터 요소를 파악하고 이를 현행 시스템의 데이터 설계에 반영할 필요성을 검토하는 것을 목적으로 한다. 이를 위하여 OAIS(Open Archival Information System) 기반 보존시스템의 SIP(Submission Information Package)와 AIP(Archival Information Package) 단계에서 요구되는 메타데이터의 구성 요소를 분석하여 제시하고 해당 분석을 통해 현행 시스템에서 필수적으로 고려되어야 할 메타데이터 요소를 도출하고자 한다.

❍ SIP (Submission Information Package)

SIP(Submission Information Package)는 생산자(Producer)가 OAIS(Open Archival Information System) 보존시스템에 제출하는 정보 패키지로서, 생산・수집・전송 단계에서 생성되는 최소한의 보존 기반 메타데이터로 구성된다. SIP(Submission Information Package)는 국립중앙도서관의 현행 업무 정보화 시스템에서 생성・운영된 데이터와 SIP(Submission Information Package) 생성 시점에 추가로 생성되는 정보로 이루어지며, OAIS(Open Archival Information System) 관점에서는 Ingest 이전 단계의 입력물에 해당한다. Ingest 단계에서 SIP(Submission Information Package)를 인수한 이후, 해당 정보는 AIP(Archival Information Package)로 확장되고 정규화된다.

SIP(Submission Information Package)의 구성 요소는 Content Information, Preservation Description Information(PDI), Packaging Information, 그리고 선택적으로 Descriptive Information으로 이루어진다. 이 단계의 메타데이터는 전달에 필요한 최소한의 범위로 한정되며, 장기보존을 위한 심층 정보는 포함하지 않는다.

SIP(Submission Information Package) 단계에서 요구되는 핵심 메타데이터 범위에는 식별자, 기본 기술 메타데이터, 기본 fixity 정보, 기초적인 provenance 정보, 기본 서술 및 관리 메타데이터, 그리고 파일 구성 정보가 포함된다. 식별 영역에서는 지적 개체를 식별하기 위한 IE Identifier, 표현 단위 식별자인 Representation Identifier, 개별 파일 단위의 File Identifier, 그리고 현행 시스템에서 부여된 Source System Identifier가 포함된다.

기술 메타데이터 영역에는 파일 포맷명과 버전, 파일 크기, 생성일과 수정일, MIME 타입이 포함되며, 비디오・오디오・이미지・문서 유형에 따라 해상도, 비트 깊이, 코덱과 같은 세부 속성이 추가될 수 있다. 또한 파일의 임시 Ingest 경로와 파일을 생성하거나 처리한 소프트웨어 및 하드웨어 환경 정보가 포함되며, 포맷 레지스트리나 PRONOM PUID와 같은 포맷 위험성 평가는 SIP(Submission Information Package) 단계의 필수 요소로 간주되지 않는다.

고정성(Fixity) 영역에는 체크섬 값, 체크섬 생성 알고리즘, 체크섬 생성 일시가 포함된다. 출처(Provenance) 영역에서는 생산자 또는 제공자, 전달 경로, SIP(Submission Information Package) 생성 날짜, 그리고 담당 에이전트에 대한 최소 수준의 정보가 기록된다. 권리(Rights) 영역은 이관에 필요한 범위로 한정되어 저작권자와 라이선스 정보가 포함된다.

서술 메타데이터는 현행 정보화 시스템에서 제공되는 정보를 그대로 활용하며, 제목, 저자, 발행 정보, 주제 또는 키워드, ISBN과 같은 식별자가 포함된다. 구조 메타데이터 영역에서는 파일 간의 관계를 중심으로 파생기록물 구조가 기술되며, 예를 들어 master 파일, access 파일, OCR 텍스트 간의 관계가 포함될 수 있다. 페이지 순서나 프레임 순서 정보는 현행 시스템의 기능에 따라 SIP(Submission Information Package) 반영 여부가 결정된다. 전달 영역에는 SIP(Submission Information Package) 패키지 버전, 체크섬 매니페스트, Ingest 프로파일 식별자가 포함된다.

❍ AIP (Archival Information Package)

AIP(Archival Information Package)는 장기보존을 위한 완전한 보존 패키지로서 OAIS(Open Archival Information System)의 핵심 데이터 구조에 해당한다. AIP(Archival

Information Package)는 Content Information, Preservation Description Information(PDI), Packaging Information, Descriptive Information으로 구성된다. Content Information은 논리적 내용에 해당하는 Conceptual Content와 비트 스트림에 해당하는 Physical Content로 구분된다.

PDI는 provenance, fixity, context, reference, 그리고 보존 목적에 한정된 접근 권한 정보를 포함한다. Packaging Information은 METS(Metadata Encoding and Transmission Standard), BagIt, XML(Extensible Markup Language) wrapper와 같은 구조를 통해 파일, 메타데이터, 이벤트 정보를 하나의 보존 패키지로 구조화한다. Descriptive Information은 주로 탐색과 식별을 목적으로 하며, 외부 메타데이터 스키마를 참조하는 방식으로 활용된다.

AIP(Archival Information Package)는 SIP(Submission Information Package) 단계에서 전달된 요소를 포함하면서, PREMIS(Preservation Metadata: Implementation Strategies) 기반의 보존 기록이 추가된다. 이 가운데 Objects, Events, Agents, Rights로 구성되는 PREMIS(Preservation Metadata: Implementation Strategies) 기본 구조가 핵심을 이룬다. AIP(Archival Information Package) 단계에서 관리되는 메타데이터 범위에는 심층 수준의 기술 메타데이터, PREMIS(Preservation Metadata: Implementation Strategies) 기반 보존 관리 메타데이터, 보존 이벤트, 보존 에이전트, 보존 관계, 식별자 확정 및 버전 관리 구조, 스토리지 및 위치 정보, 전체 이력 관리, 그리고 파일 포맷 위험성 모니터링 결과가 포함된다.

Object Technical Metadata 영역에서는 MIME 타입과 PRONOM PUID, 포맷 유효성 검증 결과, 포맷 버전, 비트 깊이, 압축 유형, 해상도, 컬러 프로필, ICC 정보, 이미지·오디오·비디오별 상세 기술, 파일 구조 분석 정보가 기록된다. Preservation Administrative Metadata 영역에는 저작권과 접근 제한에 대한 상세 기록, 원본 생산 정보, 포맷 위험성 레벨, 마이그레이션 정책 적용 정보, 접근 제어 정책, 보존 등급 정보가 포함된다.

PREMIS(Preservation Metadata: Implementation Strategies) Event는 AIP(Archival Information Package)의 핵심 요소로서 Ingest, 무결성 검증(Fixity check), 유효성 검사(Validation), 포맷정규화(Normalization), 마이그레이션(Migration), Virus check, Replication, 무결성 체크(Integrity Check)와 같은 이벤트가 기록되며, 각 이벤트는 eventType, eventDateTime, eventOutcome, eventDetail, eventAgent, eventObject 속성을 가진다. PREMIS(Preservation Metadata: Implementation Strategies) Agent 영역에는 보존 시스템, 담당자, 포맷 검증 도구, fixity 도구, 스토리지 시스템이 포함된다.

PREMIS(Preservation Metadata: Implementation Strategies) Relationship 영역에

서는 isVersionOf, hasSource, isDerivativeOf, requires와 같은 관계가 정의되며, 파일 간 구조적 관계는 SIP(Submission Information Package) 단계의 단순 목록 수준에서 AIP(Archival Information Package) 단계에서 정식 관계 구조로 확장된다. Storage 및 LCation Metadata 영역에는 저장 위치, 복사본 수, 지리적 분산 정보, 스토리지 클래스, 백업 정책이 포함된다.

식별자 및 버전 관리 영역에서는 AIP(Archival Information Package) 고유 식별자와 Object ID가 확정되며, 버전 규칙이 적용되고 논리적 구조와 물리적 구조가 구분되어 관리된다. 감사 이력(Audit trail)에는 모든 보존 행위의 누적 기록이 포함되며, AIP(Archival Information Package)는 일반적으로 변경 불가 원칙을 적용하고 변경이 필요한 경우 신규 AIP(Archival Information Package)를 생성하는 방식으로 관리된다.

❍ SIP(Submission Information Package) 이관 필수 메타데이터 요소

본 절은 SIP(Submission Information Package) 이관 과정에서 현행 사용 단계 시스템이 제공해야 할 메타데이터 요소의 중요성을 검토하는 데 목적이 있다. 현행 시스템은 보존시스템을 직접적인 목표로 데이터 구조와 메타데이터 요소를 설계한 것은 아니지만, 현행 시스템에 존재하는 일부 데이터 영역은 SIP(Submission Information Package) 단계에서 요구되는 정보의 존재 여부와 더불어 해당 정보가 가지는 정확성과 효용성에 직접적인 영향을 미친다.

이러한 이유로 SIP(Submission Information Package) 단계에서 필요한 다수의 메타데이터 요소 가운데 일부는, 보존 단계로의 이관 이전인 사용 단계에서부터 신뢰할 수 있는 형태로 충분히 제공되어야 한다. 그러므로 SIP(Submission Information Package) 단계의 메타데이터 품질은 현행 시스템에서 생성·관리되는 데이터의 수준에 의해 좌우된다고 볼 수 있다.

메타데이터 영역	메타데이터 요소	설명
Identifier	Intellectual Entity ID	자료(지적 객체) 단위의 영속 식별자로서 현용 단계에서 영구보존 단계까지 지속적인 유일식별이 가능해야 한다.
	Representation ID	동일 자료 내 디지털 표현 단위 식별로서 하나의 자료가 다수의 요소로 구분될 경우에도 개체 유일식별이 가능해야 한다.
	File ID	디지털 파일 개체 별 유일식별자 이다. 파생기록물 전체가 각 개체 별 유일 식별자를 가져야 한다.
Technical	File Format / Version	파일 포맷 식별, ingest 자동화에 필요 하다.
	MIME Type	포맷 판별 및 검증을 위한 기본 정보이다.
	File Size	Fixity 및 무결성 검증 기준 정보이다.
	Creation / Modification Date	기본 출처 및 파생기록물의 이력과 연결고리 파악에 필요하다.
Fixity	Checksum	파일 무결성 검증을 위한 최초 해시값 이다.
	Checksum Algorithm	SHA-256 등 장기보존 권장 알고리즘이다.
Provenance	Producer / Digitizing Agent	누가 어느 기관/부서가 생성, 디지털화 했는지 증명한다.
Rights	Copyright Status	보존, 가공 및 서비스의 판단 근거이다.
	Access Restriction Level	공개 여부 및 대상에 대한 판단 근거이다.
Descriptive	Title	자료 식별을 위한 기본 서술 요소이다.
	Creator	자료 식별을 위한 기본 서술 요소이다.
Structural	File Relationship Type	원본과 파생기록물 전체의 구조 및 연결정보를 파악할 수 있는 정보이다.

