디지털 라이브러리 모형에 관한 연구

A Study on Modeling of Digital Libraries

이 창 열(Chang-Yeol Lee)*

목 차

1. 서론
2. 디지털 라이브러리의 구조
3. 디지털 라이브러리의 구성
   3. 1 상호운용성
   3. 2 메타데이타
   3. 3 지적 소유권
   3. 4 분산 정보검색
   3. 5 다국어 처리
   3. 6 문헌 표현
4. 디지털 라이브러리의 사례
   4. 1 웹 가상 도서관
   4. 2 OCLC
   4. 3 ELRA
   4. 4 일리노이 대학
   4. 5 사례 비교
5. 우리 나라의 디지털 라이브러리
6. 결론

초 록

전통적 도서관은 인터넷 기술의 발달로 인하여 새로운 형태의 디지털 라이브러리로 변모하고 있다. 책은 디지털 파일로, 목록은 메타데이타로, 이용자는 네트워크가 접속되는 전세계 이용자로, 자료는 무한정 복사할 수 있는 형태로 바뀌고 있다.

국가 정보 하부구조를 위한 디지털 라이브러리 구축은 미래 정보화 사회에 매우 중요한 것으로, 본 논문에서는 기존의 디지털 라이브러리 기술에 대한 분석을 통하여 앞으로 추진할 디지털 라이브러리의 모습을 모델링한다.

ABSTRACT

With the advancement of Internet technology, traditional libraries are going through a new metamorphosis into digital libraries. Digital libraries substitute digital files for papers, metadata for catalogues, world wide users on the network for localized patrons, while offering limitless possibilities for easy downloading of information. The aim of the study is to present a model for our digital library based on the knowledge acquired from the analysis of the state of the art technology employed in building and managing digital libraries. It is our belief that the national information infrastructure for digital libraries are mandatory for an open information society.

* 첨단학술정보센터 연구개발부 선임연구원
접수일자 1998년 11월 5일

1. 서론

디지털 라이브러리(Digital Library: DL)는 인터넷 자원에 대한 정보 저장소(Data Repository)로써, 전통적 도서관과는 다른 형태(formats)의 패러다임(paradigm)을 가지고 있다. 정보 저장소의 정보는 전통적 책에 비하여 빠른 변화와 재구성이 가능하고, 이용자의 다양한 요구나, 시각에 대하여 능동적 대처가 가능하다. 이들 정보는 다양한 언어로 기술되어 있으며, 다양한 시스템에서 표준화된 기술 방식으로 재편되고 있다. 사용자는 단일 인터페이스를 사용하여 인터넷에 분산된 수많은 디지털 라이브러리로부터 원하는 정보를 쉽게 얻을 수 있다.

그것은 전통적 개념의 도서관이 단순히 컴퓨터로 장소만 이동한 것으로 단정할 수 없는 부가적인 의미를 가진다. 디지털 라이브러리에서 정보는 능동적이다. 정보는 원하는 사용자 시각에 맞게 재편되어 제공될 수 있으며, 인터넷으로 자유롭게 이동하면서, 모아지기도 하고, 흩어지기도 하며, 사라지기도 한다.

이러한 정보에 대한 저장소로써 디지털 라이브러리는 정보를 구축하는 기술과 처리하는 기술에 부가적으로, 정보 복사에 대한 경제적 보상과 법률적 보호를 필요로 한다.

본 논문에서는 정보의 저장소이며, 제공처로 디지털 라이브러리가 가져야 하는 구조와 기술을 통하여, 우리가 구축하여야 할 디지털 라이브러리의 모습을 모델링한다. 특히 DLI(Digital Library Initiative) 프로젝트와 ERCIM(European Research Consortium for Informatics and Mathematics)의 DELOS Working Group이 공동으로 작업하는 디지털 라이브러리의 연구 분야를 바탕으로 세계적 기술 동향을 살펴보고, 우리 나라에 적절한 디지털 라이브러리의 모습에 관하여 연구하기로 한다.

2. 디지털 라이브러리의 구조

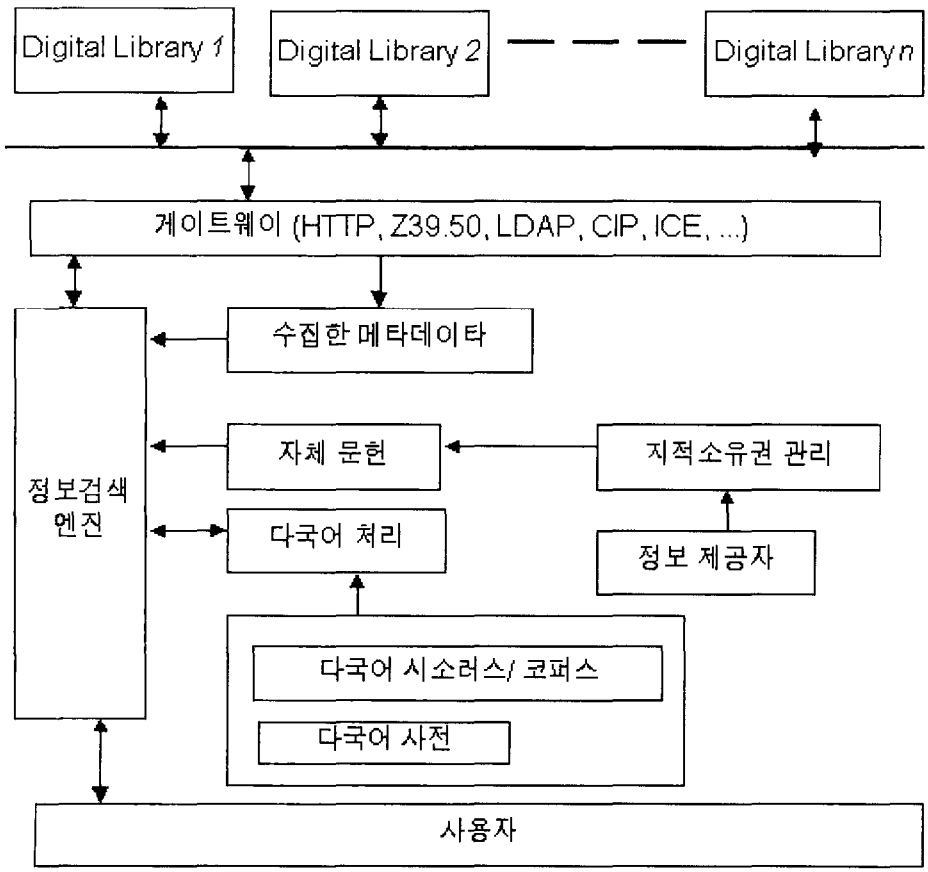
1차 DLI 프로젝트를 추진한 미국의 6개 대학이나, 미국 국회 도서관(Library of Congress: LC)의 전자도서관 프로젝트 등에서와 같이 디지털 라이브러리의 구조는 다양한 형태를 가지고 있다. 디지털 라이브러리의 형태가 다양하지만, 이들 기관이 가지는 공통적인 모습과, 미래의 디지털 라이브러리가 가지는 중요한 요소로 다국어(Multilingual) 처리와 전자 상거래(Electronic Commerce)를 고려하여 본 논문에서는 〈그림 1〉과 같은 디지털 라이브러리 구조를 제안하며, 각 모듈(module)을 중심으로 기술적인 동향을 살펴보기로 한다.

〈그림 1〉에서 제안하는 디지털 라이브러리는 '지적 소유권(Intellectual Property) 관리' 모듈, '자체 문헌'(디지탈 라이브러리에서, 보통 Repository라고 부른다) 모듈, 웹(Web, WWW)을 통하여 '수집한 메타데이타(Metadata)' 모듈, 정보검색 하부구조로 '다국어 처리' 모듈, '게이트웨이'(Gateway) 모듈, 그리고 '정보 검색(Information Retrieval) 엔진'으로 구성된다.

'정보제공자'는 디지털 라이브러리의 '지

적 소유권 관리' 모듈을 통하여 일반 사용자에게 자신의 정보를 서비스 할 수 있으며, 사용자는 질의를 통하여 원하는 정보를 얻을 수 있다. '정보 검색 엔진'은 '자체 문헌'으로부터 원하는 문헌을 추출하거나, 웹을 통하여 '수집한 메타데이타'로부터 원하는 정보를 얻거나, 또는 직접 '게이트웨이' 모듈의 표준 프로토콜(Protocol)을 사용하여 검색(예를 들어 Z39.50을 통한 검색)한다. 특히 웹(<그림 1>의 게이트웨이를 통한)을 통한 질의는 다른 디지털 라이브러리에서 사용하는 언어와 질의 언어와 다를 수 있으므로 '다국어 처리' 모듈을 통하여 목표 언어로 번역 후 질의를 보내고 또한 질의결과를 받기도 한다.

이러한 분산된 환경 하에서 디지털 라이브러리 시스템은 이질적 다른 시스템과 언어, 사용 프로토콜, S/W, 메타데이타 등의 형식이 서로 다를 수 있으므로, 이들을 보완해주는 상호운용성(Interoperability)과 수많은 자료를 사용자에게 적합한 분량으로 축소 제공하는 기능(Scalability)을 가져야 한다.



