웹기록물 아카이빙 기반기술 연구 개발

차승준, 이규철*
충남대학교 컴퓨터공학과

Research and Development of Base Technology for Archiving Web Records

Seung-Jun Cha, Kyu-Chul Lee

요 약

웹에서 생성되는 중요한 자원이 점차 증가하고 있지만, 웹이라는 특성 때문에 보존되지 않고 사라져버리는 문제점이 있다. 따라서 본 논문은 보존되지 않고 사라져가는 웹기록물의 보존을 위한 웹 기록물 아카이빙 기반기술을 연구 개발에 관한 것이다. 우선 웹기록물 아카이빙의 절차인 워크플로우에 관해 설계하였고, 그 중 보존과 전달에 필수요소인 메타데이터 요소를 개발하였다. 마지막으로 웹기록물 아카이빙의 전반적인 사항을 정의하는 아키텍쳐를 정립하였다.

Abstract

Although the important information are significantly increasing in the Web, it is easily disappeared because of the Web's characteristics. The purpose of this paper is researching and developing of base technology for archiving web records. First developing about the workflow which is the procedure of web archiving, and developing about metadata elements, the essential condition of preserving web records and accessing archived resources. Finally developing the architecture defined the overall items of web records archiving.

* 교신저자

1. 서 론

1990년대 Tim Berners-Lee가 웹을 창시한 이래로, 인터넷은 정보유통에 혁명적 변화를 가져왔다. 웹은 양방향이면서, 거의 무시할 수 있는 수준의 낮은 유통비용, 정보의 실시간 획득과 제공 가능, 개인이 직접 정보를 획득·가공·배포를 할 수 있는 등 정보유통의 모든 측면에서 한계를 극복했다

이러한 웹의 발전과 더불어 정보원으로서의 웹에 대한 의존성이 점점 증가하고 있다. 2002년 현재 검색엔진에 의해 검색되는 웹의 정보량은 167 테라바이트로 미국의회도서관 장서량의 17배이며 1999년보다 최소한 3배이상 증가하였다. 특히 컴퓨터공학분야의 경우, 온라인 논문의 인용도가 인쇄형 논문 인용도의 2.6배가 되었음을 확인할 수 있고 특히 1990년에서 2000년 사이에 4.5배 증가됨을 확인할 수 있었다[4].

하지만 웹은 편재성 뿐만 아니라 일시성이라는 한정된 속성을 가지고 있다. 2001년 한 연구에서는 웹 페이지의 평균 수명은 약 75일에서 100일 정도로 추산하였다. 따라서 중요성이 있는 웹기록물은 보존되지 않고 사라져버리는 문제점이 있다.

따라서 본 논문에서는 이러한 문제점을 해결하기 위해, 웹기록물의 저장 및 보존을 위한 웹기록물 아카이빙의 기반기술에 대해 연구하였다. 웹기록물 아카이빙의 기반기술로는 웹기록물의 선별에서 수집, 보존까지 절차를 설명하는 워크플로우 설계, 보존 및 검색에 필수 요소인 메타데이터 요소 설계, 아카이빙의 전반적인 내용을 구성하는 아키텍쳐 설계로 나눌 수 있다.

이에 따라 본 논문은 다음과 같이 구성되었아. 2장에서는 국내외 웹 아카이빙 사례에 대해 조사하였고, 3장에서는 사례조사를 바탕으로 워크플로우를 정의하였다. 4장에서는 워크플로우의 보존/전달에 중요한 요소인 메타데이터 요소에 대해 정의하였고, 5장에서는 전체적인 아키텍쳐를 설계하였다. 이러한 내용은 6장에서 결론을 맺고, 향후연구에 대해 논하였다

2. 국내외 아카이빙 사례조사

1994년 캐나다 국립도서관(National Library of Canada)의 EPPP(Electronic Publications Pilot Project)가 시작된 이후 웹 자원에 대한 관심이 여러 국가도서관으로 확산되었다. 특히 WAIS를 개발한 Brewster Kahle이 동료와 함께 1996년 4월에 인터넷 아카이브(Internet Archive)[9]를 설립한 것은 웹기록물의 아카이빙에 관한 대중적인 관심을 이끌어 냄으로서 각국의 프로젝트 추진에 큰 힘이 되었다. 인터넷 아카이브는 처음부터 공공자료를 수집하여 보존하고 역사가, 연구자, 학자 등에게 장기적으로 이용시키는 디지털도서관을 표방하였다[2]. 이후로 지금까지 웹 자원의

