4

시맨틱 웹과 의미적 연결성: 웹 사이언스를 위한 출발점

삼성전자 ▌김 학 래

1. 월드와이드웹의 등장

웹의 주창자인 팀 버너스리(Tim Berners-Lee)는 인류 사회의 보편적 가치를 지원하기 위해 웹은 기술적·사회적 진화를 거듭한다고 설명한다. 인류사회가 웹을 사회문화적 수단으로 인식하는데 소요된 지난 20여년 동안 웹 기술은 빠른 속도로 진화해 "무엇이든지 서로 연결되는 공간"으로서 웹의 이상(理想)은 실현되었다. 1989년 팀 버너스리가 제안한 "Information Management: A Proposal"은 하이퍼텍스트 시스템을 구축해 정보를 공유하는 목표를 갖고 있었고[1], 1993년 CERN은 모든 사람들이 자유롭게 월드와이드웹(World Wide Web) 기술을 사용할 수 있음을 발표한다. 학문적 영역에 있던 인터넷 기술이 일반 사용자에게 공개된 후 월드와이드웹은 웹 광고(Global Net Navigation, 1993년), 백안관의 공공정보 제공(http://whitehouse.gov, 1994년), 월드와이드웹 컨소시엄 구성(1994년) 등 다양한 분야에서 폭발적 성장을 이어가며 새로운 정보 공간으로 진화하게 된다. 이러한 과정에서 독립적이거나 폐쇄적인 환경에 있던 수 많은 데이터가 분산화된 정보 공간으로 옮겨지고, 이를 수용할 수 있는 기술의 진보를 이끌어 낸다.

팀 버너스리가 제안한 웹의 핵심은 하이퍼텍스트(Hypertext)로 공간에 제약없이 정보자원을 연결하는 것이다. 하이퍼텍스트는 정해진 목차가 존재하지 않는 주석(annotation)이 붙은 링크(link)를 통해 인용된 원문을 참조할 수 있는 새로운 형태의 텍스트를 뜻한다. 쉽게 풀이하면, 문서를 유기적으로 연결하고 상호 연관성을 규정하는 것을 뜻한다. 하이퍼텍스트의 역사는 1945년으로 거슬러 올라간다. 바네바 부시(Vannevar Bush)는 메멕스(Memex, 1945년)에서 마이크로필름으로 만들어진 문서를 서로 연결하기 위한 개념으로 하이퍼텍스트를 소개했고, 마우스를 개발해 유명해진 더글라스 엥글바트의 NLS(oNLine System), 테드 넬슨(Ted Nelson, 1965년)의 제나두(Xanadu) 프로젝트를 통해 구체화된다.

월드와이드웹에서 하이퍼텍스트 개념을 현실화하는데 HTML(Hypertext Markup Language), HTTP(Hypertext Transfer Protocol), URI(Uniform Resource Identifier) 등 중요한 기술이 소개되었다. 먼저 HTML은 하이퍼텍스트 기능을 가진 문서를 만드는 언어로, 일반적으로 홈페이지를 만들기 위한 언어로 이해된다. HTTP는 월드와이드웹에서 정보를 주고 받을 수 있는 프로토콜인데, 클라이언트와 서버 사이에서 이루어지는 요청(request)과 응답(response)을 통해 HTML 문서를 전송한다. 예를 들어, 웹 브라우저가 HTTP를 통해 서버로부터 웹 페이지나 그림을 요청하면, 서버는 해당 요청에 대해 필요한 정보를 전달하게 된다. 마지막으로 URI는 가상의 정보공간에서 공개된 정보를 참조하기 위한 문자열로 자원을 식별하기 위해 사용된다. 월드와이드웹에서 주로 쓰이는 URL(Uniform Resource Locator)은 웹 서버에 있는 파일들의 위치를 표현하기 위해 사용되는 것으로 프로토콜의 종류, 도메인명, 파일의 위치등을 포함한다. 예를 들어, 'http://www.blogweb.co.kr/scot.rdf'는 'http://'프로토콜을 사용하는 도메인 (blogweb.co.kr)의 특정 파일(scot.rdf)의 위치를 참조한다. HTML 문서에서 URL은 하이퍼링크로 표현되는데, "*소나기 블로그*"와 같이 사용된다.

