

Digital Library 구축 기술동향

1998. 9. 22

김 성 혁

숙명여대 정보과학부 문헌정보학과
첨단학술정보센터

첨단학술정보센터 <http://www.kric.ac.kr/>

1

목 차

1. Digital Library(DL) 등장 배경
2. DL 개념
3. DL 기술
4. DL 요소 기술 동향
5. 국내 DL 기술 동향
6. 결론

첨단학술정보센터 <http://www.kric.ac.kr/>

2

1. DL 등장 배경[1]

1.1. 도서관 내적 요인

- 인쇄매체의 증가로 인한 물리적 공간 부족
- 이용자의 다양한 요구에 대응하는 서비스 부족
- 도서관 자원 부족
- 전문사서 부족
- 인쇄매체의 보존성 결여로 인한 자료 손실

1. DL 등장 배경[2]

1.2. 도서관 외적 요인

- 대학 교육의 변화
- 정보의 양과 생산 비용 증가
- 정보 기술의 발전
- 유사 도서관의 범람
- 전자문헌의 등장 및 양산 체제
- 지식정보화사회로의 환경 변화

1. DL 등장 배경[3]

1.3. 정보 기술 발전

- Tera, Peta Byte의 대규모 정보 기록
- 대규모 정보에서 필요한 정보 추출 기술 향상
- 네트워크상에서 시스템간 상호 운용성 향상
- 정보 공유로 인한 정보공간 및 접근력 확대
- DL이 지식정보사회의 이상적 응용 시나리오로 등장

1. DL 등장 배경[4]

1.4. DL 추진 사업 / 기관

- 1994 Digital Library Initiative(DLI), NSF/DARPA/NASA
- 1998 DLI-2
- 1998 International Digital Libraries Cooperative Research Program, NSF
- European Research Consortium for Informatics and Mathematics(ERCIM) : 유럽판 DLI, 추진 : DELOS WG
- DL Working Groups : DELOS WG - DLI 연합

2. DL 개념[1]

2.1. DL 정의[1]

- 정보를 수집, 디지털 형태로 저장, 구조화하며, 사용자에게 친근한 방법으로 네트워크상에서 정보를 검색, 처리하는 시스템(NSF)
- 디지털화 된 도서를 네트워크에 연결시킨 저장소로 전문 및 메타 정보를 검색, 취득할 수 있는 도서관(김성혁)
- A society of agents(U. of Michigan)

2. DL 개념[2]

2.1. DL 정의[2]

- DL은
 - 산재되어 있으며
 - 전세계적인 접근이 가능하며,
 - 분산되었으며,
 - 이질적 정보이며,
 - 문제 해결의 장소(Belkin, 1998)

2. DL 개념[3]

2.1. DL 정의[3]

- DL은 단순히 디지털문헌과 이의 처리에 필요한 도구의 집합이 아니라 정보의 제작, 분류, 사용, 보존의 전 과정(life cycle)을 지원하는 장서, 서비스, 인간이 공존하는 환경(Paul Duguid, 1998)
- 관점에 따라 다양하게 정의되지만 문헌이 중심
- 문헌: 멀티미디어 디지털문헌

첨단학술정보센터 <http://www.kric.ac.kr/>

9

2. DL 개념[4]

2.2. 미래 디지털 도서관[1]

- 물리적 개념에서 논리적 개념
- 정보의 소유에서 공유
- 자료의 디지털화
- 이질적 정보의 접근과 시스템간 상호 운용성
- 인쇄매체와 디지털문헌의 상호 보완 관계

첨단학술정보센터 <http://www.kric.ac.kr/>

10

2. DL 개념[5]

2.2. 미래 디지털 도서관[2]

- Web을 이용한 Client/Server 구조
- Tera, Peta Byte 정보
- MARC에서 다양한 메타데이터로
- 이용자 중심 인터페이스
- 다국어/다문화 지원 S/W
- 다양한 유형의 정보 제공

