2018년 제6차 원내세미나

파일손상 전자기록물 복구 연구
(R&D 자체연구과제) 추진현황

국가기록원
전자기록관리과

2018.07.18.

행정안전부 국가기록원

CONTENT
진행 순서

CHAPTER 1. 개요

01. 손상파일 복구의 개념

02. 연구과제 소개

03. 주요내용 및 추진방향

01 개요 손상파일 복구의 개념

논리적인 데이터 복구

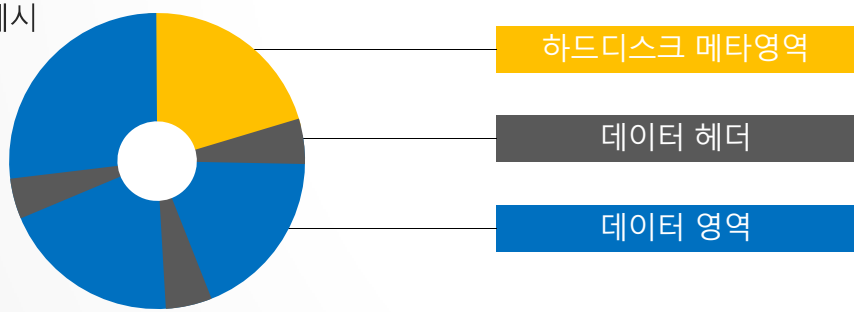
저장매체의 물리적 구성요소가 아닌 논리적인 파일시스템과 파일에 기반하여 복구하는 기법

일반적인 복구의 개념

- ✓ 손상파일이 저장된 매체를 이용하여 복구 수행
- ✓ 메타 영역이나 데이터 헤더에 손상이 있는 경우, 존재하는 데이터 영역을 확인하여 복구 실시
- ✓ 데이터에 손상이 발생한 경우 복구 가능성 낮음

일반적인 파일시스템의 저장구조

※ 예시

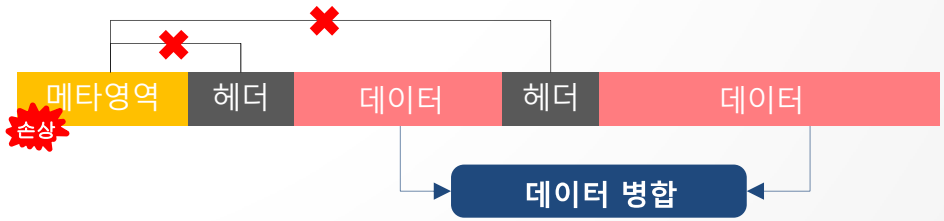


- ✓ 메타 영역 – 헤더, 데이터 영역의 위치 등, 전체를 관리하는 영역
- ✓ 헤더 영역 – 데이터에 대한 정보를 보유한 영역
- ✓ 데이터 영역 – 실제 데이터를 보유하고 있는 영역

대표적인 파일복구 방법

- **메타 정보를 이용한 복구**
  - 메타 영역에 존재하는 파일의 위치와 크기 정보를 이용한 복구

- **데이터 카빙을 이용한 복구**
  - 메타영역의 손상으로 헤더 및 데이터에 대한 연결값이 사라진 경우
  - 하드디스크에 저장된 데이터를 직접 확인하여 병합



02 개요 손상파일 복구의 개념

손상파일 복구의 개념

복구 방법과 무관하게, 데이터가 온전해야만 복구가 가능

이관된 전자기록물의 복구

✓ 단일파일의 이관으로, 이용할 수 있는 정보가 비교적 적음

✓ 파일의 고유정보만을 이용해 복구 가능
(단일파일의 고유정보는 "포맷정보", "서식정보", "데이터"로 구성)

✓ **주요 복구대상은 포맷정보, 서식정보에 손상이 있는 파일**
(e.g. 포맷정보 깨짐, 서식정보 깨짐, 확장자 오기재, 기타)

✓ 저장매체가 아닌, 파일단위로 이관
(매체에 잔존하는 정보 부재, 이를 이용한 복구 불가능)

손상 전자기록물의 복구 방안

단일파일의 고유정보

포맷정보
- 파일포맷, 버전 정보

서식정보
- 링크정보, 인코딩, 암호화

데이터
- 본문, 이미지, 사진, 표

● **포맷정보 부분이 손상된 경우**

- (가능할 경우) 포맷정보의 HEX값을 정상으로 수정
- 서식정보 및 데이터만 추출하여 정상포맷정보에 이식

● **서식정보 부분이 손상된 경우**

- (가능할 경우) 서식정보의 HEX값을 정상으로 수정
- 데이터 부분만을 추출(내용만 복구)

● **데이터 영역이 손상(손실)되었을 경우, 복구 불가능**

03 개요 연구과제 소개

과제명 : 파일 손상 전자기록물 복구 연구(자체+위탁)

연구목표 1

- **파일손상 전자기록물의 유형 분석 및 복구방안 연구**

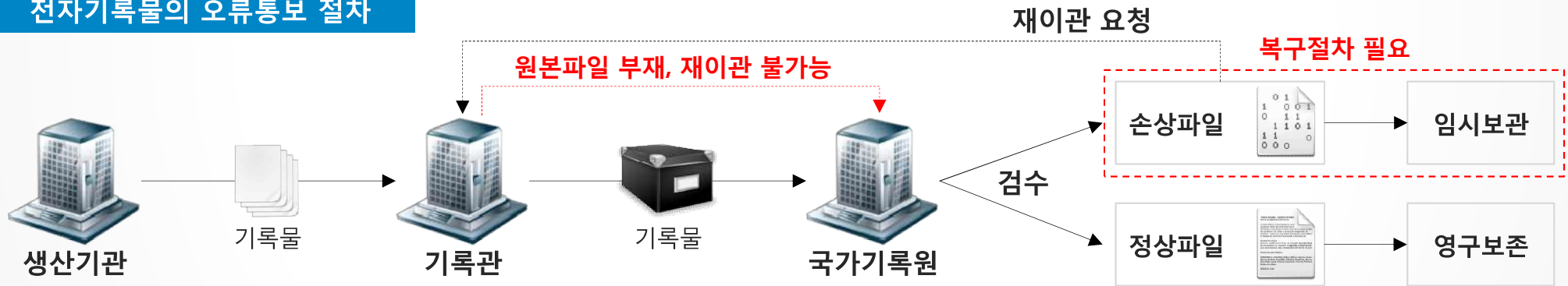
✓ 손상유형 분석 및 복구를 위한 데이터 기반의 분석
✓ 파일구조에 대한 자세한 분석을 통해 복구 가능성 및 방법 도출