그러나 웹의 사회문화적 영향력이 극대화되는 것에 비해 웹이 갖고 있는 근본적인 문제를 해결하기 위한 시도는 더디게 진행되었다. 그림 1에서 개별 객체들 사이에는 명시적인 관계가 존재한다. 예를 들어, 원(circle)으로 표시된 개체들은 '*의존적인(depends on)*', '*부분의(is part of)*', '*만든(made)*', '*의미하는(refers to)*', '*사용하는(uses)*' 등의 상호 관계로 정의된다. 팀 버너스리는 이미 이때부터 의미적 연결의 중요성을 생각하고 있었고, 궁극적으로 시맨틱 웹의 가능성을 고려했던 것으로 보인다. 그러나 월드와이드웹은 정보자원들의 의미적 관계를 명시적으로 표현

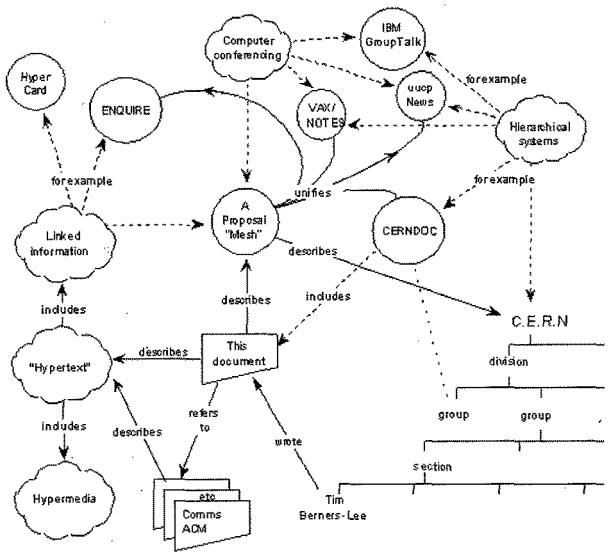


그림 1 팀 버너스리가 제안한 하이퍼텍스트 시스템

할 수 없기 때문에 이를 보완하기 위한 방법이 자연스럽게 논의되어 왔는데 대표적인 것이 시맨틱 웹이다. 시맨틱 웹에 다음 절에서 구체적으로 살펴보자.

2. 시맨틱 웹 소개

2.1 시맨틱 웹이란 무엇인가?

시맨틱 웹(The Semantic Web)이란 "기계가 정보의 의미를 이해하고 처리할 수 있는 거대한 정보의 공간"으로 정의할 수 있다[4]. 기계가 정보의 의미를 이해한다는 것은 사람이 이해하는 수준으로 기계도 세상을 이해할 수 있는 것인데, 풀어서 설명하면 세상에 대한 이해를 컴퓨터 언어 또는 인공지능 언어로 표현하고 이것을 컴퓨터가 사용할 수 있게 구현하는 것을 뜻한다. 또한 거대한 정보 공간은 하이퍼텍스트를 이용해 단순한 물리적 위치의 연결을 넘어 정보 자원들 사이의 의미적 연결이 가능한 가상공간이다. 다음 절에서 구체적으로 설명하겠지만, 온톨로지(ontology)가 시맨틱 웹의 핵심 기술로 강조되는 이유는 정보를 개념화하고 생성된 범주들과 그들 간의 관계를 의미적으로 연결해주기 때문이다.