<표 1> 국내외 아카이빙 사례 조사

국가	사업명	추진주체	수집방법	접근성	규모
호주	PANDORA	호주 국립도서관	선택	공개	353Gb
영국	Britain on the Web(Domain UK)	영국 국립도서관	선택	비공개	30Mb
일본	WARP	일본 국립국회도서관	선택	공개	524Gb
미국	MINERVA	미국 의회도서관	선택	비공개	35 사이트
덴마크	netarchive.dk	Royal Library 와 The state and University Library	선택	비공개	280Gb
프랑스	BnF Web Archiving initiative	Biblothque nationale de France	선택/포괄	비공개	1 Tb
노르웨이	PARADIGMA	노르웨이 국립도서관	선택/포괄	제한적 공개	140Gb
스웨덴	Kulturarw3	Koninklijke bibloteket(KB)	포괄	제한적 공개	6Tb
핀란드	EVA	핀란드 국립도서관	포괄	비공개	401Gb
오스트리아	AOLA	ONB/TY Wien	포괄	비공개	448Gb
미국	Internet Archive	InternetArchive	포괄	공개	150Tb

수집과 보존을 위한 시도가 다양한 형태로 이루어지고 있다. <표 1>은 그 중 대표적인 사례에 대해 국가별 사업명, 추진주체, 수집방법, 접근성, 아카이빙 된 규모 등을 조사한 것이다.

3. 워크플로우 정의

웹기록물 아카이빙에 대한 워크플로우는 [그림 1]과 같다. 이는 국내외 아카이빙 사례에서 조사한 내용과, 여러 연구들을 살펴보고 국내에 적용가능한 방안으로 설계하였다.[5] [8].

3.1 선별

웹기록물 아카이빙에 있어 적당한 선별정책을 결정하는 것은 반드시 선행되어야 한다. 현재 웹기록물 아카이빙은 도서관, 박물관, 연구기관, 상업기구 등 다양한 곳에서 시행되고 있으며 각각마다 알맞은 선별 방법을 사용해야 한다.

선택의 범위에 따라 다른 접근법들이 존재한다. 크게 비선별적(Unselective), 주제적(Thematic), 선별적(Selective)으로 나눌 수 있다.

3.2 수집방법 및 수집

수집방법(Collecting Methods)으로는 크게 컨텐츠-기반 수집(Content-driven collection)과 이벤트-기반 수집(Event-driven collection)으로 나눌 수 있다. 컨텐츠-기반 수집은 컨텐츠에 기반한, 즉 웹 사

이트의 기본적인 컨텐츠를 아카이빙하기 위한 방법이다. 이벤트-기반 수집(Event-driven collection)은 웹 서버와 브라우저 사이에 발생된 실제적인 트랜잭션(transaction)을 처리하는 것이다.

3.3 품질보증 및 목록화

품질보증(Quality assurance)은 웹기록물 아카이빙 절차에 필수적인 구성 요소이다. 품질보증의 본질 및 등급은 요구사항과 수집하기 원하는 리소스에 많이 좌우된다. 품질보증은 크게 사전-수집 테스팅(Pre-collection testing)과 사후-수집 테스팅(Post-collection testing)이 있다.

목록화(Cataloguing)는 적절한 관리나 사용자들의 접근을 제공할 수 있게 하도록 아카이빙된 컬렉션에 메타데이터와 같은 설명정보를 포함시키는 것이다.

