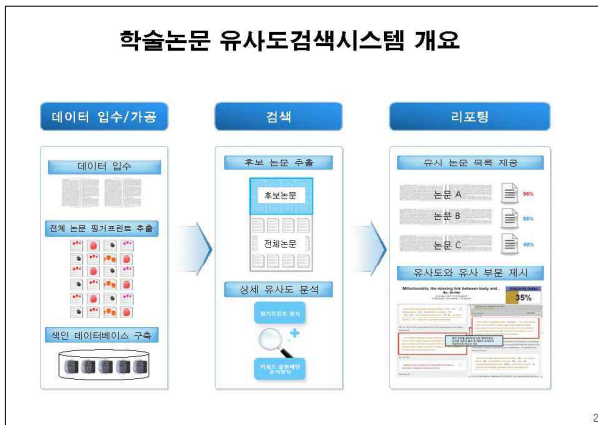


## 학술논문 유사도검색시스템 구축사례

한국과학기술 정보연구원

최 기 석



### CONTENTS

- 1 연구 개요**  
추진배경 및 필요성 | 국내외 현황 | 과제개요 | 추진체계
- 2 연구 목표 및 추진내용**  
연구목표 및 범위 | 주요추진내용
- 3 향후 추진방향**  
활용분야 | 중장기추진계획 | 자기 추진방향(안)
- 4 활용안내**

## 1 연구 개요

- 추진배경 및 필요성
- 국내외 현황
- 과제개요
- 추진체계

### 추진배경 및 필요성

- 성장 위주의 근대화 과정으로 인한 학계의 **표절 관행**이 만연한 상황임
- 현실적으로 **윤리적·제도적 차원**에서 표절을 예방하는 것에는 한계가 있음
- 최근 들어 **표절**과 관련한 **이슈**의 지속적인 발생

<표절 이슈 사례>

- 4대 학술지 표절조사 결과: 문중복하고 표절 확인가 사제발견 (06.7.17, 교수신문)
- 네이버 국내논문 표절 의뢰(2008.10, 동아일보)
- 연 600여건 중 논문 의뢰 (07.30, 동아일보)
- 표절관련 검색엔진 전략한 (08.09.7, 이데일리)
- 학생주제 표절 막겠다 저스팀 개발(08.10, 중앙일보)

※ 국내 표절이슈 증가(연호자조사) : '06년→'11간, '07년→'18, '08년→'28간, '09년→'30간, '10년→'32간

연구현장에서의 **공정사회** 구현과 바람직한 연구문화 정착을 위하여  
**“유사도검색시스템 구축” 필요 !!**

### 국내외 표절예방 활동 현황

국 내	국 외
<p><b>정부주도 연구윤리 확립 활동</b></p> <ul style="list-style-type: none"> <li>- 연구윤리에 대한 인식변화가 높아지고 있음</li> <li>- 주요기관에서 유사도검색시스템 도입 운영 중</li> <li>* 대학 → 서울대/연세대/고려대(KO)/KAIST (07), 하리 → Fjord사 가린가림(11), 홍익(08) → FTR(11)</li> <li>- 한국저학위위원회는 논문을 대상으로 과거 2년간 (08~09) 자체 구축하여 내부에서 이용 → 커티츠확보(저작권보호)의 예사 문제로 주다</li> </ul>	<p><b>영국, 미국은 공공주도로 시작하여 확립단계</b></p> <ul style="list-style-type: none"> <li>- (영국) 영국정보시스템광동위원회(BSC)가 2000년부터 표절예방 노력 시작, 2006년부터 대학에 Turnitin 기반 표절탐색체계 확립 (06년 현재 약97% 적용) ※현재는 고등학교에 적용 중</li> <li>- (미국) CIA에서 기본정책수립 및 구축/연구 지원, 민간협회(COPE, CrossRef, 대학협회 등)를 중심으로 세부정책을 수립하고 시스템 도입 추진하였음</li> </ul> <p><b>일본, 중국은 정부주도로 시작 또는 확산 단계</b></p> <ul style="list-style-type: none"> <li>- (일본) 내각부와 종합과학기술회의에서 기본정책 수립</li> <li>- (중국) 1998년부터 중국학술정보원(CNKI) 중심 국가프로젝트 수행</li> <li>* 검증 콘텐츠 확보하여 표절탐색시스템 자체개발/시험 운영 중</li> </ul>