연구목표 2

- **전자기록물 복구를 위한 제도적 개선방안 연구**

✓ 신뢰할 수 있는 손상 파일의 복구 및 관리 프로세스 정립
✓ 체계적인 손상파일의 검사, 복구를 위한 제도적 기반 마련

전자기록물의 오류통보 절차



이관되는 전자기록물의 수량 증가

- 이관되는 전자기록물의 수량이 지속적으로 증가하고 있으며, 그에 따른 손상 발생 위험 증가

이관시기에 따른 재이관의 어려움

- 이관 시기는 생산 후 10년 이상의 시간이 경과한 시점
- 생산시스템 에서 해당 전자기록물을 확인할 수 없는 경우 다수 발생 (시스템 노후화, 마이그레이션, 기타)

손상유형 분석의 어려움

- 국가기록원 이관 이후에 상세 검수 수행
- 생산 및 기록관 단계에서 손상 파악 곤란

04 개요 주요내용 및 추진방향

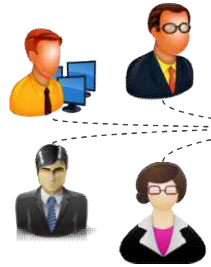
손상파일 자체분석

자체분석	방법론 도출		시스템 설계
손상유형 분석	자동분류 방안	자동복구 가능성	프로토타입 프로그램 설계

▶ 손상파일 복구 전 과정의 프로세스 설계 ▶

- 자체분석을 통해 **손상파일 자동분류 및 자동복구 프로세스 도출**
- 프로세스를 기반으로 **프로토타입 설계**

원내·외 전문가 및 관련 기관 자문



연구방향

- 복구방안
- 복구절차
- 타기관 사례
- 적용 가능성

- 원내 회의 및 전문가 자문을 통한 **의견수렴 및 검증**

국내·외 기술동향 조사

논문, 연구보고서

포맷검증 기술 | 손상발생 유형 | 자동복구 기술 | 진본성 확보 기술

진본성, 무결성 확보 절차 및 사례

디지털포렌식 증거처리 절차

- 논문, 연구보고서 등을 통해 **연구과제 관련 기술 조사**
- **신뢰성을 보장할 수 있는 프로세스 사례 및 절차 조사**

위탁사업을 통한 프로토타입 개발

프로토타입 개발 → 테스트 수행 → 시범적용

- 프로토타입 프로그램 설계를 기반으로 **위탁업체에서 개발 수행**
- 프로토타입 프로그램을 통해 **테스트 수행 및 검증**

CHAPTER 2. 추진현황

01. 손상 파일 분석결과

02. 손상파일 검사 및 복구 프로세스

05 추진현황 손상파일 분석 방법

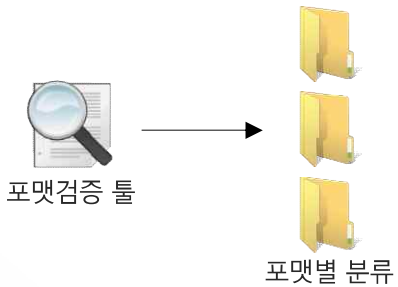
분석준비 ▶ **포맷 분류** ▶ **1차 검사(파일실행)** ▶ **2차 검사(코드확인)**

- 59,398개의 파일 중 분석대상 파일 분류
- 분석 대상 파일을 CAMS 에서 확보

실제 분석대상인 4,749개의 손상파일 확보

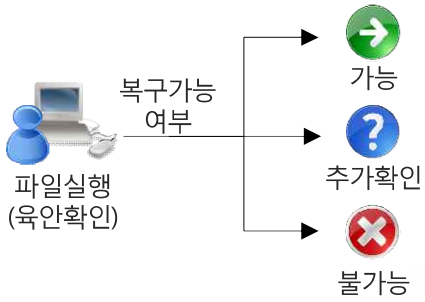
- 포맷검증 툴을 활용한 포맷별 분류

포맷검증 툴(DFR, LUPE)을 이용해 포맷별로 분류



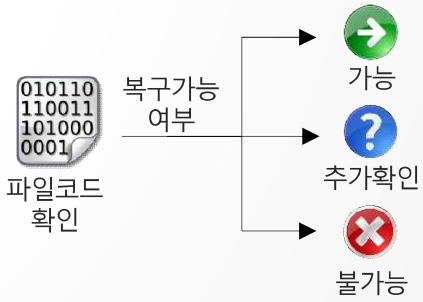
- 손상파일의 실행을 통한 오류 확인
- 자동분류 가능성 판단
- 복구 가능성 분석