[표 45] SIP 이관 필수 메타데이터 요소

또한 AIP(Archival Information Package) 단계는 SIP(Submission Information Package) 단계를 기반으로 구성되므로, SIP(Submission Information Package)에서 요구되는 메타데이터 요소에 대한 세부적인 작성 지침을 SIP(Submission Information Package) 단계에 한정하지 않고 AIP(Archival Information Package) 단계까지 고려하여 처리하는 것이 필요하다. 이러한 접근은 보존 단계 전반에 걸친 메타데이터 관리의 일관성과 효율성을 확보하는 데 기여할 수 있다.

제5장 국립중앙도서관 디지털자원 보존전략 수립

5.1. 장기보존전략

앞서 제시된 국내외 사례 분석, 국립중앙도서관 현황분석, 현행 KOLIS·OASIS(Online Archiving & Searching Internet Sources) 체계의 구조적 한계, 그리고 메타데이터·스토리지·AI 활용 여건을 종합적으로 고려할 때, 국립중앙도서관의 디지털자원 보존전략은 단기적으로는 사용과 보존이 공존하는 단계와 장기적으로는 완전한 보존으로 진입한 단계로 구성하는 것이 적절하다. 사용과 보존이 공존하는 단계는 현재의 국립중앙도서관 상황을 의미하며, 완전한 보존으로 진입한 단계는 디지털자원의 원본, 사본, 파생자원 등으로 구성된 패키지 전체가 도서관이 정한 기간동안 변화 없이, 안정적 상태를 유지함에 따라서 영구보존용 패키지로 묶어서 저장하는 단계를 의미한다.

완전한 보존 단계에서는 OAIS(Open Archival Information System) 모델과 PREMIS (Preservation Metadata: Implementation Strategies) 표준이 기본 요건이 되지만, 사용과 보존을 위해서는 KOLIS Ⅲ 등의 시스템과 상호운용성을 확보한 상태의 보존 시스템이 요구된다. 이러한 두 가지 목표를 달성하기 위해서 다음과 같은 전략을 수립한다.



[그림 42] 국립중앙도서관 장기보존전략

5.2. 실행 우선순위 설정

국립중앙도서관의 디지털자원 장기보존전략은 현재 조직의 규모와 업무 현황을 보았을 때 모든 전략을 동시에 추진할 수 없는 구조이므로, 다음과 같은 제약조건을 고려한 우선순위 설정이 필요하다.

첫째, 국립중앙도서관은 장기보존에 관한 내부정책과 운영기준 등을 명확히 정의해야한다. 특히 보존에 대한 정의, 사용과 보존의 경계와 관계에 대한 지침 및 절차, 장기보존단계에서 필요한 정책과 운영기준 등을 우선적으로 확립할 필요가 있다.

둘째, 국립중앙도서관의 현 정보화시스템 체계는 서비스(KOLIS) 중심 구조로 운영되고 있다. 따라서, 무결성 관리・스토리지 분산 저장・AI 적용 등 장기전략 수립을 위해서는 사용과 보존이 공존하는 시스템 환경에 대한 선행적 정의가 필요하다.

셋째, 장기보존은 사용단계에서 조치한 보존 활동 데이터가 PREMIS 표준에 따라 정리되어서 자원과 같이 보존되어야 한다. 따라서, 사용 단계에서 보존과 관련된 활동을 수행하고 그 데이터를 유지하는 것이 무엇보다도 중요하다.

이에 따라 실행 우선순위는 ① 정책과 운영 구조를 먼저 만들고 → ② 사용단계의 신뢰성을 확보하며 → ③ 보존단계에서의 확장과 고도화를 추진하는 단계적 논리를 따르는 방향으로 설정하도록 한다.

❍ 1순위: OAIS(Open Archival Information System) 기반 보존 책임 구조 확립

보존의 장기목표로서 가장 우선적으로 수행되어야 할 과제는 OAIS(Open Archival Information System) 기반 장기보존체계를 바탕으로 사용과 보존이 공존하는 체계를 운영할 것임을 선언하는 것이다. 이는 기술 구축 이전에 반드시 선행되어야 하며, 국립중앙도서관이 무엇에 대해, 어디까지 책임지는 기관인가를 명확히 하는 작업이다.

장기적으로는 AIP(Archival Information Package)를 보존의 기본 단위로 정의하고 단기적으로는 현행 시스템과의 상호운용성과 협업체계를 위주로 보존 시스템, 데이터 아키텍처가 구성될 필요가 있음을 선언한다.

OAIS(Open Archival Information System) SIP(Submission Information Package)-AIP(Archival Information Package)-DIP(Dissemination Information Package)의 역할과 PREMIS (Preservation Metadata: Implementation Strategies) 요건을 내부 규정으로 명확히 하지 않는 상태에서는 무결성 관리, 메타데이터, 포맷 정책, AI 도입 모두 단편적 기능 추가에 그칠

가능성이 높다. 따라서 OAIS 기반 보존 책임 구조 확립은 다른 모든 전략의 전제 조건(precondition) 이자, 국립중앙도서관 장기보존전략의 출발점으로 보여진다.

❍ 2순위: 무결성 검증과 보존 이력 관리 체계 구축

OAIS(Open Archival Information System) 구조가 개념적으로 정립되었다면, 다음 우선순위는 보존 신뢰성의 최소 요건을 확보하는 것이다. 그러므로 이 디지털 객체가 언제나 동일한가를 증명할 수 있어야 한다. 체크섬 생성, 검증 주기 설정, PREMIS(Preservation Metadata: Implementation Strategies) 이벤트 기록은 상대적으로 기술적 난이도는 낮지만 보존 효과는 매우 큰 영역이며, 초기부터 누적될수록 장기보존 가치가 기하급수적으로 증가한다. 따라서 국립중앙도서관 전체가 스토리지 관리에 무결성 검증 및 보존 이력관리 체계를 적용할 필요가 있다. 이 단계가 완료되어야만 이후 AI 기반 예측, 외부 감사, 신뢰 저장소 인증 논의가 가능해진다. 따라서 무결성 관리 체계는 OAIS(Open Archival Information System) 다음으로 가장 높은 우선순위를 가지는 것으로 판단된다.

❍ 3순위: 보존용 메타데이터 및 포맷 정책의 제도화

보존 데이터 구조와 메타데이터 체계가 정립되면, KOLIS, OASIS(Online Archiving & Searching Internet Sources) 등의 현행 시스템과의 데이터 연계/연동 및 상호운용성 확보를 위한 기술적, 제도적 장치를 정립하는 것이다. 그 다음은 포맷 관리의 정책화이다. 포맷 정책은 대규모 시스템 구축 없이도 내부 지침과 수집 기준 정비만으로 효과를 낼 수 있으며, 장기보존 리스크를 구조적으로 감소시키는 수단이다. 특히 국립중앙도서관처럼 출판물・웹자원・멀티미디어가 혼재된 기관에서는 자료유형별 우선 포맷 정의가 장기보존 비용과 위험을 크게 좌우한다.

❍ 4순위: 다계층・분산 스토리지 확장

스토리지 전략은 중요하지만, 초기 우선순위는 상대적으로 낮다. 그 이유는 저장 인프라는 비용・조직・외부 협력이 동시에 요구되며, 보존 정책과 관리 논리가 정립되지 않은 상태에서 확장할 경우 비싼 저장 공간에 무엇을 어떻게 넣을 것인가에 대한 목적이 불명확해지기 때문이다. 따라서 스토리지는 보존 단위(AIP), 무결성 관리, 포맷 정책이 정착된 이후 단계적으로 확장하는 것이 합리적이다.

○ 5순위: 조직・거버넌스 재편

조직 개편은 언제나 어렵고 비용이 크기 때문에, 기능과 필요성이 충분히 가시화된 이후 추진하는 것이 현실적이다. 앞선 전략들이 일부라도 작동하기 시작하면 장기보존 전담 조직의 필요성이 자연스럽게 드러나며, 이 시점에서 전담 조직・인력 확충은 조직 개편이 아니라 업무 안정화로 인식될 수 있다.

○ 6순위: AI 기반 고도화 전략

AI 전략은 중요하지만, 가장 마지막 단계에서 본격화되어야 한다. AI는 데이터를 만들지 못하고 구조가 없으면 학습할 수 없기 때문이다. OAIS(Open Archival Information System) 구조, 무결성 로그, PREMIS(Preservation Metadata: Implementation Strategies) 이벤트, 포맷 관리 데이터가 축적된 이후에야 AI는 의미 있는 보존 의사결정 지원 도구로 작동할 수 있다. 따라서 AI는 처음부터 도입할 기술이 아니라 앞선 전략들이 축적한 성과를 증폭시키는 가속 장치로 위치 지워져야 한다.