첨단학술정보센터 <http://www.kric.ac.kr/>

11

3. DL 기술[1]

3.1. DLI-2에서 제안된 기술[1]

- 시스템 관련 기술(Systems-Centered Research)
 - Scalability : 대규모 정보의 축소 기술
 - Interoperability : 시스템간 상호 운용성 기술
 - Adaptability and Durability
 - 다양한 사용을 위한 정보의 융통적이며 동적 변화 지원
 - Support for Collaboration
 - 정보의 Lifetime 지원을 위한 공동 협력

첨단학술정보센터 <http://www.kric.ac.kr/>

12

3. DL 기술[2]

3.1. DLI-2에서 제안된 기술[2]

- 장서 관련 기술(Content and Collection-Based Research)
 - 장서의 구조화, 인덱싱, 메타데이터
 - 정보의 수집, 사용, 인증
- 사용자 관련 기술(Human-Centered Research)
 - 문화, 언어의 장벽을 넘는 전세계적 정보 하부 구조
 - Interactive DL

첨단학술정보센터 <http://www.kric.ac.kr/>

13

3. DL 기술[3]

3.2. DL Working Groups[1]

- DELOS WG - DLI community에서 추진
- 5개 분야 DL Working Groups
- 연구 방향
 - Open issues 연구,
 - 미래 연구 방향에 대한 제안서 작성
- 현재 상태 : draft version reports

첨단학술정보센터 <http://www.kric.ac.kr/>

14

3. DL 기술[4]

3.2. DL Working Groups[2]

- **Interoperability**
 - 정보원과 사용자 사이 중개안 역할
 - 다양한 형태의 정보를 사용자에게 적절한 형태로 재구성 기술
 - DL sources와 서비스 사이 일관된 시각 제공
- **Metadata :**
 - 네트워크에서 자원 탐색, 교환하기 위한 집합
 - Ex. : Dublin Core, GILS, TEI Independent Header

청단학술정보센터 <http://www.kric.ac.kr/>

15

3. DL 기술[5]

3.2. DL Working Groups[3]

- **Intellectual Property and Economics(IPE) :**
 - 정보의 손쉬운 이동, 복제 등 부가가치 정보 서비스를 통한 정보의 법률적, 경제적, 기술적 이슈 도출 및 처리
- **Resource Indexing and Discovery(RID) in a Globally Distributed Digital Library**
 - 분산된 DL에서 정보의 검색과 인덱싱 문제

청단학술정보센터 <http://www.kric.ac.kr/>

16

3. DL 기술[6]

3.2. DL Working Groups[4]

- **Multilingual Information Access(MLIA)**

- 다국어 정보에 대한 저장, 접근, 표현
- multilingual/multicultural 정보에 대한 평가와 연구 위한 하부 구조 구축
- 효과적인 MLIA를 위한 새로운 기술 개발 및 보급
- 소규모 다수 연구자를 위한 언어적 자원 시장 형성

4. DL 요소 기술 동향[1]

4.1. Interoperability[1]

- **Introduction**

- 이질적 정보원과 client 사이 상호 운영 서비스 제공
- 정보원은 접근 방법과 모델에 따라 변형
- 자연어, QBE(Query By Example) 등을 지원하는 DL
- 즉 Interoperability는 사용자와 DL 사이 중재자 역할

4. DL 요소 기술 동향[2]

4.1. Interoperability[2]

- 정보 모델
 - 일관된 시각 제공 위한 자료 구조와 의미 표현
 - Middleware data model 지원
 - 자료 형태의 이질성을 다루는 모델 연구
 - Domain specific ontologies and thesauri
 - 정보과 사용자 사이 의미적 이질성 조정
 - 지식표현과 추론시스템 사용

첨단기술정보센터 <http://www.kric.ac.kr/>

19

4. DL 요소 기술 동향[3]

4.1. Interoperability[3]

- 연구 topics
 - 지적 중재자 서비스를 위한 메타데이터 : 질의를 위하여 적당한 source 선택, 최적화를 도와 줌.
 - middleware 정보 모델 : INRIA, FORTH,
 - Ontology : U. Michigan,
 - Metadata interoperability : U. Stanford