손상유형 확인과 더불어 자동분류 가능성을 분석



- 파일실행에서 추가확인이 필요한 대상에 대해 실제 파일의 값(HEX값)을 확인

추가확인 대상 파일의 코드를 확인하고, 복구가능성 확인



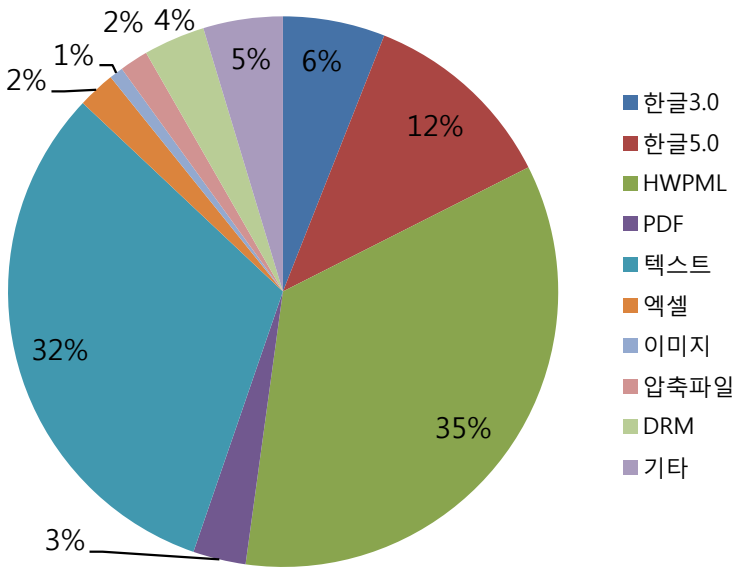
06 추진현황 손상유형 분석결과

전자기록물 손상현황

59,398개의 손상파일 중, 복구 연구를 위해 분석이 필요한 파일은 4,749개

포맷별 현황

한글3.0	285개
한글5.0	549개
HWPML	1,646개
PDF	147개
텍스트	1,508개
엑셀	104개
이미지	39개
압축파일	80개
DRM	172개
기타	219개



1차 분석 결과

단순 복구 - 정상파일

확장자 오기재로 인한 손상, 확장자 변환을 통해 102개 파일 정상 복구

복구 가능

약 1,600개의 HWPML 파일은 복구 가능

복구 불가능

대체파일, 빈파일, DRM, 링크파일 등
약 1,400개의 손상파일은 복구 불가능

추가분석 필요

PDF, 엑셀, 한글, 기타 등
약 1,700개의 손상파일은 추가분석 필요

손상으로 분류된 파일의 손상현황

- 59,398개의 손상으로 분류된 파일 중, **HTML(링크파일) 2만 9천여개**, 원본 손실 파일 **2만 5천여개 / 분석대상 파일은 4,749개**
- 4,749개의 파일에 대한 1차 분석 완료 / 현재 2차 분석 수행 중
- **약 1,600개(34%)의 파일은 복구 가능 예상 / 약 1,400개(29%)의 파일은 복구 불가능**

복구가 가능한 손상유형

07 손상유형 분석
복구가 가능한 손상유형(HWPML)

HWPML

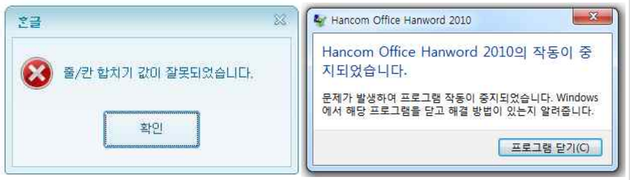
문서의 서식관련 정보만 깨지고 실제 내용은 존재하는 경우로, 복구 가능할 것으로 판단

HWPML 분류방법

- 파일의 확장자가HML, HWN 등으로 구성, HEX코드 내부에 HTML, XML과 같은 태그값이 존재
- 태그구조의 특성을 갖고 있는 HTML, XML은 태그 불일치로 에러가 발생, 서식이 깨지는 경우가 다수 존재
- 4,749개의 손상파일 중, 대상 파일의 개수는 1,646개 (HWP, HML, HWN 등의 포맷이 존재)