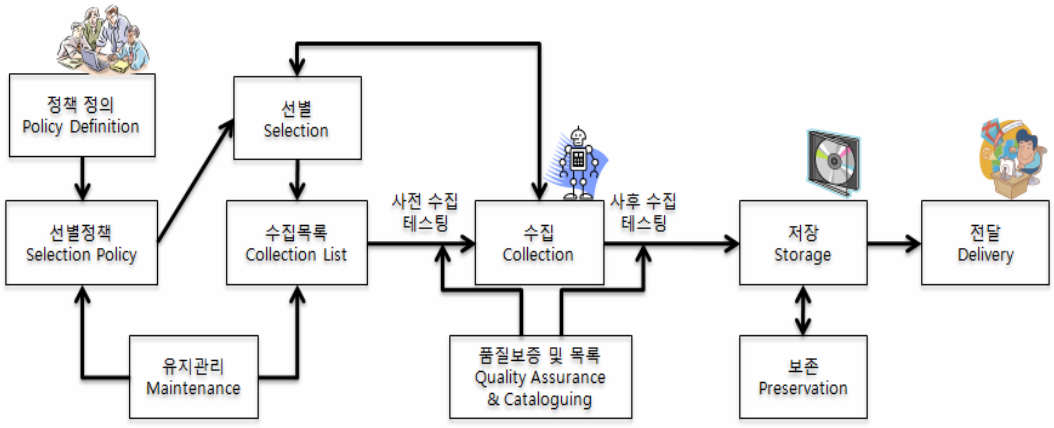
3.4 보존

보존(Preservation)의 목적은 대상의 가치가 유지되면서, 영속적인 접근을 보장하는 것이다. 즉, 성공적인 보존이란 사용자들의 접근이 가능하며 본래의 가치를 그래도 보존하여 전달해야 하는 것이다. 이러한 웹 기록물의 보존은 디지털 대상의 보존과 유사하여 같은 기술이 적용된다.

특히 장기의 보존을 위해서는 설명정보의 메타데이터와 기술정보의 메타데이터의 저장이 기본적인 필수 사항이다. 이는 메타데이터가 웹기록물에 대해 이해하고 해석하는 방법을 제공하기 때문이다.

3.5 전달

전달(Access)은 보존된 아카이빙된 웹기록물을 사용자가 사용할 수 있게 제공해주는 것을 말한다. 아카이빙된 웹기록물에 대한 발견이나 확인에 있어 기본적인 메소드는 크게 검색 접근(Searching access)과 열람 접근(Browsing access)이 있다.



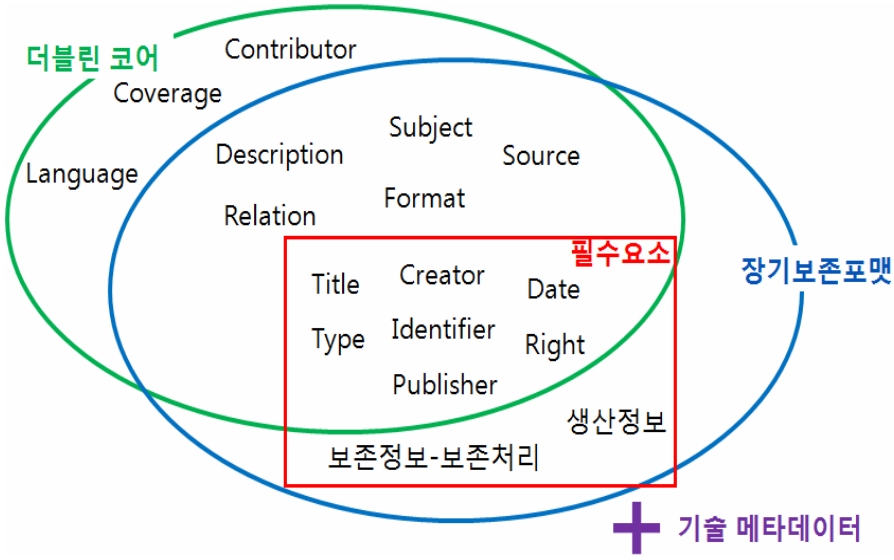
[그림 1] 워크플로우 설계

4. 메타데이터 개발

국내외 웹 아카이빙 사례에서 메타데이터를 정의할 때 설명적 정보를 기술하기 위해 더블린 코어를 사용한다는 것을 확인할 수 있었다. 즉 더블린코어를 웹기록물에 대한 기본적인 메타데이터로 정의한 것이다. 또한 웹기록물은 대부분 자동적으로 수집되어 저장되고 그 양 또한 많기 때문에 사람이 일일이 저장하는데 어려움이 있다[7]. 따라서 설계하려는 웹기록물 메타데이터가 세계적으로 상호 호환성을 유지하고, 자동화를 위해 더블린 코어 메타데이터 형식으로 정의되어야 한다.