❑ [전략 1] OAIS(Open Archival Information System) 기반 보존 책임 구조 확립

필요성		국립중앙도서관은 OAIS(Open Archival Information System)를 일부 도입하고 있으나, 이를 실제 운영 단위에서 구현할 수 있는 보존 전담 아키텍처와 책임 구조는 아직 명확히 정립되어 있지 않다. 현재 자원 생명주기 관점에서 사용 단계와 보존 단계의 경계가 구체적으로 정의되어 있지 않으며, 어떠한 조건과 기준에 따라 OAIS 기반 영구보존 체계로 전환할 것인지에 대한 운영 원칙 또한 제도화되어 있지 않다. 이에 OAIS를 기관의 실질적 운영 프레임워크로 내재화하고, 단계별 책임 구조와 시스템 전환 기준을 명문화하는 것이 장기보존 전략의 핵심 과제로 요구된다.
수행방향		• OAIS를 기관의 장기보존 운영 원칙 및 책임 프레임워크로 재정의 • 장기보존을 개별 시스템 기능이 아닌 기관 핵심 업무로 격상 • “사용 · 보존 공존 단계” 와 “보존 중심 단계” 의 명확한 구분 체계 수립 • 보존 전환 조건(보존 가치 확정, 메타데이터 완비, 기술 안정성 확보 등) 명문화 • KOLIS 서비스 시스템과 단계적 · 구조적 연계 모델 구축 • 보존 중심 단계의 OAIS 6대 기능 완전 통합 체계 지향
적용 방안	**단기**	• 현행 데이터 구조와 OAIS 기능 매핑 • SIP(Submission Information Package) 단계 도입 대비 • SIP 필수 구성요소 정의 (Content Information, Representation Information 등) • SIP 생성 기준 및 검증 절차 정책 문서화 • 기존 메타데이터 구조의 SIP 전환 적합성 점검 • 보존 전환 대상 선정 기준 초안 수립 • 데이터 생애주기 관리 개념 모델 수립
	중기	• AIP(Archival Information Package) 논리 모델 정의 • AIP 구성요소 체계화 (Content Information, PDI, Packaging Information 등) • SIP→AIP 자동 생성 · 검증 · 등록 프로세스 구현 • AIP 전용 저장소 및 관리 모듈 구축 • PREMIS 기반 보존 이벤트 자동 기록 체계 도입 • Archival Storage 기능의 독립적 운영 체계 정립 • 보존 단계와 서비스 단계의 시스템 분리 · 연계 구조 확립
	장기	• DIP(Dissemination Information Package) AIP 기반 자동 생성 체계 구현 • 서비스 시스템이 항상 AIP를 참조하는 구조로 전환 • 보존 객체 중심 서비스 아키텍처 완성 • OAIS 6대 기능(Ingest, Archival Storage, Data Management, Administration, Preservation Planning, Access) 완전 통합 • ISO 16363 기반 신뢰성 평가 체계 연계 • 국가문헌보존관 인프라에 OAIS 구조 반영 • 법 · 제도 차원의 보존 책임 명문화

[표 46] [전략 1] OAIS 기반 보존 책임 구조 확립

❑ [전략 2] 무결성 검증과 보존 이력 관리 체계 구축

필요성		디지털자원의 장기보존에서 무결성(Fixity)과 포맷 유효성 검증은 신뢰성 확보의 핵심 전제 조건이다. 사용-보존 공존 단계부터 무결성 및 포맷 유효성 검증을 독립된 보존 기능으로 정의하고, 완전한 보존 단계에서는 PREMIS(Preservation Metadata: Implementation Strategies) 이벤트 중심의 검증-기록-대응 전주기 관리 체계를 구축할 필요가 있다. 이는 장기보존의 신뢰성을 제도적으로 담보하기 위한 필수 기반이다.
수행방향		• 무결성 및 포맷 유효성 관리를 사후 점검 체계에서 상시 예방적 감시 체계로 전환 • 사용 시스템 단계부터 검증 기능을 운영하여 생애주기 전반 연속성 확보 • 무결성 검증, 포맷 검증, 오류 대응을 단일 워크플로로 통합 • PREMIS 이벤트 중심의 보존 이력 기록을 표준화하여 책임 추적성 확보 • 자동화 · 정기화 · 로그 기반 관리 체계로 전환 • 국제 선도 기관 수준의 자동화 검증 체계 도입 지향
적용 방안	**단기**	• 모든 보존 대상 파일에 대해 체크섬(SHA-256 이상) 생성 의무화 • 정기적 무결성 검증(Fixity Check) 실행 체계 수립 • 포맷 식별(DROID 등) 및 포맷 유효성 검증 도입 • 검증 결과를 PREMIS 이벤트로 기록 · 누적 관리 • 오류 발생 시 수동 대응 절차 문서화 • 검증 대상 · 주기 · 책임 부서 명확화
	중기	• 무결성 검증 주기 자동 스케줄링 시스템 구축 • 검증 실패 시 관리자 자동 알림 체계 도입 • 대체본 자동 복구(백업 · 복제본 참조) 워크플로 구현 • LTFSck, Read Test 등 저장계층 점검 절차 통합 • 포맷 유효성 검사 결과와 PREMIS 이벤트 연계 자동화 • 검증 로그 통합 대시보드 구축 • 위험도 기반 차등 검증 정책 도입(고위험 자료 우선 점검)
	장기	• 스토리지 상태 · 파일 접근 패턴 · 오류 발생 이력 분석 기반 예측 모델 도입 • AI 기반 보존 위험도 평가 체계 구축 • 잠재적 손상 가능성 사전 탐지 체계 구현 • 자동 복구 정책 및 자율 대응 시스템 고도화 • 저장계층(HDD, SSD, LTO 등)별 위험 관리 통합 모델 확립 • ISO 16363 신뢰성 기준과 연계한 무결성 관리 체계 제도화

[표 47] [전략 2] 무결성 검증과 보존 이력 관리 체계 구축

❑ [전략 3] 보존용 메타데이터 및 포맷 정책의 제도화

<table>
<tr><td colspan="2">필요성</td><td>장기보존의 안정성은 저장 인프라뿐 아니라, 자원의 구조적 정의와 기술적 표준의 명확성에 의해 좌우된다. 국립중앙도서관에는 자료 유형별 보존 메타데이터 표준, 테크니컬 메타데이터 정의, 우선 포맷 목록(Preferred Formats), 포맷 위험 관리 기준 등을 통합적으로 관리 · 갱신할 수 있는 제도적 체계가 필요하다. 현재는 사용과 보존이 공존하는 환경에 적합한 데이터 구조 정책과 포맷 관리 정책이 명확히 제도화되어 있지 않으며, 장기보존 단계로의 전환을 전제로 한 표준 연계 체계도 미흡한 상황이다. 이에 따라 메타데이터 · 포맷 · 데이터 구조를 통합한 보존 정책 체계를 수립하고, 이를 OAIS 및 PREMIS 구조와 연계성을 확보하는 것이 필수적이다.</td></tr>
<tr><td colspan="2">수행방향</td><td>• 사용 단계와 보존 단계 모두에 적용 가능한 통합 데이터 관리 모델 수립
• 특정 원본과 사본, 파생기록물 전체가 세트로 운영될 수 있는 데이터구조 정의
• 테크니컬 메타데이터 자동 생성 · 관리 체계 도입하고, 우선 포맷 목록(Preferred Format List) 및 위험 포맷 목록 관리 체계 수립
• PREMIS 메타데이터 표준과 정합성을 유지하는 보존 이력 체계 구축
• 글로벌 표준(MARC, DC, MODS, PREMIS 등)과의 정합성 확보</td></tr>
<tr><td rowspan="3">적용 방안</td><td>단기</td><td>• 자료 유형별 메타데이터 구성요소 정의(서지 · 구조 · 기술 · 보존 메타데이터)
• 컬렉션 구조 및 파생기록물 관리체계 수립
• 아이템 레이어의 다중 메타데이터 표준 수용 체계 마련
• 자원 유형별 테크니컬 메타데이터 상세 지침 제정
• 우선 포맷 목록(Preferred Formats) 1차 확정
• 포맷 위험도 평가 기준 수립
• SIP 전환 시 필수 메타데이터 매핑표 작성</td></tr>
<tr><td>중기</td><td>• 수집 단계에서 자동 포맷 식별(DROID 등) 도입
• 포맷 식별 결과와 정책 매핑 기능 구현
• 포맷 적합성 평가 자동화 체계 구축
• 포맷 변경 · 마이그레이션 정책 수립
• PREMIS 기반 포맷 이벤트 기록 자동화
• 메타데이터 검증(Well-Formed, Schema Validation) 자동화 도입
• 우선 포맷 목록의 정기적 갱신 체계 운영</td></tr>
<tr><td>장기</td><td>• 국가 차원의 포맷 레지스트리 구축 검토
• 포맷 위험 예측 및 장기 가독성 평가 체계 도입
• 국내 공공기관 간 포맷 정책 공유 플랫폼 구축
• OAIS SIP-AIP-DIP 단계와 자동 연계되는 포맷 정책 시스템 완성
• 국제 표준 기반 상호운용성 확보 및 신뢰성 체계 정착</td></tr>
</table>

[표 48] [전략 3] 보존용 메타데이터 및 포맷 정책의 제도화

❑ [전략 4] 다계층·분산 스토리지 확장

필요성		디지털자원의 지속적 증가와 장기보존 요구 확대에 대응하기 위해서는 운영용 저장소와 장기보존 저장소를 명확히 구분한 다계층 스토리지 구조가 필요하다. 단일 센터 중심 구조는 물리적 장애, 재난, 사이버 공격, 시스템 오류 발생 시 대규모 데이터 손실 위험을 내포하고 있다. 특히 장기보존 체계는 단순 저장 용량 확보가 아니라, 데이터 생존 가능성을 보장하는 구조적 안정성을 전제로 한다. 이에 물리적 · 지리적으로 분산된 저장 인프라를 구축하고, 계층별 저장 정책을 명확히 하는 것이 국가 디지털 자산 보호를 위한 필수 조건이다.
수행방향		• 스토리지를 단순 IT 인프라가 아닌 국가 디지털 자산 보호 수단으로 재정의 • 운영(Storage for Access)과 보존(Storage for Preservation) 계층의 명확한 분리 • 고가용성(Availability)과 복원력(Resilience)을 핵심 설계 원칙으로 설정 • 단일 장애 지점(SPOF) 제거를 목표로 분산 구조 확립 • 재난 · 사이버 공격 · 물리적 사고를 전제로 한 위험관리 기반 설계 • 국가 디지털 자료 보존관 개념과 연계한 장기적 분산 거점 전략 수립
적용 방안	**단기**	• 기존 스토리지를 운영 영역과 보존 논리 영역으로 분리 • 보존 대상 데이터의 이중 복제(Primary-Secondary) 구조 확보 • 저장계층별 데이터 분류 정책 수립(운영/중간보존/영구보존) • 무결성 검증과 연계된 저장계층 관리 체계 도입 • RAID, Snapshot, 백업 정책 정비 • 스토리지 자원 모니터링 체계 강화
	중기	• 오프사이트 백업 체계 구축(지리적 분리 저장소 확보) • LTO 기반 콜드 스토리지 또는 클라우드 기반 보존 스토리지 도입 • Hot-Warm-Cold 계층 구조 확립 • 스토리지 계층 간 자동 데이터 이동 정책(ILM: Information Lifecycle Management) 도입 • LTFS 기반 장기보존 테이프 운영 체계 확립 • 스토리지 장애 시 자동 복구 시나리오 구축 • 재해복구(Disaster Recovery) 훈련 정례화
	장기	• 국가 차원의 분산 저장 거점 구축 및 다중 지역 연계 구조 확립 • 단일 기관 장애가 국가 보존 실패로 이어지지 않는 다중 복제 구조 완성 • 국가문헌보존관과 연계한 영구보존 전용 인프라 구축 • 에너지 효율 · 장기 비용 최적화 고려한 지속가능 저장 전략 도입 • 사이버 보안 및 랜섬웨어 대응을 포함한 보존 보안 통합 모델 구축 • ISO 16363 신뢰성 기준에 부합하는 저장 인프라 고도화