그렇다면 시맨틱 웹에서 의미는 무엇인가? 언어학, 철학, 수학, 인지과학, 전산학 등에서 시맨틱스(semantics, 의미론)는 언어 기호와 실제 세계에서 언어 표현에 대응하는 지시 대상과의 의미적 관계에 대한 연구를 말한다. 이와 대조적으로 신택틱스(syntactics, 구문론)는 대상에 대한 의미를 고려하지 않은 채 기호와 기호들 간의 형식적 관계를 다룬다. 바꾸어 말하면, 정보자원과 이들 사이의 관계에 적절한 의미를 정의하는 것으로-굳이 표현한다면- 의미론적 웹을 뜻한다. 시맨틱 웹에서 의미는 일종의 조작적 정의(operational definition)로 대상 영역을 명시적으로 개념화하기 위한 수단이고, 동시에 개념들을 의미적으로 연결시킬 수 있는 링크(link)로 이해해야 한다. 즉, 시맨틱 웹은 종전의 웹에서 표현하지 않았던 정보자원에 의미를 부여하고 이것을 바탕으로 논리적 추론이나 향상된 정보처리의 가능성을 제안하는 것이다. 시맨틱 웹은 인공지능(Artificial Intelligence), 지식표현(Knowledge Representation)에서 추구하는 이상을 실현하기 위한 또 다른 시도인가? 이와 관련된 논란은 시맨틱 웹이 제안된 초기부터 지금까지 진행되고 있는데 시맨틱 웹은 인공지능의 한 분야가 아닌, 비록 체계화되지 않았지만 웹이라는 틀에서 실용적으로 해석된다[5].

2.2 온톨로지와 웹 온톨로지 언어

온톨로지(ontology)란 무엇인가? 일반적으로 온톨로지는 철학에서 존재론(Ontology, 대문자 'O')를 뜻한다. 존재론이란 형이상학(metaphysics)의 한 분야로, 세계를 구성하는 대상들의 존재의 본질과 유형에 관한 이론을 탐구한다. 인공지능 분야에서의 온톨로지(ontology, 소문자 'o')는 개념과 그들간의 관계를 표현하기 위해 사용되고 있다. Gruber가 내린 온톨로지의 정의를 원문 그대로 인용하면 다음과 같다[8].

"An *ontology* is a *formal*, *explicit specification* of a shared conceptualization of a domain of interest."

여기서 형식적(formal)이라는 것은 기계가 읽고 처리할 수 있는 인공지능 언어로 표현한다는 의미이고, 명시적(explicit)이라는 것은 개념들의 유형과 개념 사용에 대한 규칙을 분명하게 드러내 정의한다는 뜻이다. 공유되었다(shared)라는 의미는 개념의 사용이 관련된 사람들에게 합의될 수 있다는 것을 의미하며, 개념화(conceptualization)는 실제 세계에 대한 모형을 의미하는 것이고, 관심의 대상이 되는 영역(domain of interest)은 개념이 모형화되는 대상 영역을 말한다.

온톨로지가 시맨틱 웹에서 어떤 역할을 하는지에 대한 의견은 매우 다양하다. 지식표현의 관점에서 온톨로지는 어휘의 내용을 명확하게 정의하고 어휘들로 표현된 사실들 사이의 논리적 관계를 통하여 새로운 사실을 추출해 내는 목적으로 사용된다. 이런 종류의 온톨로지는 풍부한 표현력과 형식성(formality)을 갖고 있기 때문에 복잡한 구조의 지식을 표현하고 논리적 기법을 사용해 추론하는 것을 목표로 구축되는 경향이 있다. 이와 대조적으로 표현력을 최소화하