또한, 웹기록물은 전자적으로 되어 있다는 점에서 관계법령에서 정의하는 전자기록, 전자문서의 범주에 속할 수 있고, 단지 기록물이 웹사이트에 있다는 점에서만 차이가 있다. MINERVA에서 살펴보면 MARC를 확장한 형태로 전자기록물을 위해서 확장을 하였고, 웹기록물도 전자기록물의 일부라고 정의하여 모든 것들을 하나의 프레임워크(framework) 안에서 검색을 할 수 있도록 전자기록물 형식으로도 정의되어야 한다. 즉 전자기록물과의 호환성을 위해 전자기록물 메타데이터 표준인 장기보존포맷[3]으로도 정의되어야 한다.

메타데이터 요소를 기술할 때 텍스트 기반인 소프트웨어나 하드웨어에 의존적이지 않고 개방적인 표준인 XML을 이용한다. 또한 메타데이터 요소들간의 관계를 설정하고 이를 의미있게 하기 위해 메타데이터는 스키마에 의해 구조화될 필요가 있다. 따라서 XML Schema를 사용해서 문서구조를 정의

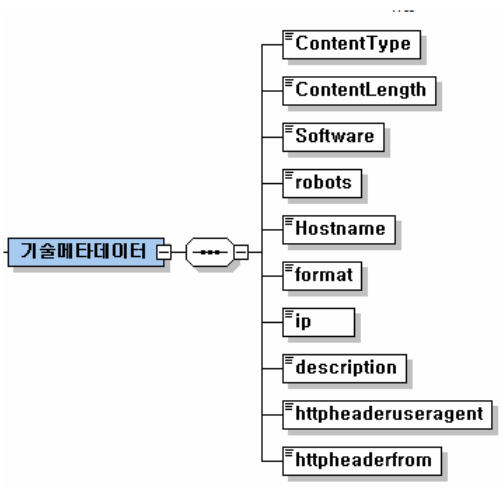


[그림 2] 메타데이터 요소 구성도

했다.

[그림 2]는 웹기록물의 메타데이터 구성도이다. 더블린코어와 장기보존포맷이 서로 공통된 부분에 대해서는 각각 따로 저장할 필요없이 기본적으로는 장기보존포맷 형식으로 정의하고 더블린 코어에서는 정의된 내용을 참조하여 사용할 수 있게 설계하였다. 장기보존포맷에는 필수 정보이지만 더블린코어에 정의되지 않은 '생산정보', '보존정보-보존처리'의 내용을 추가적으로 정의한다.

추가적으로 실제 웹기록물의 특성에 맞추어 처리하기 위한 기술(Technical) 메타데이터가 [그림 3]과 같이 정의되어야 한다. 웹기록물 아카이빙 에서는 IIPC에서 표준으로 정한 WARC 파일 포맷에서 기술 메타데이터를 추출하여 저장하도록 설계하였다.