※국내외 표절예방활동에 대한 세부현황은 [부록1] 수록

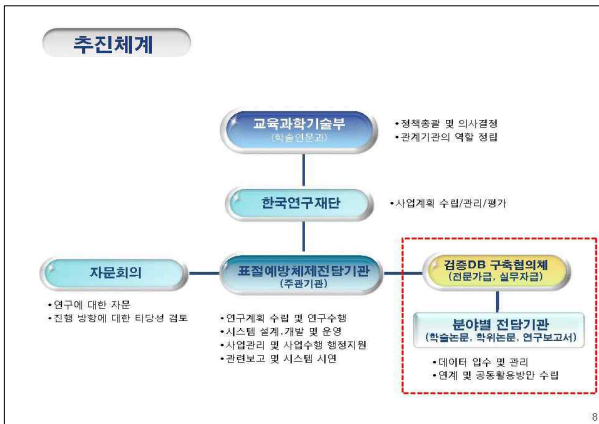
### 과제 개요

**관련추진경과**

- '10.6월 ~ 12월 (정책연구 수행) 표절예방시스템 구축방안 연구
- 표절예방체제 및 표절탐색시스템 구축방향 (안) 수립
- 중점 3개 사업\*을 중심으로 추진
- \* 국내외 현황조사, 국가차원의 체제 정립, 법제도 연구, 검증DB 보유현황 조사, 유사도검색시스템 구축방안
- KISTI 연구원 13명, 외부전문가 10인 (실무TFT 5인, 자문위원회 5인) 참여

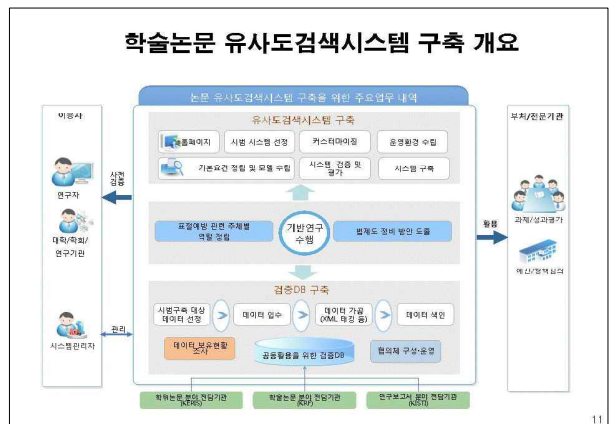
**과제개요**

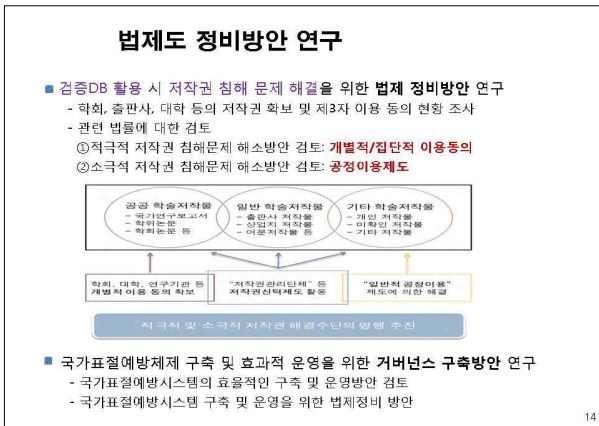
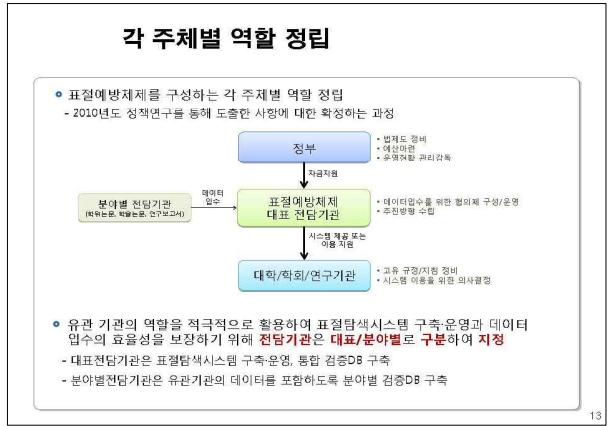
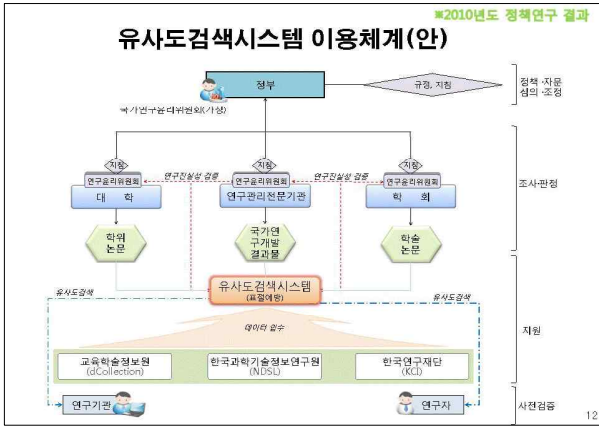
- 사업명/과제명 : 연구윤리활동지원사업 / 학술논문 표절탐색 시험시스템 구축
- 주관부처/발주기관 : 교육과학기술부 학술인문과, 한국연구재단 학술진흥지원팀
- 과제명/기간 : 학술논문 표절탐색 시험시스템 구축, '11.7~'12.4 (10개월)
- 당해연도 사업목표
- 표절예방체제의 효율적인 구축 및 지속적인 활성화를 고려하여 표절예방체제 전담기관 선정하여 시범사업 추진