는 대신 범용적으로 사용될 수 있는 온톨로지도 다수 존재하는데, 이러한 온톨로지는 데이터의 공유와 연결에 초점을 둔다. 예를 들어, Really Simple Syndication(RSS 1.0), 더블린코어 메타데이터(Dublin Core Metadata), Semantically Interlinked Online Community (SIOC)[11], Friend of a Friend(FOAF)[13], Social Semantic Cloud of Tags(SCOT)[12] 등이 대표적이다. 궁극적으로 두 가지 접근 방법은 웹에서 데이터를 연결하는데 상호보완적이지만, 관련 커뮤니티 사이에서 "온톨로지의 본질"에 대한 해석의 차이로 오랫동안 논쟁거리로 남아있기도 하다. 그러나 시맨틱 웹에서 온톨로지는 지식표현이나 인공지능에서 강조하는 논리적 완결성이나 표현력보다 정보자원을 의미적으로 표현하기 위한 범용적이면서 표준화된 언어로 보는 경향이 있다. RDF(Resource Description Framework)와 OWL(Web Ontology Language)이 대표적이다. 전자는 메타데이터를 표현하고 교환하기 위한 프레임워크로, 자원(resource), 속성 유형(property type), 진술문(statements) 등으로 구성된다. 후자는 기계나 에이전트가 처리할 수 있는 풍부한 어휘와 형식적 의미(formal semantics)를 지원한다. 웹 온톨로지 언어는 문서 차원을 넘어 웹에 존재하는 다양한 물리적 개체와 추상적 개체를 의미적 수준에서 표현할 수 있는 표준적인 방법을 제공한다.

3. 시맨틱 웹의 의미적 연결성

3.1 연결성 (connectivity)의 의미

앞에서 언급하였듯이, 웹을 현실화하는 핵심이 링크라는 점으로 볼 때, 시맨틱 웹은 정보자원 사이의 연결성을 명시적으로 정의하는 차이점이 있다. 웹의 연결성은 정보적 측면과 사회적 측면으로 나누어 볼 수 있으며, 먼저 정보의 연결성은 데이터의 유기적 연결을 지향하여 정보 자원의 생태계(ecosystem of information)를 구성하는 기본이 된다. 이와 대비하여 사회적 연결성은 사용자들의 참여에 의해 연결 관계가 형성이 되고 이를 통해 구성되는 생태계 (ecosystem of participation)를 뜻한다[2]. 정보자원의 연결성을 의미적 차원에서 구현하는 기술을 시맨틱 웹이라면, 사회적 연결성을 위한 참여와 개방적 서비스를 구현하는 것이 소셜 웹이다. 궁긍적으로 웹의 연결성은 표현되는 정보자원에 따라 다를 수 있지만 서로 보완적으로 작용한다. 예를 들어, 구글의 검색 알고리즘의 핵심인 페이지 랭크(PageRank)는 특정한 문서에 존재하는 하이퍼텍스트 링크에 가중치를 부여해 순위를 부여한다는 측면에서 정보적·사회적 연결성을 모두 포함한다.

그렇다면 연결성에 있어 의미는 무엇인가? HTML로 표현되는 모든 것은 정보자원의 본래 특징에 관계없이 모두 문서(document)로 표현된다. 다시 강조하지만 문서들 사이 또는 문서 내의 정보자원의 관계는 링크로 연결되지만 명시적인 관계를 정의하지 못한다. 예를 들어, "철수는 파리를 좋아한다"에서 '철수'와 '*파리*'가 개체이고, '*좋아한다*'라는 연결관계가 있을 때, HTML 문서에서 두 개체 사이의 연결관계는 암묵적인(implicit) 형태로만 존재하기 때문에 연결된 링크에 대한 해석은 정보 소비자에 따라 다를 수 있다. 반면 시맨틱 웹에서 개체들은 그림 2의 '좋아한다"라는 관계를 명시적으로 표현할 수 있으며, 기계가 처리 가능한 웹온톨로지언어로 구체화할 수 있다.