<그림 1> 디지털 라이브러리 구조

메타데이타는 네트워크에서 자원 탐색과 교환을 하기 위한 집합으로 문헌에 대한 서지 정보의 역할을 한다. 게이트웨이의 표준 프로토콜을 사용하여 다른 디지털 라이브러리의 문헌에 대한 메타데이타를 수집하여 축적하도록 한다.

지적 소유권은 정보 제공자로부터 받은 정보에 대하여 부가가치 정보 서비스를 제공하며, 정보의 법률적, 경제적, 기술적 이슈를 도출하여 처리 및 관리하는 기능을 제공한다.

다국어 처리 기능은 소오스(Source) 언어(질의 언어)를 목표 언어로 변환하기 위하여 기반이 되는 다국어 시소러스(Thesaurus)/코퍼스(Corpus), 그리고 다국어 사전을 필요로 한다. 인터넷에서 나타나는 상호운용성 중에서 언어적 차이는 본 모듈을 통하여 해결한다.

3. 디지털 라이브러리의 구성

디지털 라이브러리는 인터넷에 분산된 자원을 사용자에게 단일의 인터페이스를 통하여 제공하며, 분산된 자원을 자유롭게 수집·제공하기 위하여 (1)과 같은 기술적 문제에 대한 해결을 제시하여야 한다. 본 논문에서는 이들 문제를 디지털 라이브러리 구조의 각 모듈과 연계하여 살펴보기로 한다.

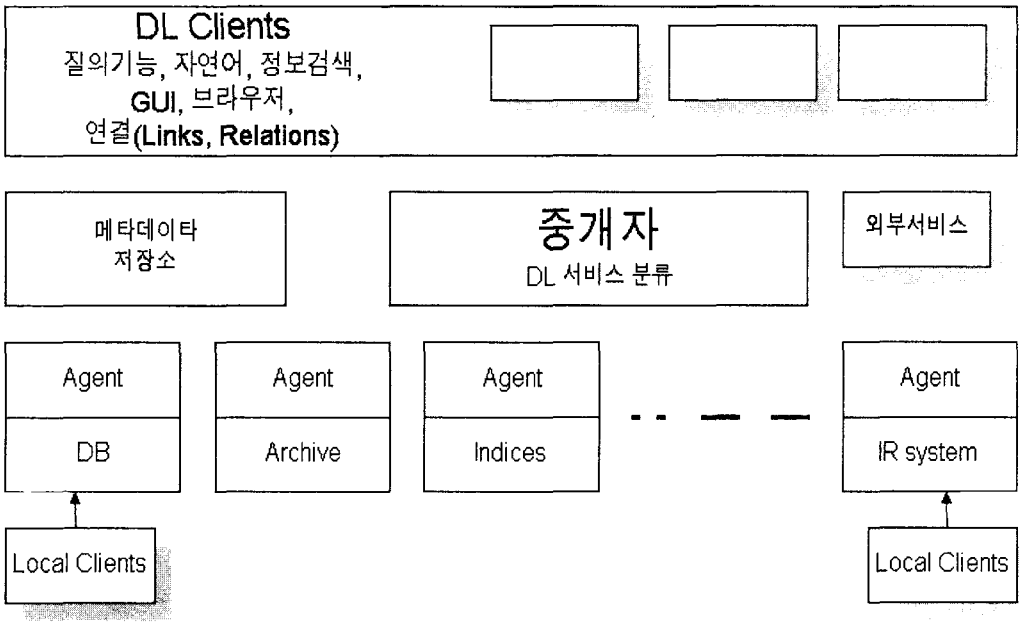
(1)- 상호운용성
- 메타데이타
- 지적 소유권
- 분산 정보검색
- 다국어 처리
- 문헌 표현

3. 1 상호운용성

상호운용성은 서로 다른 형태의 디지털 라이브러리 사이에 통합과 관련된 시각을 제공하는 기능이다. 즉 임의의 디지털 라이브러리가 제공하는 다양한 형태의 문헌에 대하여, 사용자에게 해당 문헌에 대한 단일 인터페이스를 제공한다. 이러한 방식의 통합은 커다란 형태의 분산된 디지털 라이브러리를 구축할 수 있게 한다.

〈그림 2〉는 분산 시스템에 대한 상호운용성을 나타낸 구조이다(Interoperability, 1998). 디지털 라이브러리의 사용자(DL Clients)는 GUI(Graphical User Interface) 화면에서 자연어를 통하여 질의를 하면, 인터넷상에 있는 DB, Archive, Indices, 등의 시스템(〈그림 1〉에서 Digital Library 1, 2, ..., n에 해당)으로부터 원하는 정보를 찾은 후 중개자를 거쳐서 사용자에서 제공한다. 중개자는 사용자가 원하는 정보가 있는 위치, 쓰여진 언어, 문헌 형태 등에 무관하게 정보를 제공받을 수 있게 한다. 〈그림 2〉의 상호운용성을 제공하기 위하여 (2)의 항목을 처리하는 기능을 제공하여야 한다.

(2)- 데이터 모델
- 중개와 통제
- 질의 처리
- 구현 메카니즘



〈그림 2〉 디지털 라이브러리의 상호운용성을 위한 구조

가. 데이터 모델

여러 서버 사이에 단일 시각을 제공하기 위하여 통합된 데이터 구조와 의미 표현이 가능하여야 한다. 데이터 모델을 기술하는 언어는 최대의 융통성과 상호운용성을 위한 존재구조적(ontological) 기반을 가져야 한다.

나. 중개와 통제

데이터 모델 기술 언어는 새로운 데이터 형태의 추가를 허가하여야 하며, 정보 하부구조(Infrastructure), 서비스, 인터페이스, 또는 기능의 변화에 대한 중개 역할을 고려하여 설계하여야 한다. 디지털 라이브러리의 구조 변경이나, 데이터 갱신으로 인한 불 일치성을 해결할 수 있는 시스템의 형태를 갖추어야 한다.

다. 질의 처리

질의 처리는 질의어 형식, 다른 시스템으로 질의어 전송, 변환 그리고 결과에 대한 순위 매김 문제를 가지고 있다. 적합성 재전송(relevance feedback)은 질의 결과에 대한 질을 향상시켜 준다.

라. 구현 메카니즘

디지털 라이브러리의 구현은 분산 컴퓨팅 환경과 유사한 특징을 가지고 있으며, 표준화된 CORBA(Common Object Request Broker Architecture)와 같은 프로토콜을 사용하여 구현하여야 한다.

상호운용성에 대하여 연구 결과로 (3)과 같은 사례가 있다(Interoperability, 1998).

(3) - 미들웨어(MiddleWare) 정보 모델 : INRIA(프랑스), FORTH(그리스)
- 존재구조 : 미시간 대학
- 메타데이타의 상호운용성 : 스텐포드 대학
- 변화에 적용할 수 있는 조정 메카니즘과 소오스 기술 언어 : 미시간 대학
- 설명 서비스, 시각화 : FORTH
- 질의어 형식과 서비스 정의: 스텐포드 대학
- 이질 정보사이 멀티미디어 정보 검색 : CNR(이태리)
- 중개 층에서 질의어로 시스템 최적화, 신뢰성 제공 : CNR
- CORBA 하부구조 제안 : ETH(스위스), 미시간 대학, 스텐포드 대학

3. 2 메타데이타

메타데이타는 데이터에 대한 구조화된 데이터로 웹에서 문헌에 대한 서지 정보나, 도서 목록 기록을 나타내고 있다. 웹의 메타데이타는 서지 정보에 추가적인 정보로써, 창조와 교환을 강조하고 있다. 즉 누구나 쉽게 메타데이타를 생성할 수 있으며, 교환을 위하여 표준 형식을 권장하고 있다.

가치 있는 디지털 문헌은 웹에서 소유권과 경제성 문제로 인하여 접근을 제한하고 있어서 웹 검색기나, 웹 정보 수집소는 문헌 자체에 대한 수집이 계속적으로 어려워지고 있다. 그러므로 문헌 자체보다 원 문헌과 분리되어 복사·교환될 수 있는 메타데이타에 대한 관심이 훨씬 높아지고 있다. 메타데이타를 수집하고, 사용자에게 원하는 정보를 제공하는 형태의 메카니즘은 대규모 분산된 시스템을 구성할 수 있게 하기 때문에, 메타데이타에 대한 표현과 기술에 많은 연구를 하고 있다. 분리 가능한 메타데이타는 원 문헌에 대한 정보를 필요로 하며 그 형태는 (4)와 같이 존재할 수 있다.