### 2 연구 목표 및 추진내용

- 연구목표 및 범위
- 주요추진내용





## 기관별 검증 DB 보유 현황 조사

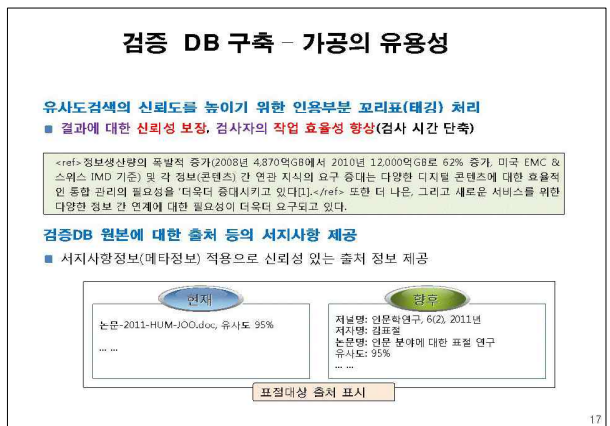
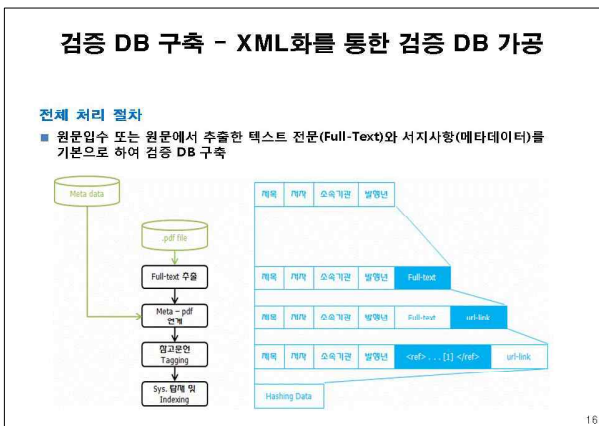
### 기관별 학술저작물 원문 확보현황

(2010년까지의 입수분 기준)

정보 유형	보유 기관	보유건수		분야	비고
		전체	텍스트		
학위논문	KERIS	888,339	244,205	전 분야	~ 2010
	NRF	524,275	524,275	전 분야	2004 ~ 2010
	KERIS	107,276	-	인문사회	2001 ~ 2010
	KISTI	523,078	45,715	전 분야	~ 2010
학술논문	누리미디어	476,655	30,092	전 분야	~ 2010
	소계	1,553,316	587,971		
보고서	KISTI	48,261	8,826	과학기술	1970 ~ 2010

### 저작권 확보

- 학술저작물 원문에 대한 이용동의문의의 해석을 놓고 의견이 분분하나, 영문해제는 이용동의문의에 대한 개량이 필요함
- KISTI는 표절탐색과 관련된 타사업(CrossRef)의 협력을 통한 CrossRef 이용)의 추진을 통해 약 70% 학회(3만5천여건 논문)에 대해 표절탐색을 위한 이용동의체결
- 이 중 48개 학회 54종 저널, 약 1만건을 선정하여 본 연구에 활용