[표 49] [전략 4] 다계층 · 분산 스토리지 확장

❑ [전략 5] 조직·거버넌스 재편

필요성		디지털 장기보존 전략은 기술적 설계만으로 실현될 수 없으며, 이를 지속적으로 운영 · 관리할 수 있는 조직적 기반을 전제로 한다. 보존은 단일 부서의 기술 업무가 아니라 정책 수립, 표준 관리, 인프라 운영, 품질 관리, 위험 대응이 통합된 복합적 과제이다. 현재 조직 구조에서는 보존 기능이 IT 운영 또는 개별 사업 단위에 분산되어 있어, 장기적 책임성과 전략적 일관성을 확보하는 데 한계가 존재한다. 따라서 보존 전담 조직과 명확한 책임 구조를 정립하고, 거버넌스 체계를 제도화하는 것이 장기보존 전략 실행의 핵심 전제 조건이다.
수행방향		• 장기보존을 IT 지원 기능이 아닌 국가 지식 인프라의 핵심 기능으로 재정의 • 보존 전담 조직 또는 센터의 위상을 제도적으로 강화 • 정책–기술–운영–품질 관리 기능을 통합하는 거버넌스 구조 수립 • OAIS 기반 기능 구조와 조직 구조를 정합적으로 정렬 • 책임(R&R) 명문화 및 의사결정 체계 일원화 • 보존분야의 협력 체계에서 허브 기관 역할 수행을 지향
적용 방안	**단기**	• 기존 조직 내 보존 전담 기능의 명확한 구분 및 업무 재정의 • 보존 책임 범위 및 역할(R&R) 문서화 • 보존 정책 · 무결성 관리 · 포맷 정책 기능을 담당하는 실무 전담 인력 지정 • 보존 관련 내부 협의체 구성(정책 · 기술 · 운영 통합 회의체) • 보존 KPI(성과 지표) 설정 및 정기 보고 체계 도입 • 외부 전문가 자문 체계 구축
	중기	• 보존 정책 · 기술 · 품질 관리 역할을 통합한 전담 조직(센터) 구성 • Archival Storage 및 Preservation Planning 기능의 조직 내 독립성 확보 • 보존 전담 예산 항목 분리 및 안정적 재원 확보 • 인력 전문성 강화 프로그램 운영(OAIS, PREMIS, ISO 16363 등) • 내부 감사 및 신뢰성 점검 체계 구축 • 대외 협력 창구 일원화
	장기	• 국가 차원의 디지털 보존 협력 거버넌스 허브 기관으로 역할 확장 • 공공기관 · 대학 · 연구기관과의 보존 협력 네트워크 구축 • 국가 디지털 보존 정책 수립 및 표준화 주도 • 국제 보존 네트워크 참여 및 표준 제정 협력 강화 • 국가문헌보존관 운영을 총괄하는 상위 거버넌스 구조 확립 • ISO 16363 등 국제 신뢰성 인증 추진

[표 50] [전략 5] 조직 · 거버넌스 재편

❑ [전략 6] AI 기반 고도화 전략

<table>
<tr><td colspan="2">필요성</td><td>국립중앙도서관의 디지털자원은 규모와 유형 측면에서 지속적으로 확대되고 있으며, 파일 포맷의 다양화, 대용량 시청각 자료 증가, 복합 객체의 확산 등으로 관리 복잡성이 급격히 증가하고 있다. 기존의 인력 중심 · 규칙 기반 보존 운영 체계만으로는 이러한 환경 변화에 대응하기 어려우며, 무결성 검증 · 포맷 관리 · 위험 예측 등 고도화된 분석 기능을 상시 수행하기에는 구조적 한계가 존재한다. 특히 장기간 축적되는 PREMIS 이벤트, 무결성 로그, 접근 이력 등은 단순 기록을 넘어 분석 자산으로 활용될 수 있음에도, 이를 체계적으로 학습 · 활용하는 환경은 아직 미비하다. 이에 따라 AI 기반 자동화 · 지능화 기술을 수용할 수 있는 기술적 · 조직적 기반을 구축하고, 이를 OAIS 기반 보존 체계와 정합적으로 연계하는 전략이 요구된다.</td></tr>
<tr><td colspan="2">수행방향</td><td>• 수집 · 인수(Ingest) – 보존 저장(Archival Storage) – 보존 전략 기획(Preservation Planning) – 접근 · 서비스(Access) 전 영역에 단계적 적용
• 사후 대응 중심 운영에서 예측 · 선제적 관리 체계로 전환
• PREMIS 이벤트 및 무결성 로그를 학습 데이터로 활용하는 구조 설계
• 인간 의사결정을 보완하는 “설명 가능한 AI(Explainable AI)” 지향
• 알고리즘 책임성(Accountability) 및 감사 가능성 확보
• 데이터 기반 보존 의사결정 문화 정착</td></tr>
<tr><td rowspan="3">적용 방안</td><td>단기</td><td>• 수집(Ingest) 단계에서 AI 기반 파일 포맷 자동 식별 고도화
• 메타데이터 자동 추출(OCR, STT, 객체 인식 등) 시범 적용
• 중복 파일 · 손상 파일 자동 탐지 기능 도입
• SIP 품질 점검 자동화 도구 도입
• Archival Storage 영역에서 무결성 로그 · 접근 이력 분석 기반 위험 징후 탐지 시범 운영
• AI 결과를 운영자 검토 후 반영하는 “인간-검증(Human-in-the-loop)” 구조 적용</td></tr>
<tr><td>중기</td><td>• PREMIS 이벤트, 포맷 변환 이력, 무결성 검증 로그를 학습 데이터로 활용
• 포맷 노후화 예측 모델 개발
• 마이그레이션 우선순위 자동 제안 시스템 구축
• 위험 포맷 자동 식별 및 경보 체계 구현
• Preservation Planning 기능과 연계한 AI 의사결정 지원 시스템 구축
• 스토리지 상태 · 에러 로그 기반 장애 발생 가능성 예측 모델 도입</td></tr>
<tr><td>장기</td><td>• OAIS 기반 보존 아키텍처 전반에 AI 기능 내재화
• 보존 리스크 종합 예측 모델 구축(포맷 · 스토리지 · 접근 패턴 통합 분석)
• 자동 대응 시나리오 생성 및 복구 경로 최적화 기능 구현
• 분산 스토리지 환경에서 복원 우선순위 자동 계산
• 보존 데이터 공유 기반 AI 협력 모델 구축
• AI 의사결정 설명 가능성 및 감사 체계 제도화
• 윤리 · 보안 기준을 반영한 AI 거버넌스 체계 확립</td></tr>
</table>

[표 51] [전략 6] AI 기반 고도화 전략

5.3. 장기보존전략 흐름도

5.3.1. 실행 로드맵 수립

❑ 단기 실행계획 : 보존 책임 구조와 최소 신뢰성 확보 단계

단기 실행단계는 국립중앙도서관의 디지털자원 관리 체계를 장기보존체계로 전환하기 위한 출발점으로서 기술 구축보다 정책·개념·운영 논리의 정비가 핵심 목표가 된다. 이 시기의 가장 중요한 과제는 OAIS(Open Archival Information System) 참조모델을 국립중앙도서관 장기보존의 공식 운영 원칙으로 명확히 선언하고 장기보존의 책임 범위와 관리 단위를 내부적으로 합의·제도화하는 것이다.

이를 위해 SIP(Submission Information Package)–AIP(Archival Information Package)–DIP(Dissemination Information Package)의 역할을 명확히 정의하고 특히 AIP(Archival Information Package)를 보존 책임이 발생하는 최소 관리 단위로 규정함으로써 이후 모든 보존전략의 기준점을 마련한다.

무결성 관리 측면에서는 모든 보존대상 디지털 객체에 대해 체크섬을 생성하고 Strategies) 이벤트를 활용해 무결성 검증 결과를 기록하는 체계를 전면 적용한다. 이는 자동화 이전 단계이지만, 장기적으로 가장 중요한 보존 이력 데이터가 축적되기 시작하는 시점이라는 점에서 전략적 의미가 크다.

동시에 보존 메타데이터 체계 전환을 위한 기반 작업으로 PREMIS(Preservation Metadata: Implementation Strategies) 핵심 요소를 시범 적용하고 자료유형별 장기보존 우선 포맷 목록을 내부 지침으로 확정하여 수집 단계부터 보존 적합성 기준이 적용되도록 하는 것도 단기 단계의 주요 과제이다.

조직 측면에서는 장기보존이 특정 담당자의 부수 업무가 아니라 기관 차원의 공식 책임 영역임을 선언하고 전담 기능과 책임을 명확히 구분한다. AI 전략은 이 단계에서 본격 도입보다는 적용 가능 영역을 식별하고 내부 가이드라인을 마련하는 준비 단계로 위치한다.

❑ 중기 실행계획 : OAIS(Open Archival Information System) 기반 보존체계의 구현·자동화 단계

중기 실행단계는 단기 실행단계에서 정립된 개념과 정책을 실제 시스템과 운영 환경에 구현하는 핵심 전환기이다. 이 시기에는 OAIS(Open Archival Information System) 기반 보존 논리가 더 이상 선언적 수준에 머무르지 않고 실제 데이터 흐름과

업무 절차에 내재화되는 것을 목표로 한다.

우선 AIP(Archival Information Package) 전용 저장소와 관리 모듈을 구축하여, SIP(Submission Information Package) 수집 이후 디지털 객체가 자동으로 AIP(Archival Information Package)로 생성・검증・등록되는 워크플로를 구현한다. 이를 통해 보존 영역과 서비스 영역은 논리적으로 분리되지만, 서비스는 항상 보존 객체(AIP)를 참조하는 구조로 전환된다. 이는 국립중앙도서관 장기보존체계가 실질적으로 작동하기 시작하는 결정적 단계라 할 수 있다.

무결성 관리 영역에서는 무결성 검증(Fixity check)을 자동화하여, 시스템이 검증 주기를 관리하고 결과를 누적 기록하도록 전환한다. 오류 발생 시 관리자 알림, 대체본 복구, 재검증으로 이어지는 일련의 대응 절차가 운영 규칙으로 정착된다. PREMIS(Preservation Metadata: Implementation Strategies) 이벤트는 이 단계에서 단순 기록을 넘어 보존 의사결정의 근거 데이터로 활용되기 시작한다.