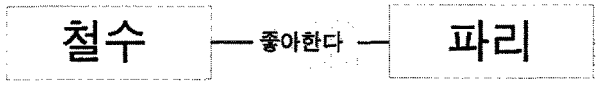
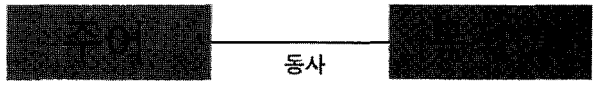


그림 2 명시적 관계의 정의

시맨틱 웹에서 의미적 연결성이 어떻게 다루어지는지에 대해 좀 더 구체적으로 살펴보자. 그림 3은 위에서 설명한 철수와 파리의 의미적 관계를 개념적으로 보여준다. 사실 이러한 모델을 트리플(triple)이라 하는데, 객체(정보자원)와 그들 사이의 관계를 ***개체(Object)-속성(Attribute)-값(Value)***의 데이터 모델 - *A(O, V)* -로 표현한다. 쉽게 설명하면, 트리플은 그림 3과 같이 주어-동사-목적어의 영어 문장구조와 일치한다.



RDF (S-P-O)의 기본 구조

그림 3 속성 중심의 데이터 구조(Subject, Predicate, Object

앞서 설명했듯이, 시맨틱 웹에서 사회문화적 데이터를 의미적으로 정의하는 방법도 그림 2와 그림 3에서 설명한 것과 동일하다. 월드와이드웹에서 연결성은 하이퍼텍스트 링크를 통한 물리적 개체들의 위치적 연결을 의미하지만, 시맨틱 웹은 물리적 개체뿐만 아니라 추상적 개체를 포함하여 정보 자원들 사이의 의미적 연결을 강조하고 있다. 사회적 연결성은 사람들의 관계를 기반으로 형성되기 때문에 HTML 문서의 물리적 표현만으로 한계가 있다. 전통적 의미의 웹에서 '철수'와 '파리'가 연결 관계를 갖고 있다해서

'어떤 사람이 특정 도시를 좋아한다'는 것을 기계가 자동적으로 그 의미를 처리하기 쉽지 않다. 전산학, 언어학, 인공지능 분야에서 개발된 다양한 알고리즘과 기법으로 의미를 찾아낼 수 있지만, 근본적으로 결과를 해석하는 것은 전적으로 정보 사용자에게 달려 있다.

그렇다면 소셜 웹의 개체를 단순히 문서가 아닌 개체 본래의 성격으로 표현할 수 없을까? 그림 4에 있는 개체에서 '철수'는 사람이고, '파리'는 곤충이 아닌 도시라는 공간적 개념이라는 것을 사람들은 직관적으로 알아낼 수 있지만, 기계는 정보자원들 간의 관계에 대해 이해하는 것이 불가능하다. 여기서 "기계가 의미를 이해한다"는 의미를 정리해 보면, 이해하는 것은 이해의 대상이 되는 새로운 것을 이미 알고 있는 다른 것과 관계를 만들어주는 것으로 지칭(referring)과 식별자(identifier)를 필요로 한다. 다시 말해, 기계가 사람처럼 하려면 정보 자원들 간의 관계를 컴퓨터가 처리할 수 있는 지칭과 식별자가 명시적인 언어로 표현되어야 한다. 시맨틱 웹에서 온톨로지의 중요성이 여기에 있다. 웹 온톨로지는 지식 표현과 개념들의 관계를 의미적으로 연결해 주고, 웹에서 처리할 수 있는 데이터 포맷을 제공한다[1]. 그림 4에서 'foaf:Person'는 FOAF(Friend-of-a-Friend) 온톨로지[13]의 클래스로 '철수'가 사람이라는 의미를 정의하였고, '파리'는 SIOC(Semantically Interlinked Online Community) 온톨로지[11]의 'sioc:Space' 클래스를 이용해 공간임을 구체화하고 있다.