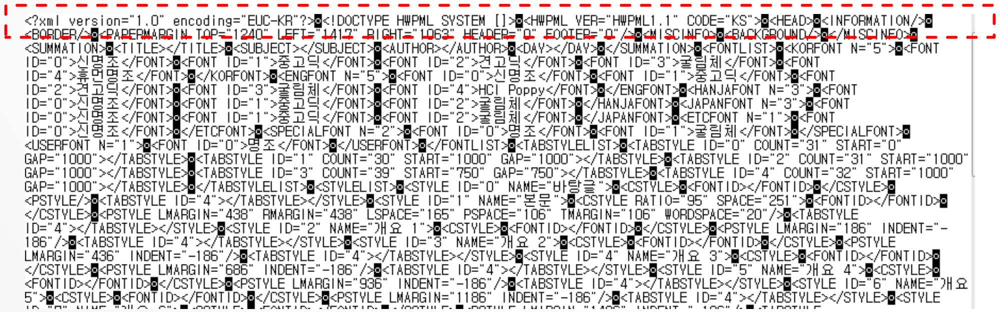
복구전

- 파일 열기에 따른 에러 발생
- ‘줄/칸 합치기 값’ 에러는 문서에 존재하는 서식정보 손상을 의미



※ 파일열기에 따른 에러 메시지

- HEX코드 내부를 확인한 결과, XML 태그 구조와 HWPML포맷임을 확인

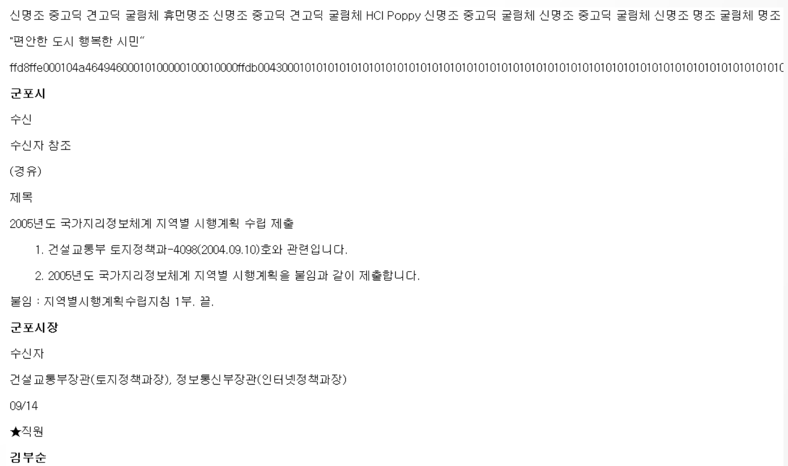


```
<?xml version="1.0" encoding="EUC-KR"?>■<!DOCTYPE HWPML SYSTEM []>■<HWPML VER="HWPML1.1" CODE="KS">■<HEAD>■<INFORMATION/>■
<BORDER/>■<PAPERMARGIN TOP="1240" LEFT="1417" RIGHT="1063" HEADER="0" FOOTER="0"/>■<MISCINFO>■<BACKGROUND/>■</MISCINFO>■
<SUMMATION>■<TITLE></TITLE>■<SUBJECT></SUBJECT>■<AUTHOR></AUTHOR>■<DAY></DAY>■</SUMMATION>■<FONTLIST>■<KORFONT N="5">■<FONT
ID="0">신명조</FONT>■<FONT ID="1">중고딕</FONT>■<FONT ID="2">견고딕</FONT>■<FONT ID="3">굴림체</FONT>■<FONT
ID="4">휴먼명조</FONT>■</KORFONT>■<ENGFONT N="5">■<FONT ID="0">신명조</FONT>■<FONT ID="1">중고딕</FONT>■<FONT
ID="2">견고딕</FONT>■<FONT ID="3">굴림체</FONT>■<FONT ID="4">HCI Poppy</FONT>■</ENGFONT>■<HANJAFONT N="3">■<FONT
ID="0">신명조</FONT>■<FONT ID="1">중고딕</FONT>■<FONT ID="2">굴림체</FONT>■</HANJAFONT>■<JAPANFONT N="3">■<FONT
ID="0">신명조</FONT>■<FONT ID="1">중고딕</FONT>■<FONT ID="2">굴림체</FONT>■</JAPANFONT>■<ETCFONT N="1">■<FONT
ID="0">신명조</FONT>■</ETCFONT>■<SPECIALFONT N="2">■<FONT ID="0">명조</FONT>■<FONT ID="1">굴림체</FONT>■</SPECIALFONT>■
<USERFONT N="1">■<FONT ID="0">명조</FONT>■</USERFONT>■</FONTLIST>■<TABSTYLELIST>■<TABSTYLE ID="0" COUNT="31" START="0"
GAP="1000"></TABSTYLE>■<TABSTYLE ID="1" COUNT="30" START="1000" GAP="1000"></TABSTYLE>■<TABSTYLE ID="2" COUNT="31" START="1000"
GAP="1000"></TABSTYLE>■<TABSTYLE ID="3" COUNT="39" START="750" GAP="750"></TABSTYLE>■<TABSTYLE ID="4" COUNT="32" START="1000"
GAP="1000"></TABSTYLE>■</TABSTYLELIST>■<STYLELIST>■<STYLE ID="0" NAME="바탕글">■<CSTYLE>■<FONTID></FONTID>■</CSTYLE>■
<PSTYLE/>■<TABSTYLE ID="4"></TABSTYLE></STYLE>■<STYLE ID="1" NAME="본문">■<CSTYLE RATIO="95" SPACE="251">■<FONTID></FONTID>■
</CSTYLE>■<PSTYLE LMARGIN="438" RMARGIN="438" LSPACE="165" PSPACE="106" TMARGIN="106" WORDSPACE="20"/>■<TABSTYLE
ID="4"></TABSTYLE></STYLE>■<STYLE ID="2" NAME="개요 1">■<CSTYLE>■<FONTID></FONTID>■</CSTYLE>■<PSTYLE LMARGIN="186" INDENT="-
186"/>■<TABSTYLE ID="4"></TABSTYLE></STYLE>■<STYLE ID="3" NAME="개요 2">■<CSTYLE>■<FONTID></FONTID>■</CSTYLE>■<PSTYLE
LMARGIN="436" INDENT="-186"/>■<TABSTYLE ID="4"></TABSTYLE></STYLE>■<STYLE ID="4" NAME="개요 3">■<CSTYLE>■<FONTID></FONTID>■
</CSTYLE>■<PSTYLE LMARGIN="686" INDENT="-186"/>■<TABSTYLE ID="4"></TABSTYLE></STYLE>■<STYLE ID="5" NAME="개요 4">■<CSTYLE>■
<FONTID></FONTID>■</CSTYLE>■<PSTYLE LMARGIN="936" INDENT="-186"/>■<TABSTYLE ID="4"></TABSTYLE></STYLE>■<STYLE ID="6" NAME="개요
5">■<CSTYLE>■<FONTID></FONTID>■</CSTYLE>■<PSTYLE LMARGIN="1186" INDENT="-186"/>■<TABSTYLE ID="4"></TABSTYLE></STYLE>■<STYLE
```


※ 해당 파일의 HEX코드

단순 복구후

- 확장자를 .html로 변경하여 파일을 열어본 결과, 해당 파일의 내부 내용에는 문제가 없음을 확인 (서식정보만 복구할 경우, 완전복구 가능)



신명조 중고딕 견고딕 굴림체 휴먼명조 신명조 중고딕 견고딕 굴림체 HCI Poppy 신명조 중고딕 굴림체 신명조 중고딕 굴림체 신명조 명조 굴림체 명조

"편안한 도시 행복한 시민"

ffd8ffe000104a46494600010100000100010000ffdb004300010101010101010101010101010101010101010101010101010101010101010101010101010101010101010

군포시

수신

수신자 참조

(경유)

제목

2005년도 국가지리정보체계 지역별 시행계획 수립 제출

1. 건설교통부 토지정책과-4098(2004.09.10)호와 관련입니다.
2. 2005년도 국가지리정보체계 지역별 시행계획을 붙임과 같이 제출합니다.

붙임 : 지역별시행계획수립지침 1부. 끝.