[그림 3] 기술 메타데이터 요소

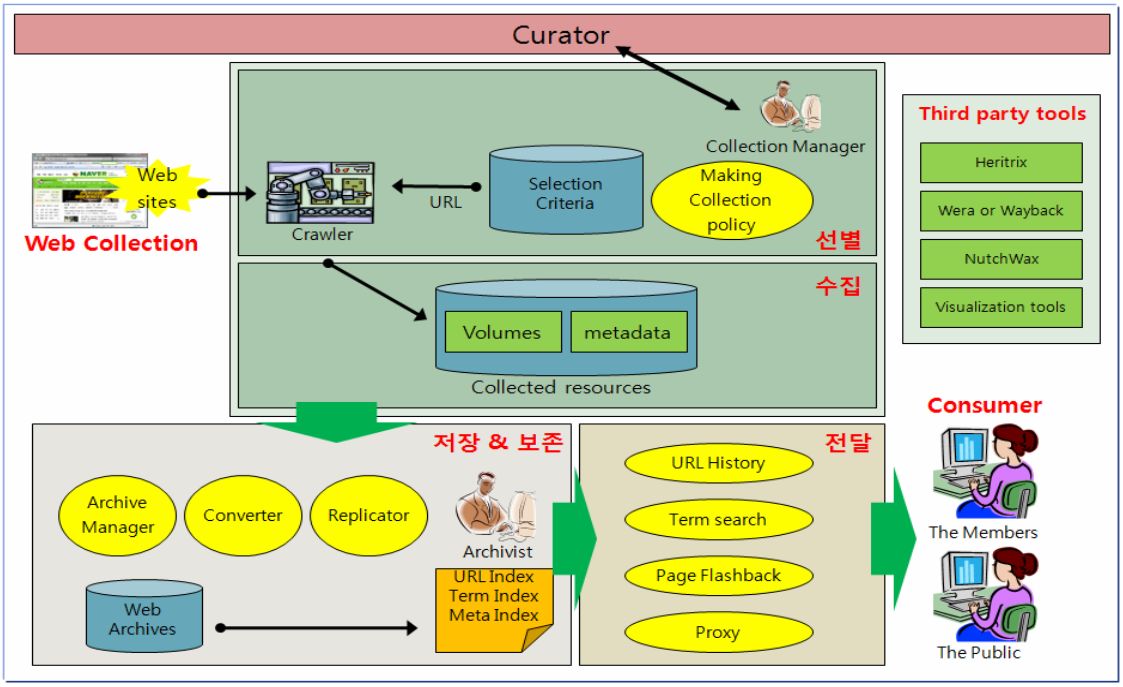
5. 아키텍처 정립

앞에서 정의한 웹기록물 아카이빙 워크플로우를 구체화하여 웹기록물 아카이빙 시스템이 가져야 할 구성 요소 및 행위자(Actor)를 정의한 아키텍처는 [그림 4]과 같다.

웹기록물 아카이빙 아키텍처는 OAIS 참조모형(OAIS Reference Model)의 내용을 준수하였다[1]. 구성요소로는 크게 큐레이터, 선별과 수집, 저장 및 보존, 전달로 정의되며, 아키텍처에 사용되는 도구와 각 단계별 행위자(Actor)도 정의되었다.

큐레이터(Curator)는 콜렉션, 시드들(seeds), URI, 크롤링 프로필 관리, 페이지 출력 등 웹기록물 아카이빙에 대한 전반적인 사항을 관리한다. 선별은 콜렉션 관리자(Collection manager)에 의해 정해진 선별정책이 결정되어 선별 기준(Selection criteria)이 저장되며, 선별정책에 의해 선정된 웹사이트는 크롤러(Crawler)가 수집하게 된다. 크롤러를 통해 수집된 데이터는 수집된 자원(Collected resource)에 실제 아카이빙 된 데이터인 볼륨(Volumes)과 아카이빙에 대한 메타데이터의 정보를 저장하는 메타데이터(Metadata)로 구성되어 있다.

저장과 보존은 기록 보존인(Archivist)에 의해 시행되며 수집을 통해 저장된 자원들(Web Archives)에 대해, 각 문서의 URL,



[그림 4] 아키텍쳐 설계

단어(Term), 메타데이터에 대한 인덱스를 작성한다. 아카이브 관리자(Archive manger)는 아카이빙된 데이터를 WARC, XML, 텍스트 형식으로 변환하여 저장해주는 역할을 하며, 변환기(Converter)는 이러한 형태로 변환시켜준다. 복제기(Replicator)는 크롤링이 끝난후 보존시에 백업이미지를 위해 이미지에 대한 복사본을 생성한다.

전달은 아카이빙된 자원을 복원하여 사용자가 사용하게 제공해는 것으로, URL 히스토리(History)를 통해 URL 별 아카이빙된 각 버전을, 단어 검색(Term Search)을 통해 해당 단어를 가진 문서 검색을 제공해준다. 페이지 복귀(Page Flashback)에서는 아카이빙에 저장된 내용을 브라우져에 출력할 수 있게 플래쉬, 오디오 등의 컴포넌트를 변환시킨다. 프록시(Proxy)에서는 아카이빙된 내용을 출력하기 위해 쉽게 접근할 수 있게 해준다.

아키텍쳐에 사용되는 도구로는 수집기로는 IIPC에서 만든 오픈소스인 Heritrix, 아카이빙된 자원을 전달하는 도구로 NutchWAX & Wera, Wayback과 다른 출력하는 도구들이 사용될 수 있다.

6. Acknowledge

본 연구는 지식경제부 및 정보통신연구진흥원의대학 IT연구센터 육성·지원사업(IITA-2008-C1090-0801-0031)의 연구결과로 수행되었다. 또한 본 연구는 행정안전부 국가기록원의 지원을 받아 기록물 보존기술 연구개발(R&D) 사업의 일환으로 이루어졌으며, 이에 감사드린다.

7. 결 론

본 논문의 목적은 사라져가는 웹기록물에 대해 수집·보존·서비스하는데 필요한 웹기록물 아카이빙 기반구조를 개발하는 것이다.