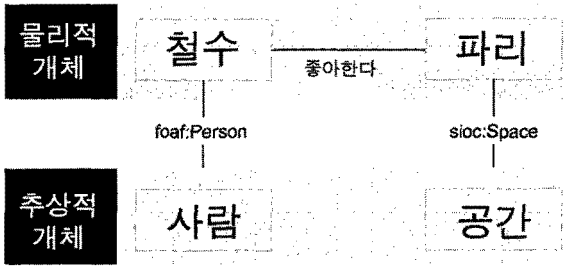


그림 4 시맨틱 웹 환경에서 물리적추상적 수준에서 정보자원의 의미적 연결

소셜 웹이 보편화되고 소셜 웹사이트에서 사회문화적 데이터가 급속히 증가하면서 시맨틱 웹 기술의 필요성도 함께 부각되고 있다. 사람들의 자발적인 참여는 특정 사이트에 한정되지 않으며 신디케이션(syndication)과 웹서비스 API를 통해 데이터의 이식성(portability)은 점차 확대되고 있다. 그러나 분산되어 있는 정보 공간에서 생성된 데이터를 일관성 있게 연결하거나 통합하는 것은 또 다른 차원의 문제이다. 최근 활발히 논의되고 있는 마이크로포맷 형식의 다양한 포맷은 사회문화적 데이터를 유통시키기 위한 대안이 되고 있다. 시맨틱 웹 기술은 이와 같은 문제를 해결하는데 중요한 역할을 할 수 있을 것으로 예상된다.

4. 소셜 웹과 시맨틱 웹의 관계

최근 웹 공간에서 사회문화적 정보활동은 인류사회가 소비자에서 참여자 또는 생산자로 거듭날 수 있는 중요한 계기를 마련해 주고 있다. '웹 2.0'으로 대표되는 소셜 웹(Social Web)은 사용자의 자발적인 참여와 사회적 상호작용에 의해 정보의 가치를 지속적으로 증대시키고, 정보 공유를 이끌어내는 일련의 움직임이다. 소셜 웹의 사회적 특징은 인류사회가 오프라인에서 추구하던 삶의 방식을 온라인 공간으로 옮기는데 큰 영향을 주고 있다. 사용자는 폭 넓은 온라인 커뮤니티에 참여하여 지식을 생산·공유하며 다른 구성원과 상호작용한다. 예를 들어, 온라인 백과사전 위키피디어(Wikipedia)는 컨텐트의 정확성에 대한 비판과 우려에도 불구하고 사용자 참여에 의해 지식이 형성되는 대표적인 사이트로 인식되고 있고, 트위터(Twitter)는 느슨한 형태의 소셜 네트워크를 기반으로 비동기적 커뮤니케이션(asynchronous communication)과 컨텐트를 융합시키는 공간으로 인기가 높아지고 있다.

그렇다면 시맨틱 웹에 있어 소셜 웹은 어떤 존재일까? 시맨틱 웹이 갖고 있던 문제를 해결하는데 소셜 웹의 역할은 무엇인가? 조금 더 구체적으로 살펴보자. 전통적으로 특정 분야의 지식이 소수의 전문가에 의해 구축되는 것과 대조적으로 일반 사용자가 지식 활동에 참여하는 것은 - 정확성에 상관없이 - 보편적이 것은 아니었다. 시맨틱 웹 커뮤니티에서 '시맨틱 데이터 부족'이라는 딜레마를 해결하기 위해 자동적으로 의미를 부여하는 방식(semantic annotation)이 연구되었지만 만족할 만한 결과를 보여주지 못했다. 시맨틱 웹 표준화 작업이 완성도를 높이는 단계에서 이러한 문제는 킬러 애플리케이션이 없다는 비난으로 연결되었고, 학문적 영역에서만 논의할 수 있는 주제로 인식되는 경향을 만들어 내기도 했다.

그러나 소셜 웹은 다르다. 소셜 웹에서 사용자의 자발적 참여와 상호작용에 의해 만들어진 데이터는 특정한 조건하에(terms of conditions 등) 공유된다. 블

1) 웹 온톨로지 언어는 온톨로지를 표현하기 표준화된 언어이지 그 자체가 온톨로지를 의미하지 않는다.