군포시장

수신자

건설교통부장관(토지정책과장), 정보통신부장관(인터넷정책과장)

09/14

★직원

김부순

08 손상유형 분석
복구가 가능한 손상유형(BASE64)

BASE64

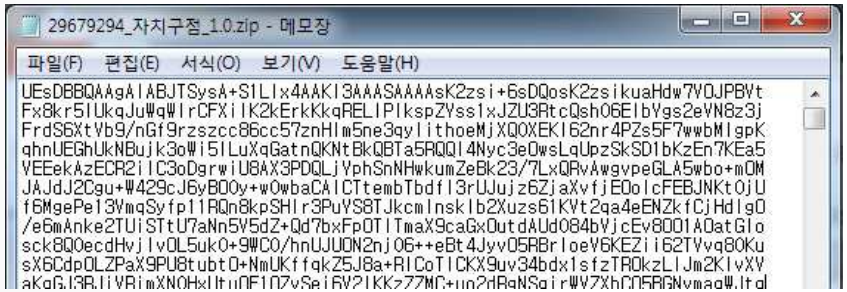
BASE64 정보를 보유한 파일로, 디코딩과 확장자 검사틀 통해 복구 가능

BASE64 분류방법

- BASE64 파일은 별도의 시그니처가 존재하지 않으므로, 값을 구성하는 문자의 패턴을 이용해 분류
- BASE64 파일은 확장자에 오류가 발생한 경우가 대다수, 확장자 변경과 BASE64 디코딩할 경우 복구 가능
- 4,749개의 손상파일 중, 압축파일의 개수는 79개 (텍스트, ZIP 등의 포맷이 존재)

복구 전

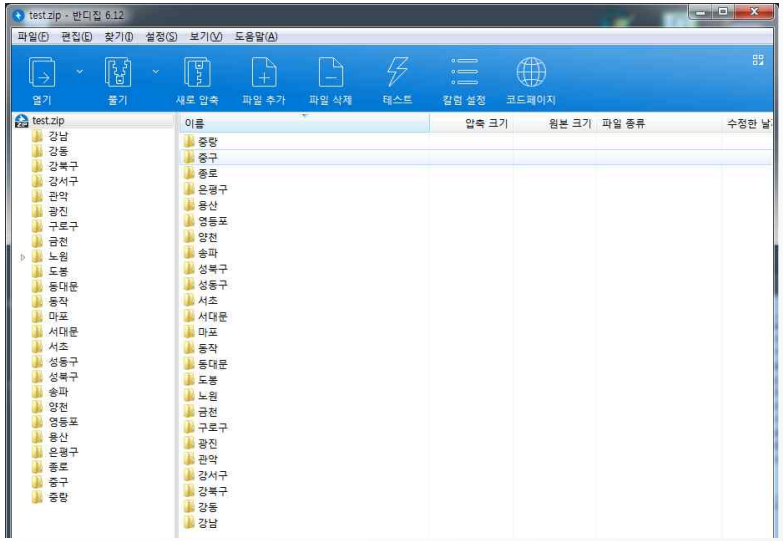
- 파일값을 확인한 결과, BASE64가 의심되는 문자들로 구성



복구 후

```
- <result identificationMessage="확장자 오기재" identificationCode="00" message="실패" code="01"
  originalFileName="test.gif" for="file1">
    - <format>
        <name>ZIP Format</name>
        <version/>
        <mime-type>application/zip</mime-type>
        <puid>x-fmt/263</puid>
        <extension>zip</extension>
        <method>Signature</method>
      </format>
      <error message="붙임손상_확장자이상" code="1104">확장자 오기재</error>
```

- BASE64 디코딩 실행 후, 파일포맷 검사툴로 확장자 확인
- 해당 포맷으로 변경 후, 정상파일임을 확인



복구가 불가능한 손상유형

09 손상유형 분석
복구가 불가능한 손상유형(DRM)

DRM

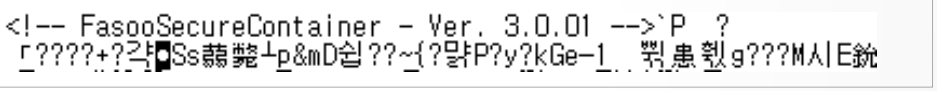
생산시스템에서 DRM이 걸린 파일 자체를 이관한 경우

DRM파일 분류방법

- DRM파일도 일반파일과 동일하게 헤더의 가장 처음에 특정 시그니처 정보를 보유
- DRM파일의 시그니처 정보를 통해 분류 가능, DRM파일은 암호화된 파일이며, 해독이 불가능함
- 4,749개의 손상파일 중, DRM파일의 개수는 172개 (한글, 엑셀 등)

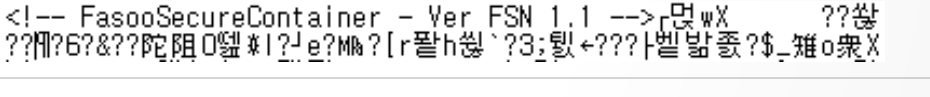
Fasoo FSD

<!-- FasooSecureContainer - Ver. 3.0.01 -->

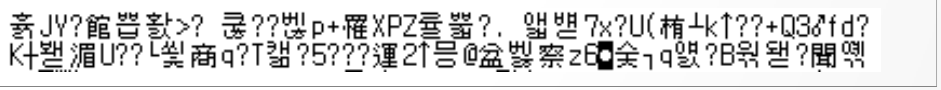


Fasoo FSN

<!-- FasooSecureContainer - Ver FSN 1.1 -->

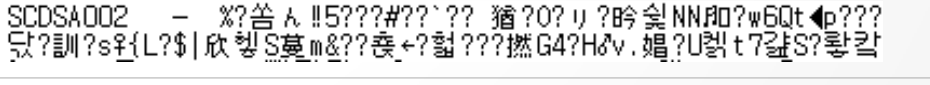


Markany Mark2



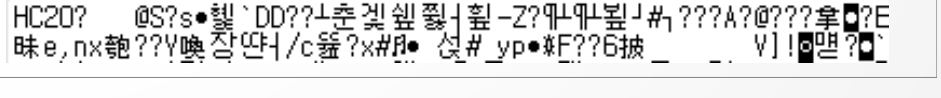
Softcamp

SCDSA002



추가확인 필요

HC20



10 손상유형 분석
복구가 불가능한 손상유형(빈파일)