이를 위해 우선 워크플로우를 연구 개발하였다. 선별위원회에서 적절하게 정해진 선별정책에 의해 사이트를 선별하여, 정해진 수집방법을 통해 수집하게 된다. 수집된 데이터는 품질보증 및 목록화에 의해 진본성을 유지하며, 일정한 매체에 저장되어 장기간 보존된다. 저장되고 보존된 아카이빙은 전달 시스템을 통해 사용자에게 전달된다.

메타데이터를 요소를 개발하기 위해 우선 국내외 아카이빙 사례에 대해 조사하였다. 대부분의 프로젝트가 더블린 코어(Dublin core)를 기반으로 하여 설명정보를 정의하였다. 따라서 세계적인 프로젝트와의 호환성을 위해 더블린 코어 기반의 메타데이터 정의가 필요하다. 또한 웹기록물도 전자기록물의 일부라고 정의하여 모든 것들을 하나의 프레임워크(framework) 안에서 검색을 할 수 있도록 전자기록물 형식으로도 정의되어야 한다. 추가적으로 실제 웹기록물의 특성에 맞추어 처리하기 위한 기술(Technical) 메타데이터가 정의되어야 한다. 웹기록물 아카이빙 에서는 IIPC에서 표준으로 정한 WARC 파일 포맷에서 기술 메타데이터를 추출하여 저장하도록 설계하였다.

또한 웹기록물 아카이빙 아키텍쳐를 연구개발하였다. 웹기록물 아카이빙 아키텍쳐는 OAIS 참조모형(OAIS Reference Model)의 내용을 준수하였다.

본 논문을 통해 웹기록물 수집에 대한 프로세스를 정립할 수 있다. 이러한 프로세스를 바탕으로 정부 및 공공기관의 웹기록물 생성 및 수집을 체계화 하고, 웹기록물을 활용한 통계 등으로 앞으로의 활용을 극대화 시킬 수 있을 것이다.

또한 아카이빙 된 웹기록물을 통해 웹 정보검색 기술을 더욱 활성화 시킬 수 있을 것이며, 보존과정을 바탕으로 행정정보 데이터베이스 보존기술로도 활용될 수 있을 것이다. 뿐만 아니라 기술을 더욱 발전시켜 요즘 많이 연구되고 있는 사용자 참여를 중심으로 한 웹 2.0의 보존기술로도 활용될 수 있을 것이다.

참고문헌

[1] 이소연, 2002, “디지털 아카이빙의 표준화와 OAIS 참조모형”, 「정보관리연구」, 33(3), 45-68

[2] 이재윤, 최원태, 이수상, 2004, “디지털 아카이빙의 현안과 과제”, KERIS 연구자료

[3] 행정안전부 국가기록원, 2008, 전자기록물 장기보존포맷 기술규격(Standard of Archival Information Package)

[4] Antonio Gulli, Alessio Signorini, 2005, "The Indexable Web is more than 11.5 billion pages" [cited 2008-11-17] <http://www.cs.uiowa.edu/~asignori/web-size/>

[5] Adrian brown, 2006, 「Archiving Websites-a practical guide for information management professionals

[6] Julien Masanes, 2005, "Web Archiving Methods and Approaches: A Comparative Study", LIBRARYTRENDS, 54(1), 72-90

[7] Julien Masanes, 2006, "Metadata for Web Archiving", Joint Workshop on Future-Proofing Institutional Websites

[8] Niels Brugger, 2005, 「Archiving Websites」

[9] The National Archives, <http://www.nationalarchives.gov.uk/>

저자소개

차승준(e-mail : juni@cnu.ac.kr)은 2006년 충남대학교 한문학 학사 및 컴퓨터공학 학사를 취득하고 2006년 하기부터 현재까지 충남대학교 대학원 컴퓨터 공학과 석사통합과정에 재학 중이다. 관심분야는 웹 서비스, Web 2.0, GIS 및 웹 아카이빙이다.

이규철(e-mail: kclee@cnu.ac.kr)은 1984년 서울대학교 컴퓨터공학과 학사를 취득하고, 1986년 서울대학교 컴퓨터공학과 석사를 취득하였으며, 1990년 서울대학교 컴퓨터공학과 박사를 취득하였다. 1989년부터 현재까지 충남대학교 교수로 재직중이다. 관심분야는 XML, 웹 서비스, 시맨틱 웹 서비스, 유비쿼터스 웹 서비스이다.