첨단기술정보센터 <http://www.kric.ac.kr/>

20

4. DL 요소 기술 동향[4]

4.1. Interoperability[4]

- 조정과 통제
 - 중재자는 사용자에게 일관된 시각 제공
 - 다양한 변화에 적용할 수 있는 조정 메카니즘과 source 기술 언어 : U. Michigan
 - 설명 서비스, 시각화 : FORTH

4. DL 요소 기술 동향[5]

4.1. Interoperability[5]

- 질의어 처리 : 이질적 DL 사이 분산 질의
 - 질의어 형식 : 다양한 질의 형식 존재
 - 질의어 형식과 서비스 정의: Stanford U.
 - 이질 정보사이 멀티미디어 정보 검색 : CNR
 - 질의어 최적화 : 다른 시스템은 다른 질의 형식 요구
 - 중개 층에서 질의어 : 시스템 최적화, 신뢰성 제공 : CNR
 - Relevance feedback : 질의 결과의 quality 향상

4. DL 요소 기술 동향[6]

4.1. Interoperability[6]

- 구현
 - 다양한 형태의 장서와 다양한 프로토콜, 다양한 질의와 결과를 처리 -> CORBA
 - 자료 교환으로 CORBA 하부구조 제안
 - 예 : ETH, U. Michigan, Stanford

4. DL 요소 기술 동향[7]

4.1. Interoperability[7]

- 레벨
 - 의미적 상호 운용성 : 다양한 시스템의 의미적 일관성을 유지하기 위한 표준 종재 S/W 사용
 - 구문적 상호 운용성 : 표준 형식으로 RDF나 HTML의 Meta tags
 - 구조적 상호 운용성 : 자원 기술하는 문화적 차이에 대한 지원
 - 이름 쓰는 순서나, 한 필드에 사용 가능한 값들의 올바른 기술법

4. DL 요소 기술 동향[8]

4.2. Metadata[1]

- Introduction

- 메타데이터 : 데이터에 관한 구조화된 데이터
- Web에서 문서 기술 정보는 HTML의 Meta tag 사용
- indexing robots 이용한 메타데이터의 자동 수집
- 문서와 메타데이터의 분리 가능

4. DL 요소 기술 동향[9]

4.2. Metadata[2]

- 메타데이터 보존소
 - 인덱싱과 보존소의 자료 사용에 관한 메카니즘 정의
- 메타데이터 생성 도구
 - 비전문가도 만들 수 있는 정보의 생성, 추출, 운영 등에 관한 도구
- Resource Description Framework(RDF)
 - 참조되는 자료를 정확하게 연결하는 상호 운용성 제공

4. DL 요소 기술 동향[10]

4.2. Metadata[3]

- **Registres**
 - 인간과 기계위한 형식과 다른 스키마에 대한 사용허가, 법률적 가치, 사상을 위한 등록 구조 제공
- **Interoperability and complexity**
 - Dublin Core 제공 : 단순하고 pidgin-like 표준 메타데이터
 - pidgin : 다른 언어 사용자간에 공동 작업위한 단순한 언어 집합
 - 복잡성 제어 위하여, ontology에 연결 기능 제공

4. DL 요소 기술 동향[11]

4.2. Metadata[4]

- **Common Metadata space**
 - Dublin Core는 다수의 보관소 사이 구조, 항해, 관리, 인증, 출처, 등에 관한 핵심 원소 제공하므로, 다양한 domain 사이 상호운용성 향상위한 의미 공간 역할
- **복잡한 자원 위한 메타데이터**
 - 장서 모임, 동적 자료, 오디오, 비디오 등에 관한 메타데이터 개발 필요

4. DL 요소 기술 동향[12]

4.2. Metadata[5]

- 평가와 측정
 - 메타데이터는 RDF를 통하여 전개되며 자료 추적 시스템 개발 필요
 - 다양한 시스템의 메타데이터 사용 종류 측정
 - 어떤 종류의 메타데이터가 효율적인지 평가

4. DL 요소 기술 동향[13]

4.2. Metadata[6]

- 정책
 - 국경을 넘은 메타데이터에 대한 법률적 문제
 - 메타데이터의 통합과 신뢰성
 - 부정확성을 탐지하기가 쉽지 않다.
- 자원 탐색에 메타데이터 사용
 - 전통적 목록과 인덱싱을 사용한 메타데이터의 효과적으로 통합