빈파일

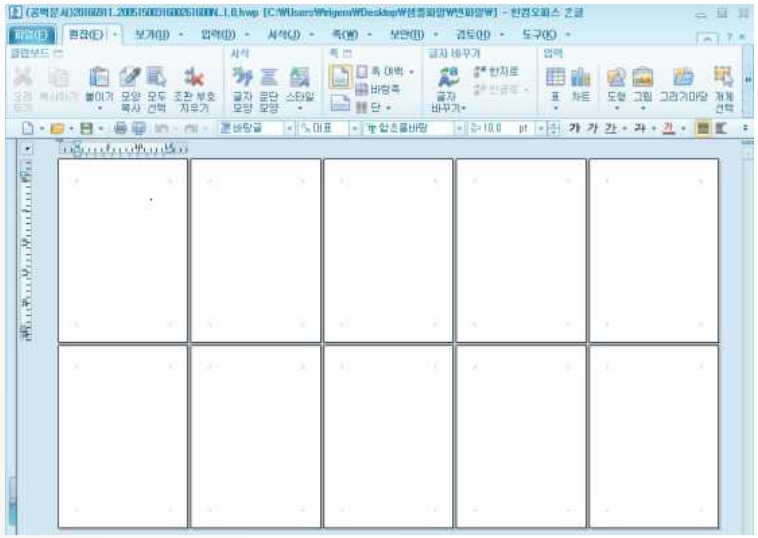
공백 혹은 내용이 없는 파일을 이관한 경우

빈파일 분류방법

- 빈파일에 존재하는 0 값을 이용한 방법(포맷구성 영역외에는 모두 0으로 존재)
- 파일용량 기반 분류 (포맷구성을 위한 기본용량만 존재)
- 4,749개의 손상파일 중, 빈파일의 개수는 95개 (한글, PDF, 압축파일, 이미지 등의 포맷이 존재)

원본형태로 열기

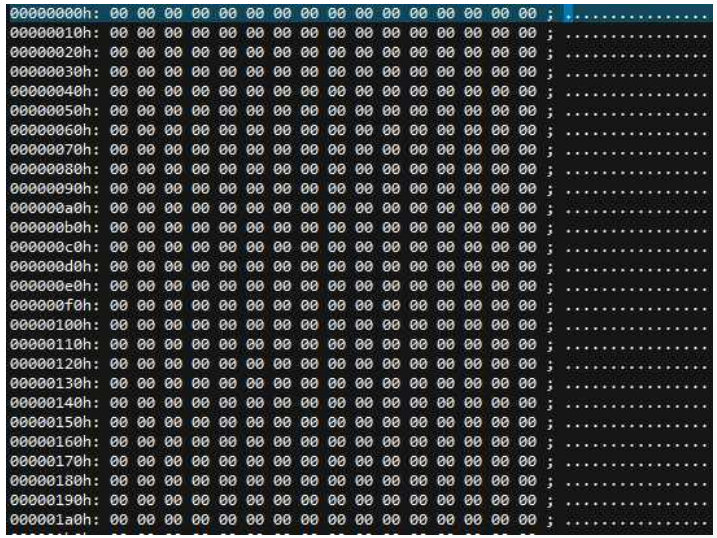
- 파일 내부에 내용이 없음을 확인(공백)



※ 열어본 한글파일 내부내용

HEX코드 분석

- HEX 값도 모두 0으로 구성되어 있는 것을 확인



※ 열어본 한글파일 내부내용

11 손상유형 분석
복구가 불가능한 손상유형(압축파일)

압축파일

압축파일의 손상부분을 제외한 나머지는 정상파일, 압축파일의 손상이기 때문에 복구가 불가능

압축파일 분류방법

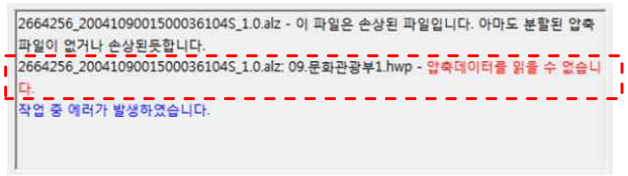
- 압축파일은 고정적으로 헤더에 특정 시그니처가 존재
- 파일별로 구분되는 압축파일 구조상 손상이 발생할 경우, 손상 대상 파일은 복구 불가능
- 4,749개의 손상파일 중, 압축파일의 개수는 80개 (ZIP, ALZ 등의 포맷이 존재)

압축파일 구조

- 압축파일은 특정 알고리즘을 이용해 여러 파일을 하나로 묶으며, 그 과정에서 비트열의 변경이 발생

ZIP파일 분석

① 압축 해제시, 특정파일 손상 메시지 확인



② 압축은 정상적으로 해제됨

③ 압축해제 메시지에 표기된 "09.문화관광부1.hwp" 파일을 실행할 경우 에러메시지 발생

12 손상유형 분석 복구가 불가능한 손상유형(이미지)