(4)- URI(Universal Resource Identifiers) : 일반적인 자원을 지칭하는 이름이나 주소. URL 또는 URN.
- URN(UR Names) : 자원 이름 스킴(Scheme)과 내용으로 구성 (Ron Daniel, 1997).
  예 : urn:isbn:0-262-12186-7, urn:inet:library.bigstate.edu:aj17-mcc
- URL(UR Locators) : URN에 의해 나타나는 자원 위치나 내용.
- URC(UR Characteristics) : IETF(Internet Engineering Task Force)에서 개발중인 스킴으로 자원을 나타내는 속성/값의 집합. URN을 URL과 연결시켜준다. SGML을 사용하여 기술.

위치 정보를 표현하는 형태에서 URC는 저자, 날짜, 위치, 데이터 타입, 등의 정보를 포함하고 있으며, 이 정보는 RDF(Resource

Description Framework)와 유사한 형태를 가진다 (Eric Miller, 1998).

디지털 라이브러리의 중요한 기능인 정보 교환을 위하여 메타데이타는 상호운용성을 가져야 한다. 즉 다양한 형태의 메타데이타는 인터넷에서 일관된 형태의 검색과 교환을 지원하기 위하여 단일 의미를 표현할 수 있는 모델을 가져야 한다. W3C(WWW Consortium)는 워릭(Warwick)구조를 일반화한 의미 표현 구조로 RDF를 제안하였다 (Lorcan Dempsey, 1996). XML을 사용하여 정의하는 RDF는 다양한 형태의 메타데이타의 구조와 의미를 나타냄으로써 시스템 사이에 상호운용성을 보장하여 준다. RDF의 응용은 (5)와 같다.

(5)- 자원 검색 : 좀더 나은 효율성 제공
- 목록화 : 웹에서 유용한 자원의 내용과 관계 기술
- 지식 교환과 공유 : 지적 에이전트 구축으로
- 내용 등급화 : 논리적 문헌의 그룹핑에 의한
- 지적 소유권 : 문헌의 소유자 등록 정보 추가
- 전자 상거래 : 전자 사인 (Signature)을 가진 RDF

대표적인 메타데이타는 (6)에 기술하였다.

(6)- DC : Dublin Core
미국 OCLC와 NCSA가 1995년 더블린에서 합의한 메타데이타.
- TEI(Text Encoding Initiative) Independent Header
  · 미국 National Endowment for the Humanities 지원하에 Association for Computers and the Humanities에 의하여 1987년 11월 TEI 제안.
  · TEI 내용에서 헤더 부분을 독립하여 메타데이타로 지칭함.
- CSDGM : Content Standard for Digital Geospatial Metadata
미국 FGDC(Federal Geographic Data Committee)에서 지형자료를 위한 표준으로 1994년 제정.
- GILS : Government Information Locator Service
미국 연방정부에 의하여 1993년 시작하여 1994년 개발한 시스템으로 정부 자원에 대한 접근, 탐색, 획득을 제공.
- MARC : MAchine-Readable Catalogue
기계 가독형 형태로 서지와 관련 정보의 표현과 통신 표준이며, 미국에서는 USMARC, 그 외에서 UNIMARC, CANMARC, UKMARC, KORMARC 등으로 변화를 가지고 있다.

3. 3 지적 소유권

지적 소유권의 대상이 되는 디지털 정보는 전통적인 종이 정보와 다른 특징으로 인

하여 새로운 시각의 개념을 제공하여야 한다. 즉 디지털 정보는 정보 손실과 추가 비용 없이 누구나 복사 가능한 특징을 가지고 있으며 기존의 법률이 다루지 않고 있어서 잘못 해석의 위험을 초래할 수 있다. 그러므로 디지털 문헌에 대한 전통적 가격, 법률 정책과 다른 새로운 형태의 가격 모델을 제시하여야 한다. 지적 소유권에 대한 문제 해결은 디지털 정보 소유권에 관한 사회 제도, 정보의 경제화에 대한 하부구조, 그리고 정보 내용 및 제공에 관한 연구가 선결되어야 한다. 각각 기능에 대하여 살펴보자.

가. 사회 제도

개인과 기관 사이에 교환되는 모든 메카니즘은 제도와 법률에 영향을 받는다. 국내에서 저작권에 대한 사항은 문화관광부 업무로 한국문예학술저작권협회에서 위탁 관리하고 있다. 저작권 처리를 효과적으로 하기 위하여 저작권 협회는 저작물의 수집 및 등록 기능을 가진 저작권 정보 시스템을 구축하여야 한다.

나. 정보 구조

정보 구조는 저작권이 등록된 정보를 수집, 가공, 유통하는 전 과정에 대한 정보 모델 개발과 더불어 시스템에 대한 보안을 제공하여야 한다.

정보 모델은 전자 상거래를 지원하는 구조와 관련된 것으로, 상품 발견, 서비스, 협상, 교환, 지불을 포함하며, 시스템 보안은 트랜잭션 보안으로 키(Key)를 사용하는 Public Key Encryption과 전자 사인을 사용하는 Digital Certificates가 있으며, 네트워크 보안으로 방화벽(Firewall)을 사용하는 방법과 웹 프로토콜을 사용하는 보안으로 구성된다.

정보 흐름 전체 과정에 필요한 기술은 (7)과 같이 분류된다.

(7)- 정보 상품을 기술하는 언어
- 정보 탐색과 제공을 연결하는 기구와 광고
- 서비스에 대한 평가와 분류 서비스 제공
- 지적 소유권 사용 및 라이센스를 포함하는 계약 언어
- 계약과 라이센스에 관한 설명 기능
- 인증 메카니즘
- Timestamping과 Watermarking 기능
- 다양한 암호화 서비스
- 전자 지불을 포함한 교환 프로토콜

다. 정보 내용

기존의 종이 정보와 다르게 디지털 정보 서비스는 어떤 서비스를 개발하며, 어떻게, 얼마 기간으로 제공하는가에 관한 어려운 결정을 하여야한다. 많은 신생 전자 저널 업체는 출판 경험이 부족하고, 그들이 직면한 전략적 결정을 뒷받침해줄 어떠한 도움도 받을 수 없다. 물론 연구자, 학습자, 그리고 이용자의 행동에 관한 연구가 있으며, 전통적 소비자/생산자 경제 모델을 적용할 수 있으나, 매우 실험적이다. 그것은 디지털 정보가 전통적 정보와 다른 특성을 가지기 때문이다.

예를 들어 캐쉬드 카피(cached copies)(클라이언트 시스템이 정보서비스를 위하여 내부적으로 임시 복사를 수행) 문제가 있다. 캐쉬드 카피는 서버에 연결 시 만 임시 복사를 허용하는 실제적 복사가 아니지만, 서버 연결상태에서 네트워크의 장애가 발생하면 복사의 문제가 생길 수 있다.

또 다른 문제는 경제적 문제와 관련되어 있다. 기존의 자료와 다르게 자료의 분류, 관리, 목록화에 비용이 소모되나, 복사비가 거의 없으므로 가격이 복사량에 비례하는 것은 부적절하다. 또한 저널인 경우 여러 볼륨으로, 각 볼륨은 여러 기사로, 각 기사는 여러 조각(헤더, 초록, 섹션, ...)으로 구성되어 있어서 초록만 묶어서 상품화하는 식의 번들(bundle) 상품이 존재한다.

그러므로 사용 정보를 일반화하여 가격 결정에 사용할 수 있는 연구가 필요하다.

3. 4 분산 정보 검색

분산 정보 검색은 정보 검색을 웹을 통하여 할 수 있게 하는 기능이다. 웹 상에서 정보 검색이 가능하게 하기 위하여 디지털 라이브러리에서 사용하는 다양한 프로토콜에 대한 기능을 지원하는 〈그림 1〉과 같은 게이트웨이 모듈을 필요로 한다. 여기서 정의하는 게이트웨이 모듈은 다양한 표준 프로토콜을 이용하여 디지털 라이브러리에서 메타데이타를 수집하는 기능을 포함한다.

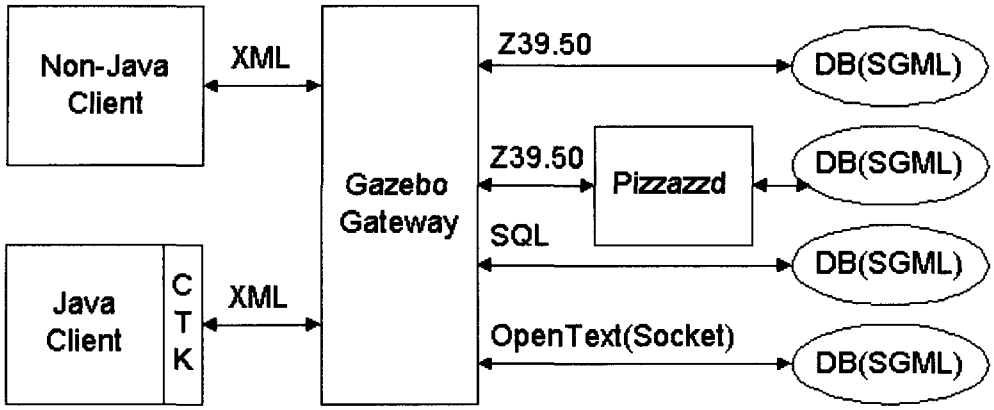
전통적 Yahoo나 Alta Vista와 같은 검색기는 하이퍼링크를 통하여 연결된 여러 웹 사이트(Sites)로부터 정보를 수집하여 기 정의된 키워드 방식에 의한 검색 서비스를 제공하고 있다. 그러나 이러한 방식은 비 공개된 가치 있는 정보를 제공할 수 없으며, 정보에 대한 구조적 검색이나 자세한 정보를 얻기가 어렵다.