4. DL 요소 기술 동향[14]

4.2. Metadata[7]

- Working group
 - W3C의 Data Object Model
 - 다수 자원에 대한 질의 결과가 다양한 레벨로 나타남으로 결과를 효과적으로 평가하기 어렵다. 일치된 결과 모델 연구
 - W3C의 RDF
 - 메타데이터 교환을 지원하는 표준 구문

4. DL 요소 기술 동향[15]

4.2. Metadata[8]

- Task Force
 - W3C의 Time-based media : audio, video
 - RTSP : 네트워크 상에서 자료 흐름 통제
 - RTP : 멀티미디어 자료에 대한 패킷 형식 제공
 - SDP : 멀티미디어 세션의 초기화 정보
 - SMIL : 복잡한 멀티미디어 구성을 위한 언어

4. DL 요소 기술 동향[16]

4.3. IPE[1]

- 디지털 정보
 - 정보 손실과 추가 비용 없이 복사 가능
 - 누구나 구축가능
 - 기존 정보 제공자와 경쟁
 - 출판사와 도서관의 역할이 협조 체제로 전환
 - 그러므로 IPE가 등장

4. DL 요소 기술 동향[17]

4.3. IPE[2]

- 법률적 이슈
 - 기존 법률이 다루지 않으며, 해석의 위험 초래
 - cached copies : client 시스템이 정보의 서비스를 위하여 내부적으로 download(임시 복사) 수행
 - cached copies는 true copy가 아니다
 - lifetime value : server에 연결 시 만 임시 복사 허용
 - server연결 상태에서 네트워크 장애 시: 처리를 어떻게 할 것인가?

4. DL 요소 기술 동향[18]

4.3. IPE[3]

- Access

- public 시스템에 protected 문서 기재 부적절
- 시스템 레벨에서 access를 제한
- 접근 통제 S/W 백업과 침입 방지 S/W설치 요망

4. DL 요소 기술 동향[19]

4.3. IPE[4]

- Traversing, Indexing, Archiving

- 공개된 인터넷 자원에 대한 indexing 및 사용 시
 - 소유자에게 사용을 알림, 소유자는 접근 권한 분류, 왜냐하면 자료를 모아서 재생시키는 것은 소유권 침해
 - hyperlink, reference는 copyright 보호하에서만 가능
 - » 초기 public인데 나중에 copyright된 것은?

4. DL 요소 기술 동향[20]

4.3. IPE[5]

- Privacy

- 도서관 사용자의 사용 정보 공개 금지
- DL S/W 개발자의 불법행동
 - 사용자 profile을 타 도서관에 제공 /판매
 - 사용자 관심분야와 정보의 적정가격 자료를 제공

4. DL 요소 기술 동향[21]

4.3. IPE[6]

- Virtual space, Agent, Active object

- 사용자의 요구는 다른 S/W를 부르고, 그것은 또 다른 S/W를 수행 시킬 수 있다.
 - 이때 정보의 불법 사용은 개인이 위반인가? S/W제공 기관이 위반인가?
 - 비록 S/W 제공이 불법이지만 그것을 작동시킨 개인이 위반인가?

4. DL 요소 기술 동향[22]

4.3. IPE[7]

- 경제적, 정치적 이슈
 - Open costs : 자료의 분류, 관리, 목록화에 비용소모
 - Revenue model : 복사비가 거의 없으며 가격이 복사량에 비례하는 것은 부적절-> 가격 등급 제시
 - bundling 상품 개척
 - 예 : 저널은 여러 volumes, 각 volume은 여러 articles, 각 article은 여러 조각(header, abstract, section, ...)이므로 abstract만 묶어서 상품화

첨단기술정보센터 <http://www.kric.ac.kr/>

39

4. DL 요소 기술 동향[23]

4.3. IPE[8]

- Market positioning
 - 누구나 정보 제공, 정보 검색 시 first class 정보를 가진 sites만 이용, 도서관 역할 애매
- Incentives
 - 전자 자료 구축을 위한 저작자 인센티브
 - 불법 시 지불하는 벌금과 사용요금 사이 균형 유지
 - 위의 시장 원리에 의한 인센티브 제공