스토리지 전략 역시 중기 단계에서 본격적으로 확장된다. 오프사이트 백업과 테이프 또는 클라우드 기반 콜드 스토리지를 도입하여, 단일 저장소 의존 구조에서 벗어나 다계층・분산 저장 구조를 갖추게 된다. 이는 재난・시스템 장애・사이버 위협에 대비한 실질적 복원력을 확보하는 과정이다.

포맷 정책 측면에서는 자동 포맷 식별 기능을 도입하고 식별 결과를 장기보존 포맷 정책과 연계하여 위험 포맷을 체계적으로 관리한다.

조직적으로는 보존 정책・기술・품질 관리를 통합하는 전담 조직을 구성하고 전문 인력을 확충한다. AI 전략은 이 단계에서 보존 데이터 분석을 통한 포맷 노후화 예측, 마이그레이션 우선순위 제안 등 보존 의사결정 지원 기능으로 제한적으로 도입된다.

❑ 장기 실행계획 : 지능형·국가 허브형 장기보존체계 완성 단계

장기 실행단계는 국립중앙도서관의 장기보존체계가 성숙 단계에 도달하여, 국가 디지털 장기보존 허브로 기능하는 단계이다. 이 시기에는 OAIS(Open Archival Information System) 전주기(SIP-AIP-DIP)가 완전 자동화된 운영 체계로 안정화되며, 외부 감사와 신뢰 저장소 인증(예: ISO 16363)을 검토할 수 있는 수준의 투명성과 신뢰성을 확보하게 된다.

무결성 관리와 보존 계획 영역에서는 AI 기반 예측 기술이 본격적으로 활용되어, 파일 손상 가능성, 포맷 노후화 위험, 스토리지 장애 가능성을 사전에 분석하고

선제적으로 대응하는 지능형 보존체계로 발전한다. 이는 장기보존을 사후 복구 중심 구조에서 예방·예측 중심 구조로 전환하는 결정적 전환점이다.

스토리지 구조는 국립중앙도서관 내부를 넘어 국가 차원의 분산 저장 거점과 연계되며, 단일 기관 장애가 국가 디지털 자산 손실로 이어지지 않는 구조적 안전성을 확보하게 된다. 포맷 정책과 메타데이터 체계 역시 국가 차원으로 확장되어, 포맷 레지스트리 구축 또는 국제 레지스트리와의 연계를 통해 자동화된 보존전략 수립이 가능해진다.

조직·거버넌스 측면에서는 국립중앙도서관이 국가기록원, 문화예술 아카이브 기관 등과 협력하여 국가 디지털 보존 협력 체계를 주도하는 중심 기관으로 자리매김하게 된다. AI 활용 역시 윤리·책임 거버넌스를 포함한 제도적 틀 안에서 안정적으로 운영하는 단계이다.



[그림 43] 국립중앙도서관 디지털자원 보존 실행 로드맵

❑ 종합 정리

국립중앙도서관의 디지털자원 장기보존 실행계획은 단기-중기-장기의 연도별 단계에 따라 보존 책임 구조 확립 및 사용보존 공존체계 구현 → OAIS(Open Archival Information System) 기반 체계 구현 → 지능형·국가 허브형 보존체계 완성이라는 점진적 발전 경로를 따른다. 이러한 단계적 실행계획은 기술 중심 접근이 아니라 정책·조직·운영 논리의 성숙을 전제로 한 현실적인 이행 경로로서, 국립중앙도서관이 국가 디지털 기억을 장기적으로 보존·보증하는 핵심 기관으로 도약하기 위한 실천적 전략이라 할 수 있다.

제6장 결론 및 제언

6.1. 결론

국립중앙도서관 현황분석, 보존·복원 영역의 수요조사, 국가문헌보존관 체계 분석을 통하여, 국립중앙도서관 디지털자원 보존체계의 특징과 방향성은 도서관 자원 생명주기 관리 전체 프로세스 중 하나의 구성요소로서 접근할 필요가 있음을 확인하였다. 국립중앙도서관의 보존체계는 현행 KOLIS 중심 서비스 시스템과 연계된 상태에서 운영되어야 하며, 사용과 보존이 동시에 이루어지는 환경을 전제로 한다. 따라서 먼저 자원의 신뢰성과 품질을 유지하는 보존 기능을 구축하고, 이후 단계적으로 OAIS 기반 영구보존 시스템으로 이관하는 이중 구조를 고려해야 한다. 이러한 관점은 국립중앙도서관의 보존 체계를 단순히 '단일 영구 보존 저장소' 구축으로 귀결시키기보다, 자원 관리의 전 과정에 보존 활동이 선행적으로 내재화되는 구조를 지향할 필요성이 있음을 시사한다. 즉, 일상적인 서비스 운영 및 관리 흐름 속에서 보존 요건을 충족한 자원이 단계적으로 영구 보존으로 이행되는 다단계 보존 관리 체계(Multi-tier Preservation Framework)로의 전환이 이루어져야 함을 의미한다.

국립중앙도서관의 디지털 장기보존체계는 이러한 특성을 고려할 때, 수집과 서비스 중심 구조에서 출발하여 점진적으로 보존 기능을 확대해 온 단계적 발전 과정에 놓여 있었다. 그러나 현 단계에서는 서비스 중심 시스템과 보존 중심 시스템 간 협업·연계 구조에 대한 정의 부족과 경계·접점의 부재, 업무 흐름의 분절, 저장소 및 데이터 구조의 물리적·의미적 한계, 사용과 보존이 공유해야 할 표준의 부재, 저작권 제약, 신기술 연계의 한계라는 구조적 문제를 동시에 가지고 있다. 따라서 국립중앙도서관의 기본 책무인 국가 지식정보 자원의 관리와 활용을 우선하면서도 국가 차원의 실질적 디지털자원 장기보존 기관으로 기능하기 위해서는 기술 인프라 구축과 함께 거버넌스 개편, 법·제도 정비, 표준 기반 운영체계 확립, 보존 전문 인력 확충이 통합적인 추진이 필요하다. 이를 위한 핵심 전략으로는 첫째 현재 KOLIS Ⅲ 시스템 내 포함 되어 있는 장기보존 시스템을 OAIS(Open Archival Information System) 기반의 독립적인 영구보존 시스템으로 분리·구축하는 것, 둘째 사용과 보존이 공존하는 시스템 환경에서 무결성 검증과 보존 이력 관리 체계를 구축하여 일상적·정기적으로 무결성 점검 절차를 이행하는 것, 셋째 다계층 분산 스토리지 및 백업 체계를 확장하는 것, 넷째 장기보존

포맷 정책을 제도화하여 자료유형별 장기보존 포맷과 메타데이터 관리를 위한 포맷 레지스트리를 운영하는 것, 다섯째 생성형 인공지능 기반의 보존체계 및 기술 고도화 전략을 추진하는 것이다.

이러한 전략을 구체화하기 위하여 국립중앙도서관 디지털자원에 적합한 보존 전략 선택 기준을 제시하였다. 선택 기준은 자료의 논리적 구조와 표현 복잡성, 이용자의 접근성과 이용 지속성, 비트 수준의 보존 안정성, 보존 책임의 증거성과 신뢰성, 저작권・법적 제약과 기술적 실현 가능성에 근거하여 설정되었으며, 이를 바탕으로 디지털 온라인자료유형별 보존전략을 핵심 보존전략과 보조적 보존전략으로 구분하였다. 그 결과 포맷정규화(Normalization), 매체전환(Refreshment), 마이그레이션(Migration)은 디지털자원이 데이터 자체의 진본성과 유용성을 유지하면서 '살아있는 상태'로 운영될 수 있도록 하는 보존전략으로서 필수 절차에 해당함을 확인하였다. 또한 자료유형별 보존전략 수립을 위해 전자책・전자저널・학위논문(PDF, ePub 등), 디지털화 인쇄자료 이미지(TIFF, JPEG 등), 웹자원(OASIS 수집 웹사이트, WARC 등), 시청각자료(영상, 음원), 데이터세트・연구데이터(CSV, XML, DB 등), 특수 디지털 객체(인터랙티브 콘텐츠, 구형 소프트웨어 등)별 선택 기준과 적용방안을 제시하였다.

안정적 보존을 위한 운영적 절차로서 보존 파일・매체・메타데이터에 대한 정기 점검체계를 NDSA 레벨 모델의 요구사항을 기준으로 제시하고 무결성 검증(Fixity Check), 파일 포맷 유효성 점검, 저장매체 상태 점검, 메타데이터 오류점검 방안을 함께 제안하였다. 특히 매체전환(Refreshment)의 대부분을 구성하는 무결성 검증 방안으로 SHA-256, SHA-512와 같은 체크섬 알고리즘 기반 무결성 검증(Fixity Check) 적용방안을 상세히 제시하였는데, 무결성 검증(Fixity check)를 단일 기능이 아닌 정기・수시적으로 수행되는 절차로 규정하고 절차에 포함되는 포맷 유효성 검사, 저장매체 상태 점검, 메타데이터 오류점검을 스토리지 저장 단계별로 실시하는 방안과 점검 절차별 상세 프로세스 및 사용 가능한 도구를 제시하였다. 또한 이러한 무결성 검증 기능과 절차는 보존 단계에만 한정되지 않고 사용 단계에서도 저장매체 노후화로 인한 비트 오류(BIT rot), 하드웨어 오류, 무단 접근에 따른 데이터 손상 가능성을 조기에 탐지하기 위해 적용할 필요성이 높다는 점을 함께 제시하였다.

스토리지와 백업에 관한 기술 구성은 매체 유형에 따른 스토리지 적용성, Local Storage・DAS・NAS・SAN 등 연결방식에 따른 적용성, 파일 스토리지(File storage)・블록 스토리지(Block storage)・오브젝트 스토리지(Object storage) 등 파일

접근방식에 따른 적용성 검토를 통해 제시하였다. 그 결과 국립중앙도서관 디지털 자원 보존 스토리지 적용방안으로 계층형 스토리지 구조 설계 필요성과 파일 스토리지 방식 적용방안을 도출하고 디지털자원의 생애주기에 적합한 단계별 적용 방안을 제시하였다. 백업 인프라 구성은 RAID, Disk-to-Disk, Virtual-Tape-Library, LTO(Linear Tape-Open) Tape-backup, Off-site-backup, 원격지 백업 방안 분석을 통해 국립중앙도서관에 적합한 6단계 다계층 백업 구조로 제시하였으며, 이를 통해 일상적 장애부터 대규모 재난까지 위험 시나리오 전반에 대응하는 보존·복구 체계를 목표로 제시하였다.