메타데이타는 저작권 문제가 없는(일반적으로) 공개 자료로 누구나 추출하여 사용할 수 있으며, 표준 기술 방식으로 정의되어 있어서 정보의 정확한 추출 및 제공이 가능한 정보이다. 이와 같은 정보를 얻기 위하여 분산 정보검색은 다양한 형태의 메타데이타 형식을 지원하여야 하며, 여러 사이트와 통신을 위한 표준 프로토콜을 이해하여야 한다.

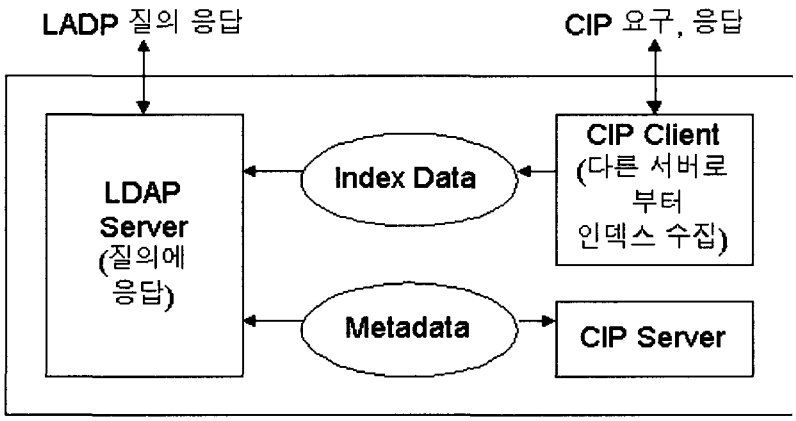
〈그림 3〉의 Emerge 시스템 모델은 DLI 프로젝트의 일환으로 Urbana-Champaign에 있는 일리노이 대학에서 개발한 시스템이다. 이 시스템은 하나의 디지털 라이브러리에서 다수의 인터넷 자원을 검색할 수 있는 기능을 가지며, 하부구조인 과학문헌 연합저장소(Federating Repositories of Scientific Literature)를 구축하였다(Bruce, 1996).

Emerge는 Z39.50을 지원하는 Pizzazzd와 다양한 프로토콜을 지원하는 Gazebo를 통하여 웹에 분산된 정보를 수집하여 인덱스(Index)한다.

또 다른 방법으로 Isaac모델을 살펴보자(Michael Roszkowski, 1998). 초기 Internet Scout Project의 Signpost는 2,000개 이상의 분류된 인터넷 사이트를 가지고 있으나 인터넷 상에서 유용한 질 좋은 정보를 얻을 수 없었다. 이와 같은 문제를 극복하기 위하여 나타난 Isaac 프로젝트는 검색 인터페이스로 웹 기반 HTTP와 HTML을, 질의 처리 및 인덱



〈그림 3〉 일리노이 대학의 분산 정보검색(Emerge) 시스템 구조



〈그림 4〉 Isaac 모델 구조

스 정보에 대한 질의로 LDAP(Lightweight Directory Access Protocol)를, 그리고 서버 사이 인덱스의 생성과 교환을 위하여 CIP (Common Indexing Protocol)를 사용한다. Isaac에서 사용하는 메타데이타는 수동으로 만들어지거나 분류된 자원을 대상으로 한다. 그것은 수동으로 분류된 자료가 자동으로 추출하는 키워드보다 좀더 적절하기 때문이다. Isaac의 전체적인 목표는 분산되고 독립적인 메타데이타 콜렉션에 대한 단일 인터페이스를 제공하는 것이다.

Isaac은 메타데이타로 15개의 DC 필드를 사용하나, DC의 각 필드와 유사한 필드를 가진 메타데이타도 허용한다. Isaac 모델은 자신의 메타데이타 정보와, 검색하여 축적한 인덱스 데이터를 가지고 있으며, LADP를 통하여 사용자 질의에 대하여 응답한다. 전체적 구조가 〈그림 4〉에 기술되었다.

Emerge나 Isaac에서 사용하는 HTTP/Z39.50, XML 방식의 ICE, LDAP, CIP이외에 Dienst, WHOIS++ 등의 다양한 인터넷 프로토콜이 있다.

3. 5 다국어 처리

가. 진행 과정

초기 언어 자원에 대한 기반 기술은 미국을 중심으로 단일어(Monolingual) 정보검색에 중점을 두고 연구되어 왔다. 그러나 다국적 문화를 가진 유럽에서는 다국어 정보 검색이 중요한 이슈로 떠오르면서 본격적으로 연구되었고, 최근 들어 인터넷의 역할이 증대됨에 따라 미국과 유럽이 공동으로 연구하고 있는 분야이다.

AltaVista와 같은 웹 검색 엔진은 수백 단어 수준의 번역 검색을 제공하고 있으며 TITAN과 MUNDIAL은 실험적인 교차언어 정보검색(Cross-Language Information Retrieval: CLIR) 기능을 제공하고 있다. 일본에서는 영어-일어 번역 소프트웨어를 사용한 웹 브라우징이 대중화되고 있다. 1991년 ESPRIT 프로젝트의 일환으로, EMIR (European Multilingual Information Retrieval)은 영어, 프랑스어, 독일어 사이 검색을 수행하는 시스템을 개발하였으며 이는 프랑스 SPIRIT로 확장되었다. EMIR은 1994년 종료되었지만 SPIRIT는 계속 진행 중에 있다. European Commission Tele-matics Application Program(DG XIII/E)의 언어 공학 분과에서 유럽은 CRISTAL을 가지고 계속적인 CLIR 연구를 진행 중에 있다. CRISTAL은 1993년과 1996년 사이 영어, 프랑스어, 이태리어에 대한 연구에 집중하였으며 계속적으로 EuroWordNet(Pie Vossen, 1998)에 연구를 집중하고 있다.

대부분의 유럽 프로젝트는 유럽 문화의 특징상 CLIR의 응용에 집중하고 있으며 2개의 Telematics 응용인 TRANSLIB와 CANAL/LS 프로젝트가 1995년과 1997년 사이에 있었다. 이들 프로젝트는 영어, 스페인어를 기반으로 CANAL/LS는 독어, 프랑스어를, TRANSLIB는 그리스어를 추가하여 교차 언어 도서 목록을 개발하였다.

Language Engineering Initiative인 MULINEX 프로젝트는 웹 검색 엔진 기술에 대한 연구를 하고 있다. 다른 진행중인 프로젝트로 TwentyOne이 다국어 문서 검색 시스템을 개발 중에 있다. 유럽 연합의 회원은 아니지만 스위스는 Swiss Federal Institute of Technology에서 다국어 정보검색을 지원하고 있다. 프랑스에 있는 Rank Xerox Research Laboratory는 유럽과 공동으로 다국어 정보검색을 진행 중에 있다.

미국의 연구도 유사한 패턴을 따르고 있다. Bellcore는 1990년 영어와 프랑스어를 대상으로 LSI(Latent Semantic Indexing) 기술을 사용하여 실험적 연구를 보고하였다. Bellcore 그룹의 회원은 계속적으로 듀크 대학과 콜로라도 대학에서 연구를 진행 중에 있다. 뉴멕시코 주립 대학은 영어와 스페인어를 대상으로 1993년부터 연구를 하고 있으며, 1994년 메릴랜드 대학이 시작하였고, 카아네기멜론 대학, 아리조나 대학 UC 버클리, 아이오아 대학, 매사추세츠 대학이 현재 연

구 중에 있다. 최근 들어 DARPA의 정보기술 사무국은 이 분야에 중점적 지원을 시작하였으며, Translingual Information Management에 대한 3개의 프로젝트를 동시에 지원하고 있다.

아시아에서는 일본의 KDD, NEC, NTT에서 실험적 연구가 있었으며, NACSIS (National Center for Science Information Systems)에서 관심을 가지고 투자하기 시작하였다. Australia의 Royal Melbourne Institute of Technology는 영어와 베트남어 사이 다국어 정보검색에 연구를 하기 시작하였다.

다국어 정보검색은 세계적 문제로 간주되어 미국과 유럽연합은 디지털 라이브러리 Working group 중에서 다국어 정보접근이라는 협력 연구 그룹을 결성하였다(Alan F. Smeaton, 1998). 이 그룹은 1998년 여름 White Paper를 제출하였으며, 최종 보고서가 10월에 보고되었다. 이는 미국 DLI-2 프로그램 모습으로 나타나는데 큰 역할을 하였다. 미국의 DARPA는 CLIR의 군사적 요구사항 분석과 Translingual Information Management에 대한 추가적 투자를 계획하고 있다.