### 학술논문 유사도검색시스템 구축 내용

다음 3단계로 구분하여 학술논문 유사도검색시스템 구축

**STEP 1**

유사도 검색시스템 기본요건 및 모델 수립

- 표절탐색을 위한 유사도 검색시스템의 제공 요건 정립

**STEP 2**

시범시스템 구축

- 유사도 검색 시스템 선정 및 기능 최적화
- 유사도 검색시스템 커스터마이징, 사용자 인터페이스 개발

**STEP 3**

시스템 평가

- 시스템 구축 및 운영 환경 제안
- 개발된 유사도 검색 시스템에 대한 검증 및 평가
- 시연계획 수립 및 시연

※ 구축한 시스템은 과제종료 후 일정기간 유지

18



### 유사도검색시스템 선정

전문가의 의견을 수렴 → 엔진보다는 콘텐츠/문화가 우선임을 확인 → iThenticate 도입

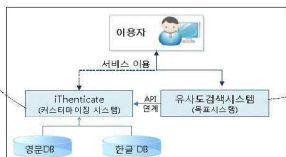
구분	iThenticate (외산)	멤체커(MemChecker) (국산)
서비스 제공 방식 및 특징	<ul style="list-style-type: none"> <li>상용시스템 도입 + 서비스 이용료 지불</li> <li>※시범 운영기간 동안은 서비스 이용료 없음</li> </ul>	<ul style="list-style-type: none"> <li>상용시스템을 구매하여 구축하는 방식</li> <li>※라이선스 범위에 따른 엔진도입 비용 지불</li> </ul>
장점	<ul style="list-style-type: none"> <li>국제적으로 검증된 시스템</li> <li>신속한 표절검색체제 구축 가능</li> <li>Turnitin과 동시 추진 시 대학수준까지의 표절예방체제 구축 가능</li> <li>“방대한 검증DB” 제공</li> <li>15억여 가지(영문: 6주요: 10만/보그서/해태이치/유니클)</li> <li>다국어 처리 및 교차언어검색 지원예정 (한국어 포함)</li> </ul>	<ul style="list-style-type: none"> <li>한글 자료에 대한 표절검색 성능우수</li> <li>표절탐색엔진의 확장(무기능 개발 등)의 용이</li> <li>외국에 대한 기술 및 체제 충족화 방지</li> </ul>
단점	<ul style="list-style-type: none"> <li>체제, 데이터, 기술 등에 대한 외국 종속화 우려</li> </ul>	<ul style="list-style-type: none"> <li>국산 제품으로서 Turnitin보다는 엔진도 및 시스템 성능 검증 기회 부족</li> <li>검증 DB의 무재로 인해서 본 사업에서 구축되는 자료 이외의 외산 학술자료에 대한 표절탐색 불가</li> </ul>
유사제출	-	<ul style="list-style-type: none"> <li>WISE Referee(와이즈넷)</li> <li>XTEN CLONE(오른베이스)</li> </ul>

19

### 유사도검색시스템 커스터마이징, UI개발

• 독립 시스템 제공  
• 독립 DB 제공



이용자 → 서비스 이용 → iThenticate (컨스터마이징 시스템) / PLPS 유사도검색시스템 (독립시스템) → DB

20

### 유사도검색시스템 성능평가

시범 구축한 논문 1만 건을 이용하여 평가를 수행해본 결과,

- 시스템은 전반적으로 우수한 성능을 보임
- 우리나라의 표절 강도를 고려할 때 소정의 목적을 달성할 수 있을 것으로 사료됨
- 한글의 띄어쓰기 문제나 교묘한 형태의 조작은 걸러낼 수 없으므로 이 부분에 대한 개선이 필요할 것으로 보임

**평가방법 (예시)**

- 한글 논문을 대상으로 표절된 문서를 1~5 조각으로 구분
- 해당 문서를 10KB~100KB 사이의 임의 문서의 임의 위치에 끼워 넣고 표절 검사

21

### 유사도검색시스템 시범개발 및 운영 환경 제안

**유사도검색시스템 구성 개요**

연구자, 부처/전문기관, 정책기획자, 정책결정자

서비스 이용

유사도검색 서비스 (참고정보, 사례조사지원 등)