인공지능 적용 영역에서는 생성형 인공지능 기반 디지털자원 관리 및 보존·복원 관점에서 제목 자동 생성, 메타데이터 자동 생성, 자동 분류·색인(주제 분류, 색인어 추천), 멀티모달 요약(문서·이미지·영상·음성), 포맷 분석 및 장기보존형 변환과 손상·위험 탐지, OCR/STT 전사 및 자동 정체, 중복·유사 디지털자원 식별, 포맷 식별 자동화, 파일 진단과 오류 자동 탐지 및 경고 서비스, 문서·이미지·영상·음원 복원 및 품질 개선 모델을 제시하였다.

결론적으로 본 연구를 통해 도출한 체계와 기준을 ‘사용과 보존이 공존하는 시스템’ 아키텍처 설계에 적용함으로써, 일상적 무결성 검증 수행, 효율적인 다계층 분산 스토리지 구성, 다양한 파생기록물의 컬렉션 구축, 다중 메타데이터 표준과 강화된 테크니컬 메타데이터 표준 적용, OAIS(Open Archival Information System) 시스템의 SIP(Submission Information Package) 및 PREMIS(Preservation Metadata: Implementation Strategies)로의 신속한 데이터 이관이 보장될 수 있는 국립중앙도서관 디지털자원 보존관리시스템 구현을 위한 가이드라인으로 활용되기를 기대한다.

6.2. 제언

보존 시스템의 저장 인프라 운영 고도화 영역에서는 계층형 스토리지 구조를 장기보존 운영의 기본 아키텍처로 정착시킬 필요가 있다. 각 저장 계층의 역할과 데이터 이동 조건을 명확히 정의함으로써, 저장매체 간 전환이 임의적으로 이루어지는 것을 방지하고 장기보존 관점에서 예측 가능한 운영 체계를 구축할 수 있을 것이다. 특히 온라인 계층과 오프라인 계층을 물리적으로 분리하되, 관리 메타데이터를 통합적으로 운영하는 방식은 유지·확장 측면에서 중요한 발전 방향이 될 것이다.

그리고 파일 스토리지 인프라의 디지털자원 생애주기 단계별 적용방안을 효과적으로 적용하기 위해서는 NAS, SAN 또는 DAS 기반 스토리지와 오브젝트 스토리지의 병렬 구성에 따른 불필요한 중복 스토리지 발생의 최소화, 서비스 성능의 최적화, 보존의 안정성 확보, 보존 시스템의 경제성 확보를 위한 스토리지 현황 모니터링 및 분석 방안의 수립을 위한 연구가 필요할 것이다.

보존 포맷 및 점검 절차의 제도화 영역에서는 보존 포맷 정책과 점검 절차는 일회성 지침이 아니라 지속적인 운영 규칙으로 내재화될 필요가 있다. 포맷 유효성 검사, 무결성 검증, 저장매체 상태 점검 결과를 PREMIS(Preservation Metadata: Implementation Strategies) 이벤트로 체계적으로 기록하고 장기적으로 축적된 이력을 기반으로 포맷 위험도와 매체 상태를 평가하는 관리 체계를 강화해야 한다. 이는 장기보존을 기술적 행위가 아닌 관리 행위로 정착시키는 데 핵심적인 역할을 할 것이다.

선제적 보존 관리 체계 구축 영역에서는 점검·검증 결과를 기반으로 한 선제적 대응 체계를 강화할 필요가 있다. 장기보존 환경에서는 기술 변화와 물리적 열화가 불가피하므로, 문제 발생 이후의 사후 복구보다는 초기 징후를 탐지하여 조기에 대응하는 방식이 중요하다. NDSA는 일상적인 시스템 접근 제어 및 보안을 비롯하여 스토리지와 물리적 저장소의 상태점검 (Media health check), 무결성 및 데이터 진본성 점검 (Fixity & Data integrity check), 메타데이터 오류점검 (Metadata check), 파일 포맷 유효성 (File format check)의 5개 영역에 대해서 레벨 별로 점검 절차를 제시하고 있다. 이들을 사용단계부터 선제적 관리 체계를 구현하기 위한 핵심 요소로 지정하여 보존 정책과 시스템 설계에 지속적으로 반영할 필요가 있다.

AI 활용 범위 및 관리 기준 정립 영역에서는 AI 기술의 적용은 단계적으로 확대

하되, 그 관리 기준을 명확히 설정할 필요가 있다. AI 산출물을 파생 메타데이터로 분류하고 생성·갱신·폐기에 대한 관리 기준을 설정함으로써 기존 보존 메타데이터 체계와의 정합성을 유지해야 한다. 이를 통해 AI 기술은 장기보존체계의 안정성을 저해하지 않으면서도, 관리 효율성과 서비스 품질을 향상시키는 보조 수단으로 기능할 수 있을 것이다.

데이터 구조 및 메타데이터 영역에서는 영구보존 패키지의 자료 구성 방안에 대한 보다 구체적인 규정과 지침의 개발이 필요함을 제시하였다. 영구보존 패키지는 PREMIS(Preservation Metadata: Implementation Strategies)의 엔터티 구성 원칙을 준수하되, 원본과 서비스본을 포함한 다양한 파생기록물들이 사용 당시의 관계 구조를 유지한 채 패키징되는 것이 적절하며, 이 과정에서 자료조직에 활용된 분류, 전거, 코드 및 스킴이 식별 가능한 상태로 유지되어야 한다. 하나의 SIP(Submission Information Package)는 단일 지적 객체 또는 보존 단위에 속하는 모든 파생기록물을 하나의 구성 단위로 묶음으로써, 의미적·물리적 관계를 지닌 자원들의 체계적 보존과 접근권한, 저작권 관리, 기술적 이용 및 재현 과정에서의 오류를 방지할 수 있다. 반면, 원본과 파생기록물 간에 공개 여부나 보존 여부가 현저히 다른 경우에는 SIP(Submission Information Package)를 분리 구성하는 것이 적절할 수 있으며, 개체별 관리 정보가 명확하게 식별되는 데이터 구조에서는 단일 SIP (Submission Information Package) 구성이 가능할 수 있다. 다만 파일 개체 단위 관리가 불가능하여 파일별 SIP(Submission Information Package)를 구성할 경우, 무결성 검증의 부담 증가와 AIP(Dissemination Information Package)의 과도한 세분화로 인해 관계 정보 관리 구조가 복잡해지는 문제를 방지할 수 있는 방안을 강구해야 한다.

끝으로 사용과 보존이 공존하는 시스템 환경에서는 아직 국내 기록관리 분야에서 경험하지 않은 새로운 이슈에 대한 대응 방안 수립이 필요할 것이다. 사용과 보존이 병행될 경우, 기록관리 분야에서 적용되어 온 기록물 유형별 보존 기간과 획일적인 이관 조건에 따른 기계적 이관이 어려울 수 있으며, 장기보존 단계에서 전제가 되는 기록물의 안정성과 불변성이 사용 단계의 인용, 참조, 주석 활동으로 인해 훼손될 가능성도 고려해야 한다. 이에 따라 SIP(Submission Information Package) 및 AIP(Dissemination Information Package)의 패키징과 재패키징이 실행되는 조건과 기준, 사용 단계와 보존 단계에 동시에 걸쳐 있는 자원의 처리에 대한 의사결정 기준, 그리고 AIP(Archival Information Package) 재작업을 전제로 한 보존관리 업무 지침과 시스템 및 데이터 아키텍처의 개발이 반드시 필요하다.

참 고 문 헌

국가기록원 (2023). 시청각 유형별 관리 프로세스 재정립 및 메타데이터 체계 재구축 연구.

국립중앙도서관 (2017). 국립중앙도서관 국가문헌보존관 건립 기본연구.

국립중앙도서관 (2018). 국립중앙도서관 디지털자원 보존전략 수립 연구.

국립중앙도서관 (2018). 국가도서관 장서 디지털화 및 디지털 보존에 관한 정책 사례 조사.

국립중앙도서관 (2020). 디지털자료 보존을 위한 보존패키지 분석결과 보고.

국립중앙도서관 (2021). 국립중앙도서관 국가문헌보존관 종합운영계획 연구.

국립중앙도서관 (2022). 시청각자료의 디지털화 보존을 위한 데이터 구축 시범연구.

국립중앙도서관 (2022). KOLISIII 온라인 사용자지침서.

국립중앙도서관 (2024). 2024 국립중앙도서관 연보.

국립중앙도서관 (2024). 2024 국립중앙도서관 온라인자료 정리 지침 (2025.04 개정).

국회도서관 (2014). 정보 공유를 위한 메타데이터 표준 및 활용 방안 연구.

행정안전부 (2019). 공공기관 데이터 자원 보존 개선방안 연구.

국립세계문자박물관 (2024). 국립세계문자박물관 디지털아카이브 정보화 전략계획(ISP).

강민정 (2021). OAIS 참조모형 기반 민중가요 메타데이터 설계에 관한 연구, 한국기록관리학회지, 22(1), pp. 211-230, https://doi.org.10.14404/JKSARM.2021.21.1.211

박옥남 (2012). PREMIS 기반 보존 메타데이터 요소 개발에 관한 연구, 한국문헌정보학회지, vol.46(2), pp. 83-113, https://doi.org/10.4275/KSLIS.2012.46.2.083

박진호 (2021). DDI 메타데이터를 활용한 METS 설계에 관한 연구, 정보관리학회지, vol.38(4), pp. 153-171, https://doi.org/10.3743/KOSIM.2021.38.4.153

박희진 (2020). 디지털 보존을 위한 PREMIS 3.0 기반 메타데이터에 관한 연구, 한국기록관리학회지, vol.20,(4), pp. 185-203, https://doi.org/10.14404/JKSARM.2020.20.4.185

소정의 (2018). 국외 전자기록물의 장기보존 정책 비교 분석, 한국기록관리학회지, vol.18(4), pp. 125-148, https://doi.org/10.14404/JKSARM.2018.18.4.125

오상훈 (2009). ISO14721 OAIS 참조모형을 활용한 웹아카이빙의 메타데이터 구조 및 요소 정의, 정보처리학회논문지, vol.16(5), pp. 651-660

오선경 (2024). 국립중앙도서관의 장서개발 전략 수립에 관한 연구, 정보관리학회지, vol.42(2), pp. 29-56, https://doi.org/10.3743/KOSIM.2025.42.2.029

오세라 (2016). 공개포맷에 기반한 전자기록 보존 포맷 재설계 방향 연구, 한국기록관리학회지, vol.16(4), pp. 79-120, https://doi.org/10.14404/JKSARM.2016.16.4.079

이소연 (2012). 국내 디지털 보존 연구의 동향 분석, 한국기록관리학회지, vol.13(2), pp. 247-283

이승민 (2015). 디지털 아카이빙을 위한 보존 메타데이터 패키지 구축, 정보관리학회지, vol.32(3), pp. 21-47, https://doi.org/10.3743/KOSIM.2015.32.3.021

Bermès. E. (2010). A data-first approach to digital preservation: the SPAR project.