이미지

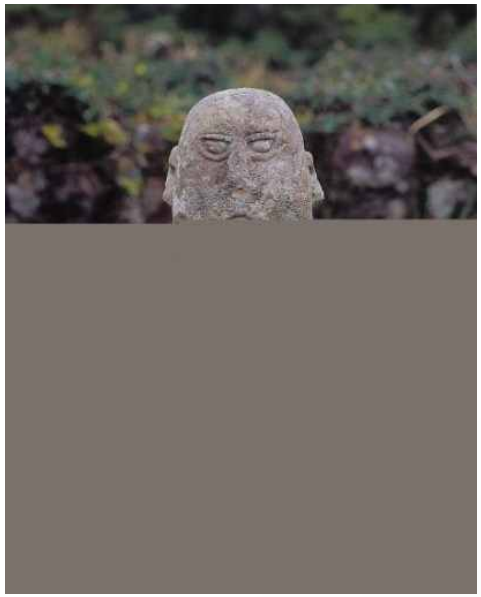
이미지 파일은 별도의 압축 알고리즘을 사용하는 포맷으로 손상된 경우, 복구가 불가능

이미지 파일 분류방법

- 이미지 관련 시그니처 정보를 기준으로 1차 분류, 파일 끝에 존재하는 시그니처 정보 및 파일사이즈 정보를 통해 탐지
- 이미지 파일은 특정 알고리즘을 통해 압축하는 포맷으로 손상시 복구가 불가능
- 4,749개의 손상파일 중, 이미지파일의 개수는 39개 (JPG, JPEG, TIFF, BMP 등)

이미지 파일 손상

- 파일 비트열의 일부분에 손상이 발생한 경우
- 손상지점부터 이미지 소실



손상파일 복구 프로세스

13 손상파일 검사 및 복구 프로세스
손상파일 복구 관련 외국사례 조사

국외사례

NAA(호주), NARA(미국), TNA(영국), LAC(캐나다) 등 4개 기관 조사

NAA (National Archives of Australia)

- 기록원 단계의 복구절차 없음
- 각 기관에서 검사 후 이관 수행

NARA (National Archives and Records Administration)

- 디지털 파일에 대한 복구절차 없음
- 전자매체에 관련된 복구절차만 존재 (직접 복구가 불가능할 경우, 외부에 요청하여 복구수행)

TNA (The National Archives)

- 재이관이 불가능할 경우, 손상파일(the corrupt copy)과 함께 수정파일(a fixed up copy) 생성하여 보존
- 원본 파일(the original file)의 보존과 동시에, 파일 내부의 정보(information within the file)를 접근 가능하게 함

LAC (Library and Archives of Canada)

- 손상파일 및 매체에 대한 복구절차 존재(표준, 절차 등의 문서화 추진 중)
- 가능한 경우 이관 전에 checksum 요청, 원칙적으로 손상파일(damaged files)은 이관 받지 않음

14 손상파일 검사 및 복구 프로세스 설계안

손상파일 검사 및 복구 프로세스

가능한 빠른 시점에 손상 검사를 수행하기 위하여 검사모듈과 복구모듈로 구분하여 개발 필요

처리과 ▶ 각급 기록관 ▶ 영구기록물관리기관

손상파일 검사 및 복구를 위한 신규 이관절차

① 생산기관에서 각급 기록관으로 기록물 이관

② 기록관에서 영구기록물 관리기관으로 이관

③ 장기보존 수행

① 검사모듈을 통해 손상 확인시 자동으로 재이관 요청

② 재이관 불가시, 관련 정보와 함께 RMS에서 보존 후 국가기록원 이관

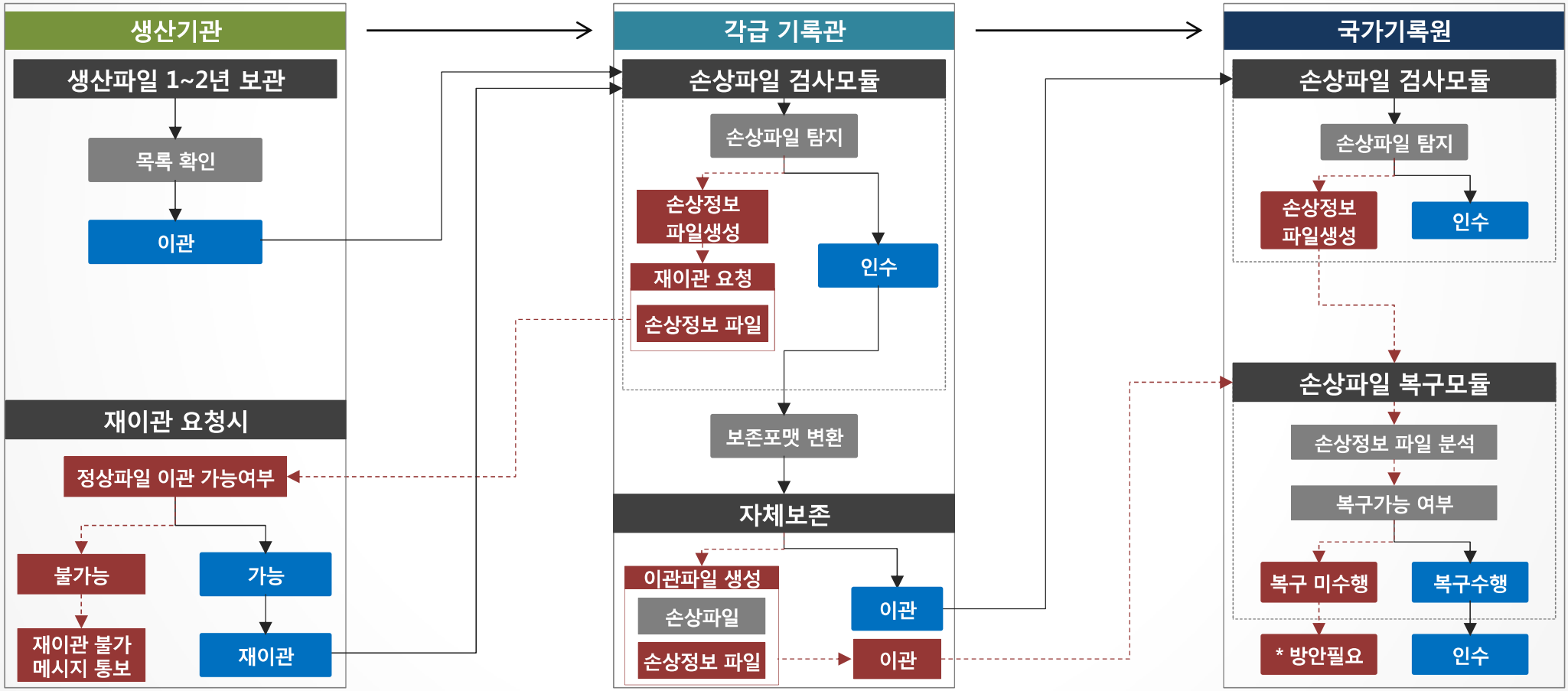
③ 손상복구 모듈을 통해 복구수행

손상파일 검사 및 복구 프로세스

15 손상파일 검사 및 복구 프로세스 설계안(상세)