서비스 인터페이스

연진 커스터마이징

시스템기본보급 및 모델 수립, 분석능력향상, 연진 최적화, 커스터마이징

상용 유사도검색 연진

표절탐색 기반 실시간 모니터링, 유사도분석, 유사부분 하이라이팅, 비교분석

원문여유, 해안정보, 부가정보

**운영환경 구성**

Internet

원문여유, 원문확백, 원문저장

보안장벽

LG, LS 스위치

이동통신 장비

검색서버, WEB서버, WAS서버, DB서버

백업장비

NAS, SAN, 백업 디스크, 대용백업

22

### 3 향후 추진방향

- 활용분야
- 중장기 추진계획 (TRM)
- 자기 추진방향(안)

### 활용분야

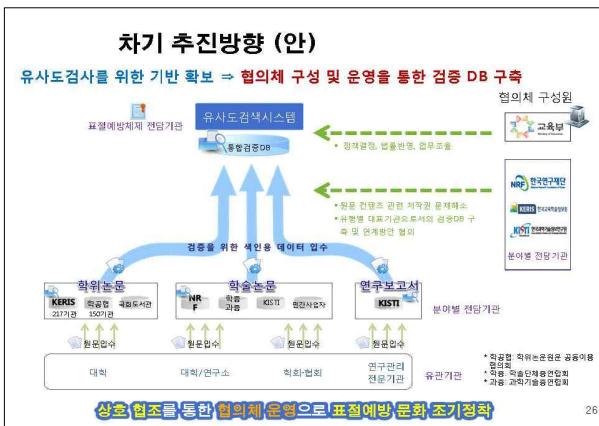
- 자율적 표절예방 문화 조성 및 확산**
  - 표절탐색시스템의 존재만으로도 표절예방에 효과를 발휘
  - 표절검사를 통해 불이익을 증가보다는 **사전검사 도구로 제공**하여 학계 및 연구계의 **자율적 표절예방활동 지원**
- 각종 유관업무에 활용**
  - 연구진실성 검증 절차** 중 예비조사, 분조사, 제조사 과정에서 사용함으로써 심사자의 **표절 판단 효율성을 높임**
  - 각종 학회 및 연구기관의 **논문심사**에 사용
  - 심재물전담기관, 연구관리전문기관, 국가과학기술지식정보서비스(NIS)에 각종성과를 제출 시 표절탐색시스템을 이용한 검사결과를 첨부하도록 함으로써 **성과검증 레벨업**
- 정책수립에 반영**
  - 표절 추이분석**을 통한 **연구유리** 관련 정책수립에 반영
  - 학계 및 연구자들에게 고품질의 표절탐색시스템을 무료(혹은 저가비용)로 제공함으로써, **올바른 연구성과 창출을 통한 국가경쟁력 향상**에 기여함

24

### 중장기 추진 계획 (TRM)

	1차년도	2차년도	3차년도	4차년도	이후
<b>정책 지원</b>	재정용 이용용의(안) 마련 유사도검색시스템 개발/운영	제작용 이용용의(안) 정비 유사도검색시스템 의무이용 강제화 표절예방체제 확장을 위한 국제컨퍼런스 개최 (연 1회)			
<b>DB 구축</b>	검열DB인수 협의체 구성 및 운영 학술논문 데이터 압수 및 가용 24만건~10만건 학술논문 데이터 압수 및 가용 10만건~12만건	검열DB인수 협의체 구성 및 운영 학술논문 데이터 압수 및 가용 24만건~10만건 학술논문 데이터 압수 및 가용 10만건~12만건	검열DB인수 협의체 구성 및 운영 학술논문 데이터 압수 및 가용 10만건~12만건	검열DB인수 협의체 구성 및 운영 학술논문 데이터 압수 및 가용 10만건~12만건	학술대회(한국) 논문 제공 대학/대중 학술논문 데이터 압수 및 가용 10만건~12만건
<b>시스템 개발</b>	유사도검색시스템 이용을 위한 포털 구축 및 운영 유사도검색시스템 시범구축	유사도검색시스템 이용을 위한 포털 구축 및 운영 대표/분야별 전담기관 간의 자동연계 체계 구축	유사도검색시스템 이용을 위한 포털 구축 및 운영 유사도검색시스템 개선	유사도검색시스템 이용을 위한 포털 구축 및 운영 유사도검색시스템 개선	자재연진 개발 및 시험 적용 기반 정비 운영/유지보수