첨단기술정보센터 <http://www.kric.ac.kr/>

40

4. DL 요소 기술 동향[24]

4.3. IPE[9]

- 기술적 측면
 - 정보 유통의 전체 사이클 지원 모델 개발
 - 정보의 lifetime 동안 암호화 제공
 - Digital Encryption Standard, Public key Cryptography
 - 저자의 (전자) 사인이 있는 문서

4. DL 요소 기술 동향[25]

4.3. IPE[10]

- 계약
 - Kerberos Architecture : 인증과 접근정보를 가진 토큰이 시스템간 교환되고, 허가를 받으면 무한정 이용
 - Kerberos Architecture 의 변형으로 융통성 있는 계약, 인증, 접근 권한, 가격이 필요
 - Resource Discovery Framework
 - Standard for the Exchange of Product Data(STEP)
 - » 각 정보에 대한 사용항목과 조건을 가진 토큰 사용
 - » 사용에 관한 가격, 허가 등의 감시 메커니즘 제공

4. DL 요소 기술 동향[26]

4.4. Resource Indexing and Discovery in a Globally Distributed Digital Library(RID) [1]

- Introduction

- 전세계 분산된 DL 자원 인덱싱과 발견
 - local 처리의 제반 문제를 Internet에 분산된 환경으로 확대 적용

4. DL 요소 기술 동향[27]

4.4. RID [2]

- 중앙 집중식 인덱싱과 검색
 - Alta-Vista 같이 대부분의 web 인덱싱 시스템 사용
 - 주기적으로 갱신
- 분산 인덱싱과 중앙 집중식 검색
 - Harvest Information Discovery and Access System에서 사용
 - 지역 인덱싱 정보를 검색 엔진에 전달

4. DL 요소 기술 동향[28]

4.4. RID [3]

- 분산 인덱싱/분산 검색
 - Dienst/NCSTRL System에서 사용
 - 인덱싱 엔진, 사용자 검색의 분산

4. DL 요소 기술 동향[29]

4.4. RID [4]

- Scalability
 - 많은 정보를 Task에 적절한 규모로 조정
- Reliability
 - 분산 시스템은 네트워크 자원의 신뢰성을 요구
 - 중앙 집중화 된 구조는 시스템 failure나 새로운 인덱스 정보 갱신 문제 발생

4. DL 요소 기술 동향[30]

4.4. RID [5]

- **Heterogeneity**
 - 다양한 자원, 이질적 검색 엔진에 대한 질의어와 수행 결과에 대한 조정
- **서비스 질**
 - 각 사이트의 검색 응답 시간의 다양성에 대한 적절한 서비스

4. DL 요소 기술 동향[31]

4.4. RID [6]

- **Right management**
 - 가치 있는 정보는 접근 리스트에 의하여 보호 되고 있음
 - 검색은 그 내용의 관리자에 의해 허가된 사이트 만 가능
- **특정 영역**
 - 중앙집중식 인덱싱은 자원 발견위한 일반 모델을 가짐
 - 복잡한 과정을 거쳐서 정보를 발견
 - 결과적으로 내용과 분리된 전문화된 인덱싱 자료 구축

4. DL 요소 기술 동향[32]

4.4. RID [7]

- 질의어의 선택적 분배
 - Completeness
 - 질의어의 결과를 돌려줄 시스템을 선택하여 질의
 - 내용 기반 선택
 - GLOSS 시스템, STARTS 프로토콜

4. DL 요소 기술 동향[33]

4.4. RID [8]

- Preferred :
 - 시간 : 빠른 시간에 결과 도출 사이트 선정
 - 비용 : 적은 비용 지불 사이트 선정
 - 최신성 : 가장 최근 정보를 가진 사이트 선정
 - 문제 :
 - metadata : 인덱스 사이트의 메타 정보 필요
 - metadata 배분 : 선호되는 질의 처리에 필요한 분산 정보 제공
 - Routing : 선호도, 메타 정보를 이용하여 각 사이트 노선 결정