손상파일 검사 및 복구 프로세스

생산시스템→RMS 이관시 손상파일 검사 및 재이관을 통해 오류 전자기록의 이관 최소화

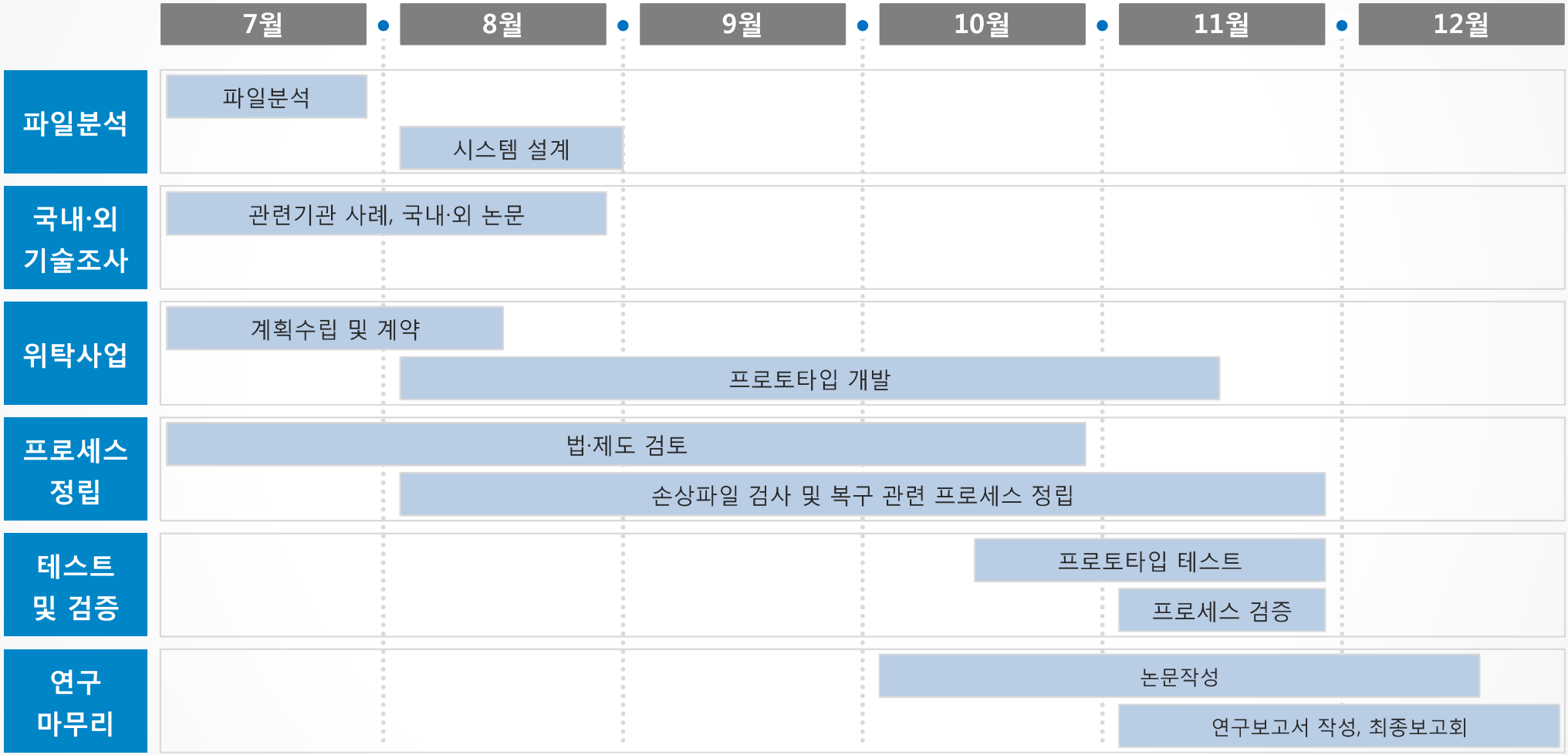


CHAPTER 3. 향후 추진계획

01. 향후 추진일정

02. 기대효과

16 향후 연구계획
향후 추진일정



17 향후 연구계획 기대효과

기대효과

기록관 및 영구기록물관리기관의 이관.검수 프로세스 개선을 통한 전자기록의 품질 확보 및 업무효율 증가

2018년
- 손상유형 분석
- 복구 프로세스 마련
- 프로토타입 개발

→

2019년
- RMS, CAMS 연계방안 연구
- 프로세스 정립
- 지침, 표준 작성

→

2020년
- 시스템 개발
- 법, 제도, 표준 개선

손상파일 감소 및 복구를 통한 중요기록물 멸실 방지

이관 및 검수 프로세스 개선을 통한 기록관리업무 효율 증진

검사 프로세스 개선
사전 검사를 통해 손상파일의 수량을 감소

전자기록물 복구
신뢰성 있는 전자기록물의 장기보존 방안 마련

이관 프로세스 개선
기록물 보존단계부터 영구보존단계까지 관리

시스템 자동화
기록관 및 국가기록원의 업무효율 증진