최근까지 다국어 테스트 문헌의 제한된 유용성은 CLIR 시스템 사이의 성능 테스트를 어렵게 하였다. 1997년부터 미국 NIST(National Institute of Standards and Technology)와 DARPA의 TIPSTER 프로그램에 의하여 지원하는 TREC(Text REtrieval Conference) 회의 시리즈 중에서 TREC-6(TREC 프로젝트는 결과에 대한 공헌자만 참가할 수 있는 closed conference로 TREC-6는 TREC의 6번째 회의)은 25개의 일반적 주제에 대한 다국어로 된 문헌을 마련하였으며, CLIR에 대한 12개의 실험적 그룹의 결과를 제시하였다. 1998년 11월초에 열린 TREC-7에서 CLIR 시스템에 대한 평가가 있었다.

나. 분류

다국어 정보 검색은 질의 언어와 상관없이 다양한 언어로 만들어진 정보를 검색하는 시스템이다. 다국어 정보 검색은 '다국어', '언어간'(Translingual), '교차언어' 등의 용어로 사용되고 있으며, 이들 사이에 약간의 의미상 차이를 가지고 있지만 전부 같은 분야에 대한 연구로 정의되고 있다.

다국어 문헌을 검색하기 위한 방법은 문헌 번역 방식과 질의어 번역 방식으로 나눌 수 있다. 문헌 번역은 다국어 문헌을 질의 언어로 미리 번역한 뒤에 처리하는 방식이나 현재 기계 번역의 낮은 품질로 많이 이용하지 않고 있다. TREC-6에서 문헌 번역 방식에 의한 정보 검색 작업을 시도하였다. 결과로 질의어 문장이 긴 형태와 제한된 영역에서 좋은 효과를 발휘하였다(장명길, 1998)

질의어 번역 방식은 사전 기반, 시소러스 기반, 코퍼스 기반 방식 등으로 다시 분류된다. 질의어 번역 방식은 질의 언어를 검색 대상 문헌의 언어로 번역한 뒤 검색하는 방식을 사용하는 것으로, 인터넷의 분산된 시스템 환경 하에서 적절하다. 질의어 번역 방식은 질의어를 대상 문헌 언어로 번역할 때 발생하는 모호성 문제(예를 들어, '눈'이 영어의 'eye'인지, 'snow'인지 구별)에 대한 해

결이 중요한 역할을 한다.

사전 기반 방법은 대역 사전(bilingual dictionary)을 사용하는데 사전 획득이 쉽고 번역 방식이 단순하나 모호성 문제로 인하여 검색 효과는 다른 방법에 비하여 매우 낮다.

시소러스 기반 방법은 모호성 문제를 시소러스가 가지는 유사 개념의 어휘 확장을 통하여 질의를 하고 있으나 시소러스 또는 다국어 시소러스 구축 자체가 쉬운 문제가 아니다. 현재 프린스턴 대학에서 구축한 WordNet(George A. Miller, 1990)과 유럽 8개국어로 구축한 EuroWordNet(Pie Vossen, 1998)이 시소러스를 통한 방식에 대한 해결책을 제시하고 있다.

코퍼스 기반 방법은 시소러스 기반 방법에 비하여 구축하기가 쉽다. 코퍼스는 일정한 원칙 하에 모아 놓은 텍스트를 지칭하는 것으로, 병렬 코퍼스(Parallel Corpus) (두 언어 사이에 line-by-line 형태의 번역을 통하여 구축)를 사용하거나, 비교 코퍼스(Comparable Corpus) (두 언어 사이에 동일한 주제를 가지고 자유롭게 구축)를 사용하여 모호성 문제를 해결하고 있다. 이들 두 가지 형태의 코퍼스는 한 언어에 대하여 번역의 사례를 보여주고 있기 때문에, 모호성 문제를 기 번역된 예제(코퍼스)를 통하여 해결하는 방식을 가진다. TREC-6 자료에 일본어 15개 질의어로 평가한 GDMAX (Ge-neralized Double MAXimize criteria based on Comparable Corpus) 방식은 일-영 사전에 기반한 시스템에 비하여 12%, 기계 번역 질의보다 6%의 효과를 발휘하였다(Akitoshi Okunura, 1998).

다. 응용

다국어 정보처리는 다양한 분야에 응용될 수 있으며 특히 다음 (8)과 같은 분야에 유용하게 사용할 수 있다.

(8)- 정보 검색 : 질의 처리, 검색, 메타 데이타 만들기, 인덱싱
- 기계번역 : 비교와 병렬 텍스트 조정, 언어간 기술
- 계산 언어학 : 형태소 분석, 파싱, 의미 분석, 언어 생성
- 문헌 처리 : 문헌의 분류, 필터링, 분리,
- HCI(Human-Computer Interface) : 문헌의 시각화, 다수 문헌 요약

3. 6 문헌 표현

가. 진행과정

문헌의 구조화는 1960년대 후반, 미국의 GCA(Graphic Communications Association)의 GenCode Committee에서 개발한 generic coding에 기반을 두고 있으며, IBM에서 추진한 Generalized Markup Language (GML) project로 구체화되었다. 미국 표준협회인 ANSI에서 이를 바탕으로 SGML 초안을 작성하여 1986년 ISO 8879 Information Processing -Text and Office System-Standard Generalized Markup Language로 완성하였다. ISO/IEC JTC1/WG4에서 담당하고 있으며 SGML에서 문헌 처리 구조와 형태, Standard Page Description Language, Font 구조, 교환 형식(DSSSL),

Hypermedia 문헌에 관한 표준을 연구하고 있다.

문헌에 대한 마크업(Markup)은 전통적으로 단어나 문단에 대한 지시 사항으로 표기하는 주석 달기에서 시작하였으며, 문헌에 대한 구조 기능을 표기하는 형태로 발전하였다. SGML은 전자화 된 문헌의 마크업을 기술하는 메타언어로써, 시스템 독립적인 기능을 가지며 오늘날 대부분의 문헌 기술에서 채택하고 있는 형식이다.

특히 미 국방성의 CALS(Commerce At Light Speed) 프로젝트에서 표준으로 사용하고 있으며, 미국 출판 협회, 유럽 공동체, 일본 등에서 SGML을 문헌 기술의 표준으로 사용하고 있다.

최근 들어 복잡한 SGML 기능을 간략화한 XML(eXtensible Markup Langauge)이 문헌 교환 표준으로 등장하였다. W3C에서 제안한 XML은 기존의 SGML이 가지는 단점을 보완한 간단한 형태를 하고 있으며, 차세대 문헌 기술 표준으로 Netscape나 Microsoft 회사에서 차기 웹 브라우저로 지원을 하고 있다. 대부분의 기존 SGML 업체들도 XML 관련 S/W를 개발·제공하고 있다.

이에 따라 ISO에서는 SGML이 XML의 superset을 진정한 의미로 가능하게 하는 기능으로 'WebSGML Adaptation Annexes'인 Annex K와 L을 발표하였다(Bob DuCharme, 1998). 또한 Hytime 기능 확장, DSSSL 표준 확정, SGML의 Unicode 지원을 포함하는 계속적인 SGML 개정을 추진하고 있으며 일부 업체에서 개정된 기능을 반영하는 S/W를 출시하고 있다.

XML은 SGML의 subset으로 W3C에서 추진하는 문헌 교환 표준으로 XML 자체에 대한 연구, XLL(XML Linking Language)로 HyTime (ISO/IEC 10744) (Llord Rutledge, 1996)과 Text Encoding Initiative (TEI) 기반 설계인 Xpointer와 Xlink 기능 제공, 그리고 XSL(XML Style Language)인 DSSSL (ISO/IEC 10179)을 이용, 표준 stylesheet 설계 활동에 관한 표준을 마련하고 있다.

어휘 정보를 기술하는 TEI/CES와 더불어 최근 들어 SGML과 XML 문헌 기술에 새로운 기반 기술로 사용자 중심 시각을 제공하는 TNM(Topic Navigation Maps)과 문헌 기술의 DOM(Document Object Model)에 대하여 살펴보자.

나. TEI/CES

TEI는 전자 문헌 작성과 교환을 위한 가이드라인으로 1987년 뉴욕 Poughkeepsie에서 개최된 회의에서 제시한 몇 가지 원칙에서 출발하여 1994년 5월 3번째 안(proposal)인 TEI P3이 제안되었다.

TEI는 기존의 전자 문헌 작성 시 서로 상이한 인코딩(encoding) 방식에 따른 낭비를 제거하고 다양한 유형의 문헌에 적절한 코딩 집합을 기술하기 위한 전자문헌기술 방식이다. 현재 제공되는 문헌의 유형은 산문(prose), 운문(verse), 드라마(drama), 스피치(speech), 사전(dictionaries), 용어사전(terminology database), 텍스트 분석(text analysis), 텍스트 비평(text criticism) 등이며, 약 400여 개 이상의 원소(elements)를 사용하

여 다양한 유형의 문헌을 인코딩 할 수 있도록 개발되었다.