로그, 사진, 동영상, 북마크, 기타 컨텐트 등 다양한 데이터가 해당 사이트를 넘어 전파 및 재생산되고 심지어 통합되기도 한다. 이러한 특징은 시맨틱 웹 커뮤니티에 커다란 변화의 원동력이 되었다. 즉, 시맨틱 웹의 병목점이었던 데이터 부족 문제는 소셜 웹에서 생산되고 공유되는 데이터를 통해 어느 정도 해결되었다. 데이터의 다양성 측면에서 전통적으로 메타데이터를 생산하기 쉬웠던 학술논문이나 서지 정보가 아닌 사회문화적 데이터가 의미화되는 기회를 얻게 되었다. 기술적 측면의 변화는 시맨틱 데이터를 만들고 관리하기 위해 대용량 온톨로지를 구축하는 방식에서 벗어나, 전통적인 관계형 데이터베이스에서 필요한 데이터를 의미화하는 방식으로 전환하게 하였다. 요약해 보면, 2006년부터 시작된 이러한 시도는 – 꼭 구분해야 한다면 – 인공지능 및 지식표현 중심의 표현력이 풍부한 온톨로지를 만들고 추론하려는 전통적인 접근 방식과 다르게 웹을 중심으로 시맨틱 웹을 적용하려는 전환점이 되었다. 이를 바탕으로 웹에서 온톨로지를 표현하고 연결하기 위한 구체적인 논의가 이루어진다. 여기서 중요한 것은 시맨틱 웹 기술의 적응력(adaptability)이다. 지식표현이나 인공지능에서 시도되었던 의미있는 연구 결과들이 현실 세계에 적용되는데 한계가 있었던 것은 변화에 효과적으로 대처하지 못한 이유가 크다. 시맨틱 웹은 표준화된 방법으로 소셜 웹의 문제를 해결하는데 유용할 수 있는데 이와 관련된 다양한 RDF 어휘와 응용 도구들이 개발되고 있다. 소셜 웹을 통해 생성된 데이터들이 "Linked Data"를 통해 표현되어 공유되는 것이 좋은 예이다.

5. 사회문화적 정보의 Linked Data

5.1 링크드 데이터(Linked Data)

사회문화적 데이터에 의미를 부여하는 것은 관계형 데이터베이스의 테이블을 설계하듯이 적절한 수준에서 데이터를 의미적으로 담아내는 과정으로 볼 수 있다. 적절한 수준이란 대상 영역의 정보를 표현하는데 요구되는 표현력을 뜻하며, 이렇게 표현된(웹 온톨로지 언어를 이용해) 의미적 데이터는 서로 다른 정보자원과 의미적으로 연결되거나 통합될 수 있다. 최근 활발히 논의되고 있는 링크드 데이터(Linked Data)는 소셜 웹에서 만들어진 다양하고 폭 넓은 데이터의 상호운용성을 증진시켜 궁극적으로 의미적 데이터의 분산성을 향상시킬 수 있다.

링크드 데이터는 웹에 존재하는 다양한 정보자원을 노출(expose), 공유(share), 연결(connect)하기 위한 방법(method)인데[9], 사전적 의미에서 "Linked"는 "링크로 연결된(connected by a link)"이란 의미를 뜻하기 때문에 "링크로 연결된 의미적 데이터"로 해석할 수 있다. 팀 버너스리는 링크드 데이터의 네 가지 원칙을 다음과 같이 제시하고 있다[6].

- 특정 개념이나 대상을 URI로 명명한다.
- HTTP를 통해 URI로 명명된 정보자원에 접근할 수 있게 한다.
- URI로 접근했을 때 RDF로 정의된 URI에 포함되어 있는 상세정보를 제공한다.
- RDF에 포함되어 있는 다른 정보자원에 접근할 수 있어야 한다.