4. DL 요소 기술 동향[34]

4.5. Multilingual Information Access(MLIA) [1]

- Introduction

- 하부 구조 구축
 - 다수언어, 문화 정보 접근을 위한 평가와 연구
- 새로운 기술 개발
 - 효율적 MLIA를 위한 새로운 기술 필요
- 언어 자원 시장
 - 많은 수요자를 위한 언어 자원 시장 구축

첨단기술정보센터 <http://www.kric.ac.kr/>

51

4. DL 요소 기술 동향[35]

4.5. MLIA [2]

- 목표

- 효율적 MLIA 요구사항 분석
- MLIA의 state-of-the-art 기술 평가
- 중요한 연구 항목 제시
- 다수 언어 처리 위한 제안서 제공

첨단기술정보센터 <http://www.kric.ac.kr/>

52

4. DL 요소 기술 동향[36]

4.5. MLIA [3]

- 인터넷에서 문제
 - 외국어로 쓰여진 정보 발견
 - 외국어로 쓰여진 정보의 해석
 - 다른 언어 정보와 혼합
 - 다수 언어 처리의 제안서 제공

4. DL 요소 기술 동향[37]

4.5. MLIA [4]

- 범위
 - 자료 교환: 인코딩, 디스플레이, 브라우저
 - 정보 처리: 다수 언어간 NLP 기술
- 공동 연구 분야
 - 정보 검색: 질의 처리, 검색, 메타데이터 만들기, 인덱싱
 - 기계번역: **comparable & parallel** 텍스트 조정, 언어간 기술
 - 전산 언어학: 형태소 분석, 파싱, 의미 분석, 언어 생성

4. DL 요소 기술 동향[38]

4.5. MLIA [5]

- 문서 처리 : 문헌의 분류, 필터링, 분리
- 자원 : 사전, 시소러스의 구축과 유지, 어휘 사전, ontology
- HCI : 문서의 시각화, 다수 문서 요약
- 기술
 - 인덱싱 도구 : 다수언어 인덱싱 도구 개발
 - 질의어 처리 도구 : 다수언어 측면에서 연구
 - 문서 분류 : 교차언어 문서 검색위한 문서 분류
 - 요약 : 다수 언어 검색 결과 요약
 - 시각화 도구

첨단학술정보센터 <http://www.kric.ac.kr/>

55

5. 국내 DL 기술 동향 [1]

- Interoperability
 - Z39.50 기술(학술정보서비스시스템, 국가 전자도서관 사업)
 - CORBA (Quality-of-Service Management Framework-POSTECH)
- Metadata
 - Dublin Core (학술지논문 종합목록 시스템)
 - TEI Independent Header(21세기 세종 프로젝트)
 - 국가 지리 정보 체제(NGIS) 사업-한국전산원
- IPE
 - 저작권 관리 시스템 (국가 전자도서관 사업)

첨단학술정보센터 <http://www.kric.ac.kr/>

56

5. 국내 DL 기술 동향[2]

- RID
 - Web 검색기(NAVER, 심마니)
- MLIA
 - 기계번역(MATE/E-K)

6. 결론[1]

6.1. 국내 현황

- DL 관련 기술에 대한 기초연구 및 응용개발이 독자적이고 소규모로 추진 중
- 많은 기관이 이미지 기반으로 원문 구축
- 표준화 없이 정보 구축
- 기관 및 시스템간 상호 운용성 보장 없음
- 정보 인프라 미비

6. 결론[2]

6.2. 앞으로 DL 연구/개발 방향

- DL: 지식정보화 사회의 기반구조
- 국가 주도하의 DL 연구/개발
 - 장기간의 기초 연구와 응용 연구
 - 국가적 표준화 요구
 - 정보 하부 구조 구축
- 정보통신 기업의 DL 연구 및 개발 참여 유도
 - 관련기업, 출판사, 도서관, 저작권 협회의 적극적 참여
 - 정보 구축/서비스 관련 기술 개발
 - 정보 경제화 도입