CES(Corpus Encoding Standard)는 TEI 기능에서 다국어 코퍼스, 사전에 대한 기능을 강화할 목적으로 1996년 만든 SGML DTD이다. 원 문헌에 증가적으로 마크업 하면서 생기는 복잡성과 비 계층적 문제에 대한 해결책을 제시하고 있으며, 문헌 번역시 원 문헌과 비교할 수 있는 배치(alignment)에 관한 마크업을 가지고 있다. 현재 XML을 사용한 DTD(Data Type Definition)를 개발하고 있다.

다. TNM

인덱스, 상호 참조, 어휘(glossary), 카타로그 같은 전통적 인쇄물을 위한 네비게이션 도구는 온라인 정보에 적절하지 못하다. TNM은 국제 표준(ISO/IEC 13250)으로 향상된 온라인 정보 검색을 위하여 만들어진 정보 구조로써 사용자가 자신의 네비게이션 전략을 정의하게 하며, 살아있는 문헌 저장소를 유지하게 하는 기능을 제공한다(Michel Biezunski, 1997).

구조화된 문헌에서 사용자는 지정된 구조 이외에 아무 것도 할 수 없으며, 비 구조화된 문헌에서는 자동 처리에 의하여 얻어지는 결과를 변경시킬 아무 권한도 없다. TM(Topic Maps)은 정보 저장소의 밖에서 사용자가 구조화하는 정보 모델을 제공한다. 그러므로 같은 정보 집합에 많은 TM이 가능하며, SGML은 TNM에서 하나의 소오스에 다수의 결과를 제공하는 도구로 사용된다.

TM은 SGML의 응용으로 적용되는 정보의 의미적 구조화를 기술한다. 또한 주소화(addressing)와 연결 기능(linking)을 사용하기 때문에 HyTime의 응용으로 구조적 형태의 집합으로 기술된다.

XML은 SGML의 단순화된 형태이므로 TM 표현에도 적절하다. TM은 XML 처럼 특별한 DTD를 요구하지 않는다. XML의 연결 부분인 XLL은 주소화 언어(Xpointer)와 연결로 구성된다. 연결은 단순 연결과 확장 연결로 되어있으며, 확장 연결은 HyTime에서와 같이 독립적 연결을 가진다. 이들은 TM 정보를 충분히 묘사할 수 있으며, XLL의 확장 연결은 TM 정보 교환을 위하여 적합하다. TNM은 1996년 CApH에서 제안되었으며 1999년 2월에 최종버전이 발표될 것이다.

TM 구조는 토픽(Topic), 토픽 관계, 필터로 (9)와 같이 구성된다.

(9)- 토픽 : 토픽은 주어진 주제에 관한 정보를 지적하는 연결이다. 앵커의 집합은 토픽의 제목이다. 토픽은 여러 개의 제목을 가질 수 있으며 다른 언어로 표현될 수 도 있다. 상호 참조는 두 앵커 사이에 연결로 제목이 없는 토픽이다. 상호 참조에 제목을 추가하여 하나의 토픽으로 향상시키는 것은 복잡한 문헌 저장소를 단순하게 유지할 수 있게 한다.

- 토픽 관계 : 토픽 관계는 토픽들의 관계로 지식 표현의 하나로 볼 수 있다. 관계 DB에서 묘사하는 정보

와 유사하다.

- 필터 : 필터는 정보를 포함하거나 배제시키는데 사용하는 연결의 제3의 범주이다. 토픽은 토픽 타입, 관계, 그리고 개인적 앵커에 적용된다. 필터의 사용은 언어, 의미, 사용자 인적사항, 보안 레벨, 정보의 유용성을 포함한다. 필터는 사용자 정의이며 표준으로 고정되지 않는다. 동시에 여러 개의 필터를 적용할 수 있으며, 네비게이션 결과는 이들이 적용 가능한 정보 네트워크의 교집합(Intersection)으로 보여질 수 있다.

High Text 회사의 EnLIGHTeN는 Topic Map Organizer로써 개발되었다.

라. DOM

최근들어 HTML이나 XML을 사용하는 문헌 사이에 웹 상호운용성 제공을 위하여 W3C에서 DOM(Document Object Model) 레벨-1 사양(Level-1 Specification)을 정의하였다. DOM은 XML이나 HTML로 쓰여진 문헌을 만들고, 네비게이션하며, 추가, 삭제, 변경을 가능하게 하는 표준 사양이다. 즉 HTML과 XML을 위한 API(Application Programming Interface)를 정의한 것이다. 예를 들어 문헌에서 특정한 원소(Element)를 가지는 부분을 제거하거나 특정한 속성을 가지는 원소를 추출할 수 있는 기능이 있다.

DOM은 문헌이 가지는 구조적 모델을 기술한 것으로 Java나 ECMAScript를 사용하여 구현하거나, COM(Component Object Model)이나 CORBA를 사용하여 구현할 수 있는 언어 독립적 사양이다(Elaine Brennan, 1998).

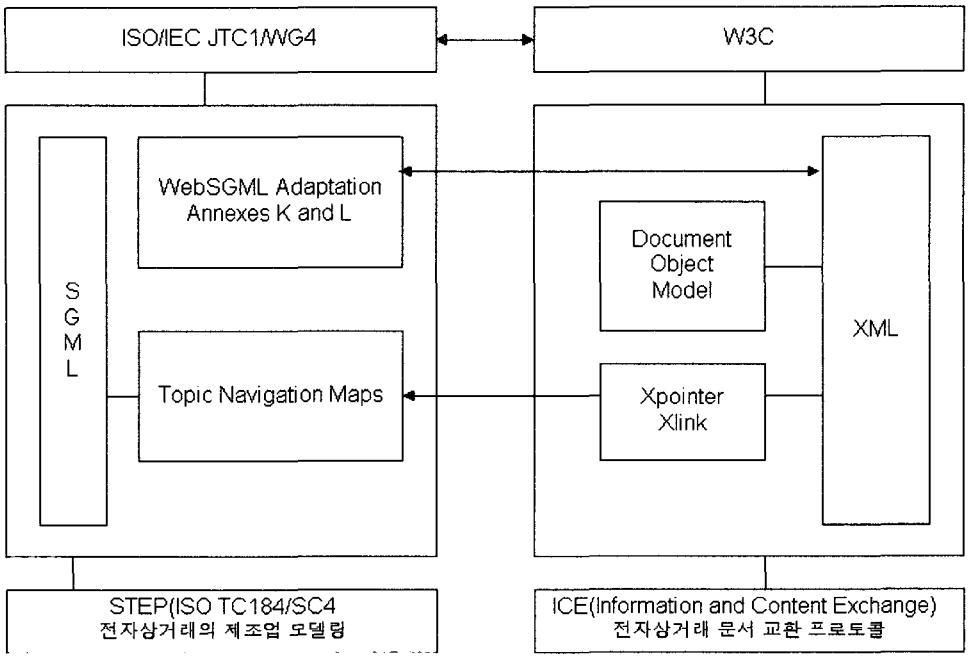
마. SGML과 XML 동향 비교

문헌 기술을 위한 SGML과 XML의 발전 동향과 서로의 관계를 살펴보면 〈그림 5〉와 같다. SGML은 ISO에서, XML은 W3C에서 제안된 사양이며 SGML은 'WebSGML Adaption'을 통하여 XML을 수용하는 방향으로 나아가고 있다. 이들은 Hypermedia 기능을 수용/응용하기 위한 기술로 SGML은 TNM을 XML은 HyTime과 TEI 기능으로부터 Xpointer와 Xlink 기능을 제공하고 있다. 전자상거래에서는 XML 구문을 사용한 ICE를 문서 교환 프로토콜로 제안하였으며, 제조업 생산 과정에 대한 모델링 언어인 STEP/EXPRESS는 SGML과 기능적 유사성으로 인하여 협력을 추진하고 있다(김철환, 1996).

4. 디지털 라이브러리의 사례

4. 1 웹 가상 도서관

웹 가상도서관(WWW Virtual Library)은 전문 분야에 관한 학술적 문헌, 자원, 정보자료에 대한 서지적, 하이퍼텍스트 기능을 제공하는 세계적 협력 프로젝트로 1991년 시작되었다. 웹의 창시자 Tim Berners-Lee에 의해 시작된 가장 오래된 목록으로 상업적



〈그림 5〉 ISO의 SGML과 W3C의 XML의 연관도

목록과 다르게 자원자 연합에 의한 느슨한 연결로 구성되어있다.

웹 가상도서관의 목록 분류에서 지역 연구 항목에 있는 Asian Studies WWW VL은 아시아 지역의 학술 정보 자료에 관련된 서지적, 하이퍼텍스트 기능을 제공한다. 1994년 3월 24일 시작되었으며 개인 및 단체 사이트가 일정 원칙 하에 가입·연결된 지식 그룹으로써 역할을 하고 있다. 현재 40개 프로젝트가 연계되어 있으나, 한국의 어떤 기관도 가입하지 않았다. 6개의 mirror sites를 운영하고 있으며 78개의 정보 모듈, 상위에 3,127개의 외부 연결 기능이 있다. 운영은 호주 국립대학이 운영하고 있으며, 제공 정보는 다른 정보와 중복되지 않는 원칙을 가지고 있다. 아시아-태평양 자원, 지역적 자원, 국가별 자원으로 분류 제공하고 있다.