25



### 활용 안내

**연구과제 결과를 중 iThenticate 부분의 활용에 대해 안내드립니다**

- 이용범위:** 정부 후원을 받는 연구자, 대학 내 연구자  
※ 학생 또는 기타 대학 직원은 이용 불가
- 이용목적:** 학문관련 연구용 (제안서 작성, 일반연구, 저널용 논문 작성 등)  
※ 이용목적 외 사용(지체 확인 등)은 불가
- 이용건수:** KISTI가 보유한 총 한도 내에서 이용 가능 (연간 10만개의 문서)  
※ 공동활용에 초점을 맞추고 있어, 현재는 이용량 제한에 대한 정책은 수립하지 않은 상황
- 이용기간:** ~14년말
- 이용비용:** 무료
- 이용방법**
  - ① 이용자가 KISTI에 문의 및 신청서 요청 (choi@kisti.re.kr / joo@kisti.re.kr)
  - ② 이용자가 신청서를 받아 작성 후 제출 (※신청서는 KISTI에서 제공)
  - ③ KISTI에서 ID 발급, 이용자 이메일로 관련 정보 발송
  - ④ 이용자가 사이트(http://plp.i.thenticate.com)에 접속하여 이용

28

### 활용 안내 - cont.

29

**[부록1] 표절예방활동 관련 국내외 현황**

구분	한국	영국	미국	일본	중국
방식	정부 주도형	공공·민간 주도형	공공·민간 주도형	정부 주도형	정부 주도형
단계	발생기	인정기	성장기	발생기	발생기→성장기
	해의 학술논문 및 레포트 표절검색 시스템 도입 운영	대학 수준의 보급은 이뤄졌고 고교 교육으로 진행	지벌적 협의체를 구성하여 대학 및 기관 중심의 적용	정부 차원의 관련 규정·지침수립 정도로 국한됨	-
주체별 역할	정부/공공	공공의 처벌 가이드라인 제시 대학별로 시스템 도입	연방정부의 정책 수립 및 가이드라인 제시 대학의 협의체 구성 /운영 시스템도입	내각의 기본지침 및 각 부처의 세부지침 수립	중앙정부의 강력한 정책 수립 및 시스템 개발 제공
	민간	국내 민간에 경우는 레포트 표절검색 중심 으로 시스템 개발 및 유지보수	표절관련 교육 및 시스템 운영 및 유지보수	-	-
시스템 보급 및 이용현황	해의 논문은 해외 민간 시스템 이용 레포트 표절검색 시스템 도입 운영 시작	대학의 97%가 민간 시스템 이용	대부분의 대학 및 주요 단체에서 민간 시스템 이용	이용률 낮음	이용률 낮음

30

**[부록2] 국가차원의 추진 당위성**

- 1. 연구자층 대상(대학교수/대학원생/연구원 4,305명)으로 한 설문조사 결과 (91.6%)**
  - 국가차원에서 시스템을 구축해야 한다는 지배적인 의견 도출 (91.6%)
- 2. 외국의 선진사례 조사결과, 초기검증체계 마련은 국가에서 추진**
  - 영국은 국가가 초기에 자금을 투입하여 표절예방을 위한 체제 마련
  - 중국은 정부 주도로 표절검사를 위한 DB와 시스템을 구축, 보급
  - 미국은 공공협의체(대학협의체, COPE, CrossRef 등)를 중심으로 세부정책을 수립하고 시스템 도입을 추진
- 3. 관련 규정에 따라**
  - 「국가연구개발사업의 관리 등에 관한 규정」 제30조(연구부정행위의 금지 등) 제31조(연구부정행위의 검증 및 조치)에 따라 중앙부처 및 연구기관의 연구부정행위에 대한 효율적인 판단을 위한 시스템 제공 필요
- 4. 연구 논문 콘텐츠 입수 문제 해소**
  - 세계적인 사용프로그램, 국내의 사용프로그램은 한글논문에 대한 입수 문제로 원활한 서비스를 제공하지 못하는 실정임
  - 국가차원(정부 또는 공공기관)에서 추진하면 분산 구축되어 있는 콘텐츠 확보 용이
  - 정부주도의 조속한 표절예방 문화 정착 필요

31