Bermès. E. (2008). Semantic Web technologies for digital preservation: the SPAR project.

Emmanuelle Bermes (2008). Digital preservation at the National Library of France.

IFLA(2017). IFLA Library Reference Model: a conceptual model for bibliographic information.

Jonas Arnold (2024). Machine learning in the archive: depth indexing in the service of access to archives

Library and Archives Canada(2017). Strategy for a Digital Preservation Program Library and Archives Canada.

웹사이트

국립중앙도서관. https://www.nl.go.kr

국가기록원. http://www.archives.go.kr

Dublin Core. http://dublincore.org

Data Catalog Vocabulary. http://www.w3.org

Trove. https://trove.nla.gov.au

National Archives . https://www.archives.gov

Government CANADA. https://www.canada.ca

Digital Preservation Coalition. https://www.dpconline.org

The Library of Congress. https://www.LOC.gov

The National Library of Australia. https://www.library.gov.au

The National Archives. https://www.nationalarchives.gov.uk

[부록] 용어정리

용어	설명
AIP (Archival Information Package)	장기보존을 위해 보존 시스템 안에 저장·관리되는 공식 보존 단위이다.
API (Application Programming Interface)	시스템 간 기능이나 데이터를 연동하기 위한 연결 방식이다.
CRC (Cyclic Redundancy Check)	데이터 전송이나 저장 중 오류를 확인하기 위한 검사 방식이다.
DIP (Dissemination Information Package)	이용자에게 제공하기 위해 가공·변환된 정보 묶음이다.
Dublin Core (더블린 코어)	기본적인 자원 설명을 위해 널리 사용되는 국제 메타데이터 표준이다.
HDD (Hard Disk Drive)	자기 디스크를 이용해 데이터를 저장하는 장치이다.
JSON (JavaScript Object Notation)	가볍고 이해하기 쉬운 데이터 표현 형식이다.
LTO (Linear Tape-Open)	장기보존에 널리 사용되는 테이프 저장 기술 표준이다.
METS (Metadata Encoding and Transmission Standard)	다양한 메타데이터를 하나의 구조로 묶어 관리하기 위한 국제 표준이다.
MODS (Metadata Object Description Schema)	서지 정보를 보다 상세하게 기술하기 위한 메타데이터 표준이다.
OAIS 모델 (Open Archival Information System)	디지털자원을 장기보존하기 위한 국제 표준 개념 모델로, 보존 시스템이 갖추어야 할 기능과 역할을 제시한다.
PREMIS (Preservation Metadata: Implementation Strategies)	디지털 장기보존을 위해 널리 사용되는 국제 표준 보존 메타데이터 규격이다.
SHA-256/512	무결성 검증에 널리 사용되는 해시 알고리즘의 한 종류이다.
SIP (Submission Information Package)	디지털자원이 보존 시스템에 들어올 때 함께 전달되는 파일과 설명 정보의 묶음이다.
SSD (Solid State Drive)	반도체 메모리를 사용하는 고속 저장 장치이다.
WARC (Web ARChive format)	웹 아카이빙 자료를 저장하기 위해 사용되는 국제 표준 파일 형식이다.
XML (Extensible Markup Language)	구조화된 데이터를 표현하기 위한 표준 문서 형식이다.
가상 머신 (Virtual Machine)	실제 컴퓨터처럼 동작하는 가상 환경이다.
가상화 (Virtualization)	하나의 물리적 장비에서 여러 시스템을 논리적으로 운영하는 기술이다.

용어	설명
계층형 스토리지 (Tiered Storage)	데이터의 중요도나 이용 빈도에 따라 여러 단계의 저장 공간을 나누어 사용하는 구조이다.
관리 메타데이터 (Administrative Metadata)	권한, 이용 조건, 관리 책임 등 행정적 정보를 담은 메타데이터이다.
구조 메타데이터 (Structural Metadata)	디지털자원의 내부 구성과 요소 간 관계를 설명하는 정보이다.
국가 지식정보자원	국가적으로 관리 · 보존할 필요가 있는 지식과 정보 자산을 의미한다.
기계학습 (Machine Learning)	데이터로부터 규칙을 학습하여 스스로 성능을 개선하는 AI 기술이다.
기술 노후화 (Technological Obsolescence)	사용 중인 기술이나 파일 형식이 더 이상 지원되지 않아 이용이 어려워지는 현상이다.
기술 메타데이터 (Technical Metadata)	파일 형식, 해상도, 코덱 등 기술적 특성을 설명하는 정보이다.
네임스페이스 (Namespace)	이름 충돌을 방지하기 위한 식별 영역이다.
노멀라이제이션 (Normalization)	여러 가지 파일 형식을 장기보존에 적합한 표준 형식으로 통일하여 관리하는 방법이다.
다계층 분산 스토리지 (Multi-tier Distributed Storage)	데이터를 여러 계층과 위치에 분산 저장하여 안정성과 확장성을 높인 저장 구조이다.
데이터 교환 (Data Exchange)	서로 다른 시스템 간에 데이터를 주고받는 과정이다.
데이터세트 (Dataset)	특정 목적이나 주제에 따라 구조화된 데이터의 집합이다.
디지털 복원 (Digital Restoration)	손상되었거나 품질이 저하된 디지털 자료를 기술적으로 복구하는 작업이다.
디지털 자료 (Digital Materials)	디지털 형식으로 존재하는 자료를 의미하며, 디지털자원 가운데 실제로 수집 · 관리 대상이 되는 개별 자료를 가리킬 때 사용된다.
디지털자원 (Digital Resources)	컴퓨터나 인터넷을 통해 만들어지고 저장 · 이용되는 모든 형태의 정보 자원이다. 전자책, 웹사이트, 사진 · 영상 파일, 데이터 파일 등이 여기에 포함된다.
디지털자원 생애주기 (Digital Resource Lifecycle)	디지털자원이 생성되어 수집 · 관리 · 보존 · 이용 · 폐기되기까지의 전체 과정을 의미한다.
디지털 장기보존 (Digital Long-term Preservation)	기술환경이 바뀌어도 디지털자원을 오랫동안 안전하게 보존하고 미래에도 계속 이용할 수 있도록 관리하는 활동이다.
디지털 콘텐츠 (Digital Contents)	디지털 환경에서 이용을 목적으로 제작된 정보나 창작물로, 영상 · 음원 · 전자출판물 등이 이에 해당한다.
디지털화 자료 (Digitized Materials)	종이책, 사진, 필름 등 아날로그 자료를 디지털 형식으로 변환한 자료이다.
딥러닝 (Deep Learning)	다층 신경망을 이용해 복잡한 패턴을 분석하는 기계학습 기법이다.
로그 파일 (Log File)	시스템 동작과 처리 이력을 기록한 파일이다.
매체전환 (Refreshment)	데이터는 그대로 두고, 노후된 저장 매체만 새 매체로 교체하는 보존 방법이다.
마이그레이션 (Migration)	오래된 파일 형식이나 시스템의 데이터를 새로운 형식이나 환경으로 변환하는 보존 방법이다.
메타데이터 (Metadata)	디지털자원을 설명하고 관리하기 위한 정보로, ‘데이터에 대한 데이터’ 라고 할 수 있다.

용어	설명
모니터링 (Monitoring)	시스템 상태를 지속적으로 관찰하고 이상을 감지하는 활동이다.
무결성 (Integrity)	디지털자원의 내용이 손상되거나 변경되지 않고 원래 상태를 유지하고 있는 성질이다.
무결성 검증 (Fixity Check)	디지털 파일이 보존 과정에서 손상되거나 변조되지 않았는지를 주기적으로 확인하는 검사이다.
무결성 이력 (Fixity History)	무결성 검증 결과를 시간 순서대로 기록한 정보이다.
무손실 압축 (Lossless Compression)	원본 데이터의 손실 없이 압축하는 방식이다.
문자 인코딩 (Character Encoding)	문자를 디지털 데이터로 표현하는 방식(예: UTF-8)이다.
배치 처리 (Batch Processing)	여러 작업을 한 번에 묶어 처리하는 방식이다.
백업 (Backup)	데이터 손실에 대비해 동일한 데이터를 별도로 복사해 두는 작업이다.
법적 납본 (Legal Deposit)	법률에 따라 출판물이나 디지털 자료를 국가기관에 제출하도록 한 제도이다.
보존 (Preservation)	디지털자원이 손상되거나 사라지지 않도록 관리하고 보호하는 모든 활동을 의미한다.
보존 개입 (Preservation Action)	장기보존을 위해 수행되는 마이그레이션, 점검, 복구 등의 개별 작업을 의미한다.
보존 거버넌스 (Preservation Governance)	디지털자원 보존을 위해 기관 간 역할과 책임을 정한 운영 체계이다.
보존 단계 (Preservation Phase)	디지털자원이 수집된 이후 관리 · 보존 · 제공되는 과정에서의 단계 구분이다.
보존 대상 (Preservation Objects)	장기보존을 위해 관리해야 하는 디지털자원이나 자료를 의미한다.
보존 메타데이터 (Preservation Metadata)	디지털자원의 보존 이력과 처리 과정을 기록하는 메타데이터이다.
보존 범위 (Preservation Scope)	보존 정책에 따라 실제로 보존 대상으로 포함되는 자원의 종류와 한계를 뜻한다.
보존 시스템 (Preservation System)	디지털자원을 장기적으로 안전하게 보존하기 위해 구축된 전용 정보시스템이다.
보존 위험 (Preservation Risk)	기술 변화, 저장 매체 노후화, 관리 미흡 등으로 인해 디지털자원이 손실될 가능성을 의미한다.
보존 이벤트 (Preservation Event)	보존 과정에서 발생한 주요 작업과 변경 사항을 기록한 이력 정보이다.
보존전략 (Preservation Strategy)	디지털자원을 어떤 방식으로 보존할지 선택하는 방법과 방향을 의미한다.
보존 정책 (Preservation Policy)	어떤 디지털자원을, 어느 수준까지, 어떻게 보존할 것인지에 대한 기관의 공식적인 기준이다.
보존 책임 기관	특정 디지털자원의 장기보존을 담당하는 공식 기관이다.
보존체계 (Preservation Framework)	디지털자원을 보존하기 위해 정책, 조직, 기술, 절차를 체계적으로 구성한 전체 구조이다.
보존 패키지 (Preservation Package)	디지털자원과 메타데이터를 하나로 묶은 장기보존 단위이다.
보존 포맷 (Preservation Format)	장기간 안정적으로 보존하기에 적합하도록 선택된 파일 형식이다.