한국에 관한 정보 제공은 현재 미국 듀크대학의 Moo-Young Han 교수가 한국 홈페이지를 운영하고 있다. 그는 IEKAS(The Information Exchange for Korean - American Scholars)의 운영자이기도 하다. IEKAS는 한국에 관한 각종 정보와 의견을 전자 뉴스 형식으로 매주 회원에게 전자메일(E-mail)로 제공하고 있다.

4. 2 OCLC

OCLC(Online Computer Library Center)는 오하이호주 소재 54개 대학 협의체의 제안에 의해 1967년 도서관 정보의 소유를 위한 비영리 단체로 출발하였으나 1980

년 이후부터는 상용 데이터베이스를 저렴한 가격으로 도서관과 일반이용자에게 제공하고 있다. 현재 65개국 30,000명 이상이 OCLC 서비스를 사용하고 있다. OCLC는 FirstSearch, EPIC, OCLC EJO등의 64개 이상 전문학술 DB를 제공하고 있으며, 이들 자료에 대한 전문(Full-text), 팩스, 우편, 온라인 목록(Online Catalog)과 Resource Sharing서비스를 제공하고 있다.

4. 3 ELRA

ELRA(European Language Resource Association)는 1995년에 설립된 비영리 단체로 유럽 사회의 언어, 텍스트, 전문용어에 관한 자원과 도구에 대한 생산, 분배, 운영을 목적으로 설립되었다. ELRA는 유럽공동체가 지원하는 언어자원(Language Resource: LR)에 대한 저장소 역할을 한다. LR은 스피치 자료, 어휘 사전, 문법, 텍스트 코퍼스, 전문용어를 포함하는 자료이며, 정보 기술의 넓은 응용인 스피치와 텍스트 처리 시스템 개발을 위한 다양한 소프트웨어 도구를 보유하고 있다.

ELRA는 개발된 모든 자원에 대하여 유효화 검증을 실시한다. 유효화 검증은 연구목적을 제외한 모든 언어자원에 대하여 시장에 적합성 검증, 표준화 준수, 그리고 품질 검사과정을 포함하고 있다. 유효화 전문가 그룹이 유효성 검증 단위를 정의하였고, 스피치, 텍스트, 전문용어 분야로 각각 나누어서 검증을 한다. 자원에 대한 개발 및 분배는 ELDA(European LR Distribution Agency)를 통하여 이루어지며, ELDA는 ELRA의 언어자원에 대한 개발 분배에 관한 자문 역할을 하며 법적인 정보를 제공한다.

4. 4 일리노이 대학

〈그림 3〉은 일리노이 대학에서 구축한 과학문헌 연합 분산 저장소의 구조를 나타내고 있다. 연합 분산 저장소의 문서는 SGML을 사용하여 기술하며, 빈도에 기반을 두고 개념공간의 규모 변경(scalable) 기술을 사용한 의미 추출에 중점을 두었다. 이용자에게 다수의 저장소에 대한 단일 인터페이스를 제공하고 있다.

시스템에서 문헌 처리 과정은 (10)과 같다.

(10)- 다양한 프로토콜을 사용하는 다수의 저장소로부터 이질적 SGML로 기술된 문헌을 수집
- 이들 자료를 휴리스틱(Heuristics)을 사용하여 canonical DTD로 변경 및 태깅
- 인덱스 엔진을 사용하여 인덱스
- 질의어 확장을 위한 INSPEC(OCLC의 FirstSearch로부터 서비스)의 subject thesauri(10,000 terms) 및
- 개념 공간 기술인 INSPEC의 공기(co-occurrence) 리스트(200,000 terms) 사용
- 단일 인터페이스를 사용하여 다수 연합 저장소 문헌을 처리

일리노이 대학은 다수 저장소에 분산된 인덱스 된 자원에 대한 검색을 제공함으로써 연합 정보 저장소 구조를 갖추고 있다. 〈그림 3〉처럼 구축한 시스템 Gazebo는 다양한 통신 프로토콜을 지원하며, Pizzazzd는 Z39.50을 지원하는 C library 기능을, XML은 통신 프로토콜을 표현 구문으로 사용된다. Gazebo, Pizzazzd는 공개된 소프트웨어로 자유롭게 제공되고 있다.

4. 5 사례 비교

〈표 1〉은 상기에 기술한 디지털 라이브러리에 대한 비교 내용을 기술하고 있다. 이들 정보로부터, 디지털 라이브러리가 문헌 정보만을 가진 형태도, 메타데이타만 가진 형태도, 프로토콜만 가진 형태도 아닌 다양한 레벨의 자원이 기관의 목적에 맞게 구성된 구조를 가지는 것을 알 수 있다.

5. 우리 나라의 디지털 라이브러리

가. 현황

디지털 라이브러리는 단순한 정보 운영 도구로써 디지털화 된 장서만을 의미하는 것은 아니며 자료, 정보, 지식에 대한 생성, 분배, 사용, 보존의 전 사이클을 지원하는 장서, 서비스, 그리고 사람을 포함하는 환경이다. 그러므로 많은 기관이 디지털 라이브러리를 제공하고 있지만 광범위한 환경적 의미를 가지는 기관은 드물다.

예를 들어 국립중앙도서관, 국회도서관, 한국과학기술원 과학도서관, 첨단학술정보센터, 법원도서관, 연구개발정보센터(KORDIC), 한국산업기술정보원이 참여하는 국가주요 전자도서관 구축 사업의 내용은 주로 소장 자료에 대한 DB 구축에 중점을 두고 있다(국립중앙도서관).

디지털 라이브러리의 대표적인 기관으로 첨단학술정보센터(KRIC)가 있다. 첨단학술

〈표 1〉 디지털 라이브러리의 사례 비교

	웹가상도서관	OCLC	ELRA	일리노이 대학
시작	가상도서관	도서관 공동 이용	다국어 언어자원구축	분산 자원 검색
분야	다양	다양	다국어 언어자원	과학분야
정보 구축 방법	하이퍼링크	상용 + 자체	자체	하이퍼링크, 자체 인덱스구축
표준화 관련	관계없음	메타데이타	메타데이타, 문헌, 스피치, TEI, CES	Internet Protocol (Z39.50, XML), 문헌(SGML)
정보내용 및 용도	일반 정보에 대한 하이퍼링크	다양한 수준으로 정보 검색자료	전문정보, 대규모 코퍼스 자료로 테스트베드, 언어 이용 도구	분산정보저장소를 연결하여 단일 서비스 제공
목적	지역적정보 연결	상호대차, 정보제공	다국어 언어자원활용	분산 자원 활용

정보센터는 OCLC와 유사한 형태를 가진 기관이지만, 지금까지 기술한 디지털 라이브러리의 기능으로 볼 때 매우 초보적 형태의 시스템을 제공하고 있다. 제공하는 서비스를 살펴보면 (11)과 같다.

(11)- 제공 기능
  · OPAC과 공동 목록을 통한 도서 종합목록
  · 해외 DB
  · 일부 학회지 원문
- 제공 예정 기능
  · 학술지 논문 종합목록
  · 상호 대차
- 사용하는 디지털 라이브러리 기술
  · 분산 처리 : Z39.50 사용
  · 목록 : 메타데이타(DC와 MARC로 표현)
  · 문헌 표현 : SGML로 구축

나. 방향

디지털 라이브러리의 목적은 단기적으로는 학술적 연구와 교육을 목표로 하지만, 장기적으로 광범위한 연구, 학습, 상업 활동에 적용할 것이다. 그러므로 현재 제공하는 학술지 정보와 목록을 바탕으로 새로운 상황에 적응할 수 있는 다방면적인 기능을 제공하여야 한다.

학술적 연구에 필요한 테스트베드, 연구 현황 정보를 제공하는 클리어링 하우스, 일반인을 상대로 하는 가상 교양 대학, 인간에게 적절하게 조정된 대화적 시스템, 인터넷 무역을 가능하게 하는 전자상거래 등의 응용에 적용할 수 있는 장기적 차원의 디지털 라이브러리를 구축하여야 한다.

그러나 현재 국내에서 서비스되고 있는 시스템은 대부분 DB 위주의 정보를 구축·제공하고 있기 때문에 메타데이타 구축, 분산 정보검색, 다국어 처리, 저작권 관리, 문헌 표현에 대한 세계적 기술 동향을 바탕으로 우리 시스템에 적절한 표준 기술 연구와 구축을 강화하여야 한다.

연구 및 구축하여야 할 적절한 기술적 사항을 제시하면 (12)와 같다.

(12)- 메타데이타 구축 : 다양한 문헌을 지원하는 메타데이타 정의(DC, TEI Independent Header). DTD 개발.
- 분산 정보검색 : 공개된 Z39.50, ICE 프로토콜 제공. 메타데이타 검색 기술 개발. 분산 인덱싱 기술 개발.
- 다국어 처리 : 표준 테스트베드 구축. 다양한 분야의 비교 및 병렬 코퍼스 구축. 대역 사전 구축. 다국어 시소러스(예를 들어 WordNet의 한국어 버전) 구축.
- 저작권 관리 : 국가적인 저작권 관리 시스템 개발.
- 문헌 표현 : SGML/XML 관련 S/W 개발. TNM, DOM에 대한 기술적 연구. 다양한 문헌을 위한 DTD sets과 일반인도 사용 가능한 Templates 개발.

(12)의 기능은 적은 시장(Market)을 가진 우리 나라와 같은 현실을 고려할 때 공개 사양으로 구축·제공되어야 하며 이러한 모델을 바탕으로 각 기관과 업체는 자신에 맞는 디지털 라이브러리 응용 시스템을 개발하여야 할 것이다.

6. 결론

디지털 라이브러리는 구축 기관의 특징에 따라 다양한 형태로 정의될 수 있으며, 문제 해결에 관한 기술적 방향도 다양하다. 디지털 라이브러리는 분산 환경과 정보 공유, 그리고 전자상거래의 발전으로 기존의 전산학, 문헌정보학을 넘어서 법률, 사회, 경제, 문화적 이슈가 계속적으로 도출되고 있다.

본 논문에서는 인터넷 사용자의 다양한 요구에 적합한 디지털 라이브러리의 구조를 제시하였으며 구성 기술에 관한 동향을 살펴보았다. 시스템 사이에 일관된 시각을 제공하는 상호운용성, 문헌에 대한 정보의 수집과 교환을 위한 메타데이타, 정보의 저작권과 경제적 측면에 관한 지적 소유권, 인터넷에 분산된 정보의 수집과 검색을 위한 분산 정보 검색, 다국적 문헌에 대한 처리를 가능하게 하는 다국어 처리, 그리고 디지털 라이브러리의 가장 중요한 사항으로 자체 문헌 기술에 관하여 살펴보았다.

디지털 라이브러리는 정보의 하부구조로, 미국, 유럽 연합, 일본 등에서 정보 기술의 선점을 위하여 국가적으로 투자하고 있는 분야이다. 국내에서도 첨단학술정보센터, 연구개발정보센터 등에서 연구를 시작하고 있으나 아직 세계적 수준에 비하여 미비한 실정이다. 장기적으로 국가적 재정 지원 하에 산·학·연이 합동으로 우리 현실에 맞는 디지털 라이브러리 모델을 설정하여 추진하여야 할 것이다.

참 고 문 헌

국립중앙도서관, 홈페이지, 〈http://www.nl.or.kr〉

김철환, 김규수, 신영인, 1996, "한국적 CALS 표준화 구축방안", Journal of the Korean Institute of CALS/EC, Vol 1., No. 1.

장명길, 김영길, 박영찬, 1998, "다국어 정보검색", 정보과학회지 제16권 8호, p21-31

Akitoshi Okunura, Kai Ishikawa, Kenji Satoh, 1988, "Translingual Information Retrieval by a Bilingual Dictionary and Comparable Corpus", LREC First Internaltional Conference on Language Resources and Evaluation, Workshop on Multilingual Information Management, p26-30

Alan F. Smeaton, 1998, Summary Report of the Series of Joint NSF-EU Working Group on Future Directions for Digital Library Research, URL 〈http://www.iei.pi.cni.it/DELOS/NSF/Brussrep.htm〉

Asian Studies WWW VL, Australian National University, URL 〈http://coombs.anu.edu.au/WWWVL-AsianStudies.html〉

Bob DuCharme, 1998, SGML Revisions and XML, 〈TAG〉 Volume11, Number 9.

Bruce Schatz, et al, 1996, "Federating Diverse Collections of Scientific Literature", 50 years of service IEEE Computer Society.

CES 1996, Corpus Encoding Standard URL〈http://www.cs.vassar.edu/CES/〉

CORBA, Common Object Request Broker Architecture, URL 〈http://blue.wonkwang.ac.kr/〉

CSDGM, Content Standard for Digital Geospatial Metadata, FGDC, URL 〈http://www.fgdc.gov/metadata/constan.html〉

DLI, Digital Library Initiative, URL〈http://dli.grainger.uiuc.edu/national.htm〉

DELOS, ERCIM Digital Library Working Group, URL 〈http://ntserv.iei.cnr.it/DELOS/ REPORTS/annual/9798.htm〉

DOM Working Group, 1998, Document Object Model(DOM) Level 1 Specification, W3C, URL〈http://www.w3c.org/DOM/〉

Dublin Core Metadata Initiative, URL 〈http://purl.oclc.org/dc/〉

Elaine Brennan, 1998, The DOM for Non-Programmers, 〈TAG〉 Volume 11, Number 10.

ELRA, European Language Resource Association, URL〈http://www.icp.grene.fr/ ELRA/home.html〉

EMIR, European Multilingual Information Retrieval, URL〈http://albion.ncl.ac.uk/esp-syn/text/5312.html〉

Eric Miller, 1998, "An Introduction to the Resource Description Framework" D-lib magazine URL〈http://www.dlib.org/dlib/may98/miller/05miller.html〉

GCA, Graphic Communications Association, URL〈http://www.gca. org/〉

George A. Miller et al, 1990, "Introduction to WordNet: An On-line Lexical Database", CSL Report 43, Cognitive Science Lab., Princeton University.

GILS, U. S. Geological Survey, Government Information Locator Service, URL〈http://www.usgs.gov/gils/〉

ICAME, International Computer Archive of Modern and Medieval English, URL〈http://www.hd.uib.no/icame.html〉

IETF, URN Working Group Discussion Space, URL〈http://www.bunyip.com/research/ietf/urn-ietf/〉

Interoperability, 1998, Position Paper of the Zurich Meeting, "EU-NSF Digital Library Working Group on Interoperability between Digital Libraries" URL〈http://www.si.umich.edu/UMDL/EU__Grant/interop/int__rep1.htm〉

ISO/IEC JTC1/WG4, URL〈http://www.ornl.gov/sgml/sc34/sc34home.htm〉

IPE, Intellectual Property and Economic Issues for Digital Libraries : A Framework for Research, URL〈http://www.si.umich.edu/UMDL/EU__Grant /ipe/ dl.ipe-white-paper.html〉

KORDIC, Korea R&D Information Center, URL〈http://www.kordic.re.kr〉

KRIC, Korea Research Information Center, URL〈http://www.kric.ac.kr〉

LC, the Library of Congress, URL〈http://www.loc.gov/〉

LDC, Linguistic Data Consortium, URL〈http://www.ldc.upenn.edu/〉

Lloyd Rutledge, 1996, HyTime: ISO 10744 Hypermedia/Time-based Structuring Language General Description of HyTime, URL〈http://dmsl.cs.uml.edu/standards/hytime.html〉

Lorcan Dempsey, 1996, "The Warwick Metadata Workshop: A Framework for the Depployment of Resource Description", URL〈 http://www.oclc.org/oclc/research/publications/review96/warwick.htm〉

MARC, Library of Congress, Machine-readable cataloging(MARC), URL〈http:// lcweb.loc.gov/marc/〉

Michael Biezunski, 1997, "General Introduction to Topic Mapping", SGML Europe 97 Conference,

Barcelona

Michael Roszkowski and Chirstoper Lukas, June 1998, "A Distributed Architecture for Resource Discovery Using Metadata", D-Lib Magazine, URL〈http://www.dlib.org/dlib/june98/scout/06roszkowski.html〉

Moo-Young Han, URL 〈 http://www.duke.edu/~myhan/〉

OCLC, Online Computer Library Center, URL 〈http://www.oclc.org/〉

Penn TreeBank, URL〈http://www.cis.upenn.edu/~treebank/home.html〉

Pie Vossen, Laura Bloksma, 1998, "The EuroWordNet Base Concepts and Top Ontology", Version 2, Final, EuroWordNet : LE2-4003, University of Amsterdam

Ron Daniel, 1997, Resolution of Uniform Resource Identifiers using the Domain Name System, Internet Draft, URL〈http://www.acl.lanl.gov/URN/ naptr.txt〉

TEI, Text Encoding Initiative, URL〈http://www-tei.uic.edu/orgs/tei/〉

TIPSTER Text Program, URL〈http://www.nist.gov/itl/ div894/894.02/related__ projects/tipster/〉

TREC, Text REtrieval Conference, URL 〈http://trec.nist.gov/〉

URC, Universal Resource Characteristics, URL 〈http://www.acl.lanl.gov/URC/〉

URI, Universal Resource Identifiers, Naming and Addressing: URIs, URL〈http://www.w3.org/Addressing/Addressing.html〉

W3C, WWW Consortium, URL〈http://www.w3c.org/〉

WWW VL, WWW Virtual Library, URL〈http://vlib.stanford.edu/Home.html〉

XML, The W3C XML Extensible Markup Language Working Group, URL 〈http://www.w3c.org/XML/〉