용어	설명
보존 품질 관리 (Preservation Quality Management)	디지털자원의 정확성과 신뢰성을 유지하기 위해 지속적으로 점검하고 관리하는 활동이다.
블록 스토리지 (Block Storage)	데이터를 일정한 크기의 블록으로 나누어 저장하는 방식으로 성능이 중요한 환경에 사용된다.
비공개 포맷 (Proprietary Format)	특정 회사나 소프트웨어에 종속된 파일 형식이다.
사용과 보존의 공존 단계	디지털자원이 이용 서비스에 제공되면서 동시에 장기보존 관리도 함께 이루어지는 상태를 말한다.
상호운용성 (Interoperability)	서로 다른 시스템이나 기관 간에 데이터가 문제없이 연계 · 교환되는 능력이다.
서지 메타데이터 (Bibliographic Metadata)	제목, 저자, 발행 정보 등 자료의 기본적인 정보를 담은 메타데이터이다.
손실 압축 (Lossy Compression)	용량을 줄이기 위해 일부 정보를 제거하는 압축 방식이다.
수명 주기 (Media Lifecycle)	저장 매체가 사용되고 교체되기까지의 기간이다.
스케줄링 (Scheduling)	정해진 시간에 점검이나 작업이 실행되도록 설정하는 기능이다.
스키마 (Schema)	데이터 구조와 규칙을 정의한 설계도이다.
스토리지 (Storage)	디지털 데이터를 저장하고 관리하는 시스템 또는 공간을 의미한다.
스토리지 리프레시	노후화된 저장 장비를 새 장비로 교체하는 작업이다.
스토리지 인프라 (Storage Infrastructure)	디지털자원을 저장 · 보존하기 위해 구축된 하드웨어와 소프트웨어 환경 전체를 말한다.
시청각자료 (Audiovisual Materials)	영상과 음성으로 구성된 자료로, 영화 · 방송 · 음원 등이 포함된다.
실행환경 (Execution Environment)	디지털자원이 작동하는 운영체제와 소프트웨어 환경을 의미한다.
에뮬레이션 (Emulation)	과거의 소프트웨어나 하드웨어 환경을 가상으로 재현하여 디지털자원을 원래 모습대로 이용하는 방법이다.
연구데이터 (Research Data)	연구 활동 과정에서 생성된 데이터로, 재사용과 검증을 위해 보존 가치가 있다.
영구보존 (Permanent Preservation)	국가나 기관이 영구히 보존할 가치가 있다고 판단한 디지털자원을 변경 없이 장기간 보관하는 보존 단계이다.
예측 분석 (Predictive Analysis)	과거 데이터를 바탕으로 미래의 상태나 위험을 예측하는 분석 방법이다.
오류 검출 (Error Detection)	데이터 손상 여부를 찾아내는 기술적 과정이다.
오류 복구 (Error Recovery)	손상된 데이터를 복구하거나 대체하는 처리 과정이다.
오브젝트 스토리지 (Object Storage)	데이터를 객체 단위로 관리하는 방식으로 대용량 데이터를 장기보존하는 데 적합하다.
오프사이트 백업 (Off-site Backup)	재난 상황에 대비해 원본과 다른 장소에 데이터를 저장하는 방식이다.
온라인 자료 (Online Resources)	인터넷을 통해 생산 · 유통되는 디지털 자료를 의미한다.

용어	설명
온라인자료 수집 제도	온라인에서 생산되는 자료를 국가 차원에서 수집 · 보존하기 위한 제도적 장치이다.
원본성 (Originality)	디지털자원이 최초 생성된 원본임을 의미하는 개념으로 진본성과 함께 사용된다.
웜 티어 (Warm Tier)	중간 정도의 접근 빈도를 가진 데이터를 저장하는 영역이다.
웹 아카이빙 (Web Archiving)	웹사이트와 온라인 콘텐츠를 수집 · 저장하여 장기적으로 보존하는 활동이다.
웹 자원 (Web Resources)	웹사이트, 웹페이지, 온라인 서비스 등 웹 환경에서 제공되는 정보 자원이다.
웹사이트 수집 (Website Harvesting)	특정 기준에 따라 웹사이트의 내용을 자동으로 수집하는 과정이다.
이관 (Ingest)	디지털자원이 보존 시스템으로 공식적으로 전달되는 과정이다.
이벤트 기반 처리 (Event-driven Processing)	특정 사건이 발생했을 때 자동으로 작업이 실행되는 방식이다.
이상 탐지 (Anomaly Detection)	정상 상태와 다른 이상 징후를 자동으로 찾아내는 기술이다.
인공지능 (Artificial Intelligence, AI)	사람의 사고나 판단을 모방하여 데이터를 분석 · 처리하는 컴퓨터 기술이다.
인공지능 기반 보존	AI를 활용하여 디지털자원의 분류, 점검, 관리 업무를 자동화하는 방식이다.
인코딩 (Encoding)	데이터를 컴퓨터가 이해할 수 있는 방식으로 표현하는 규칙이다.
읽기 전용 매체 (Write Once Media)	한 번만 기록하고 이후 수정할 수 없는 저장 매체이다.
자동 메타데이터 생성 (Automatic Metadata Generation)	파일 내용을 분석해 메타데이터를 자동으로 만들어주는 기술이다.
자동 무결성 점검	시스템이 주기적으로 파일 상태를 자동으로 확인하는 기능이다.
자동 분류 (Automatic Classification)	디지털자원의 유형이나 주제를 자동으로 구분하는 기능이다.
자동 색인 (Automatic Indexing)	검색을 위해 필요한 색인 정보를 자동으로 생성하는 기술이다.
자동화 워크플로 (Automated Workflow)	반복적인 보존 작업을 시스템이 자동으로 처리하는 흐름이다.
장기보존 (Long-term Preservation)	수십 년 이상의 장기간을 전제로 디지털자원을 안정적으로 관리하는 보존 개념이다.
재해 복구 (Disaster Recovery)	시스템 장애나 재난 발생 시 디지털자원을 복구하기 위한 절차와 체계를 말한다.
재현 (Reproduction / Replay)	과거의 디지털 환경이나 실행 결과를 다시 재현하여 보여주는 보존 · 이용 방식이다.
저작권 관리 (Copyright Management)	디지털자원의 보존 · 이용 과정에서 저작권을 보호하고 조정하는 활동이다.
저장 매체 (Storage Media)	디지털 데이터를 실제로 저장하는 물리적 장치로, 하드디스크, SSD, 테이프 등이 있다.
저장 매체 상태 점검 (Storage Media Health Check)	하드디스크나 테이프 등 저장 장치의 상태를 점검하여 장애를 예방하는 관리 활동이다.
전자저널 (E-journal)	온라인으로 제공되는 학술지나 정기 간행물이다.

용어	설명
전자책 (E-book)	디지털 형식으로 제작되어 전자기기를 통해 읽을 수 있는 책이다.
접근 제한 (Access Restriction)	저작권이나 개인정보 보호를 위해 이용을 제한하는 조치이다.
접근성 (Accessibility)	이용자가 디지털자원에 접근하여 사용할 수 있는 정도를 의미한다.
진본성 (Authenticity)	디지털자원이 위·변조되지 않았고, 원래 생성된 그대로임을 신뢰할 수 있는 속성이다.
체크섬 (Checksum)	파일의 무결성을 확인하기 위해 계산되는 값으로 해시값과 유사한 개념이다.
캡슐화 (Encapsulation)	디지털자원과 그 자원을 설명하는 정보를 함께 묶어 보존하는 방식이다.
코덱 (Codec)	영상이나 음원을 압축하거나 해제하는 기술이다.
콜드 티어 (Cold Tier)	장기보존 목적의 접근 빈도가 낮은 데이터를 저장하는 영역이다.
테이프 스토리지 (Tape Storage)	자기 테이프를 이용한 대용량 장기보존용 저장 방식이다.
파일 스토리지 (File Storage)	파일과 폴더 형태로 데이터를 저장하는 가장 일반적인 스토리지 방식이다.
파일 포맷 (File Format)	디지털 파일이 어떤 구조와 규칙으로 저장되었는지를 나타내는 형식이다.
패키징 (Packaging)	디지털자원과 관련 정보를 체계적으로 묶어 보존 단위를 만드는 과정이다.
포맷 노후화 (Format Obsolescence)	특정 파일 형식이 더 이상 사용되지 않거나 지원이 중단되는 상태를 말한다.
포맷 레지스트리 (Format Registry)	파일 형식의 특성과 보존 적합성 정보를 체계적으로 관리하는 데이터베이스이다.
포맷 유효성 검사 (Format Validation)	파일이 표준 규격에 맞게 생성되었는지를 확인하는 점검 과정이다.
표준 포맷 (Standard Format)	국제적으로 널리 사용되고 공개된 규격을 가진 파일 형식이다.
하이퍼바이저 (Hypervisor)	가상 머신을 생성·관리하는 소프트웨어이다.
핫 티어 (Hot Tier)	자주 사용되는 데이터를 저장하는 고성능 저장 영역이다.
해시 알고리즘 (Hash Algorithm)	파일의 내용을 바탕으로 고유한 값을 계산하는 수학적 방법이다.
해시값 (Hash Value)	파일 내용을 수학적으로 계산해 만든 고유한 값으로 파일이 변경되었는지 확인하는 기준으로 사용된다.
환경 의존성 (Environment Dependency)	특정 환경에서만 파일이나 프로그램이 작동하는 특성이다.
환경 재현	과거의 실행환경을 다시 구성하여 디지털자원을 이용하는 방식이다.

디지털자원의 장기보존전략 및 인프라 구축 방안 연구

인 쇄 : 2025년 12월 15일
발 행 : 2025년 12월 15일
발행처 : 문화체육관광부 국립중앙도서관
서울특별시 서초구 반포대로 201(반포동)
http://www.nl.go.kr

연구기관 : ㈜디아이엔티
ISBN 979-11-6513-553-9

인쇄 : 더프린트콜(Tel : 02-343-6868)

<비매품>



비매품/무료
03020
9 791165 135539
ISBN 979-11-6513-553-9