앞서 설명했듯이, 문서(documents)를 포함해 모든 사물(things)은 URI로 표현되고, HTTP를 이용해 분산된 환경에서 접근할 수 있는 고유한 이름으로 정의된다. URI로 정의된 대상이 RDF 형식이면, 내부에 기술된 정보를 열람하거나 다른 URI와 연결이 가능해진다. 서로 다른 공간에 존재하는 사물이나 개념 사이에 특정한 관계는 RDF 어휘(예: rdfs:seeAlso, owl:sameAs 등)를 이용해 구체화한다. 웹 브라우저에서 HTML 문서를 보여줄 것인지, RDF 문서를 보여줄 것인지는 HTTP 프로토콜에 정의된 내용 협상(content negotiation)에 의해 이루어진다. 이것은 클라이언트와 서버가 서로 협의하여 클라이언트에서 처리가능한 객체(media types), 부호화(encoding), 자연어 등을 서버가 제공해 주는 것을 말한다. 예를 들어, 서버에서 특정 페이지에 대해 영어와 한글로 만들어진 HTML 문서가 있는데 사용자가 한글 페이지를 선호한다면 클라

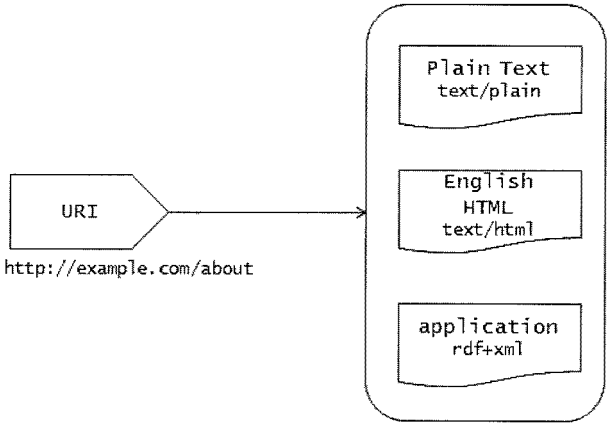


그림 5 내용 협상 (content negotiation)을 통한 정보자원의 접근

이언트 요청에 의해 한글 페이지를 자동으로 보여줄 수 있다. 그림 5에서 'URL'은 데이터 브라우저를 지향하는 Tabulator[2]가 설치된 브라우저는 RDF URI를 인식하여 메타데이터를 열람할 수 있게 해 준다.

링크드 데이터에 대한 개념은 팀 버너스리에 의해 정립되었지만, 웹에서 실현시킨 것은 크리스 비처(Chris Bizer)의 공로가 크다. 크리스는 2006년부터 관계형 데이터베이스에서 온톨로지를 추출해 낼 수 있는 연구를 진행하며 D2R 서버[3]를 개발했고, 이것을 바탕으로 링크드 데이터의 허브 역할을 하는 DBPedia를 만들게 된다. 그림 6에서 보듯이, D2R 서버는 규칙이 정의된 파일(Mapping file)을 관계형 데이터베이스에 연결하여 RDF 어휘를 추출해 주는 기능을 갖고 있으며, 브라우저에서 요청이 있을 때 HTML 또는 RDF로 결과를 보여준다. DBPedia[4]는 위키피디아(wikipedia)에 있는 정보들을 시맨틱 웹 기술을 이용해 구조화된 데이터로 변환한 것으로, SPARQL을 이용해 위키피디아에 대한 의미적 질의를 지원해 준다. DBPedia의 가치는 서비스 기능 및 완성도보다 링크드 데이터를 실현할 수 있는 기반을 제공했다는 측면에서 높게 평가받는다. DBPedia의 데이터는 다른 정보자원들과 의미적으로 연결되는 중요한 자료원이 되고 있다. 더불어 링크드 데이터를 실현하기 위해 검토해야 하는 많은 이슈들을 테스트할 수 있는 역할을 제공한다. 한 가지 강조한다면, 링크드 데이터는 정보자원을 의미적으로 생산하고 공유하기 위한 시맨틱 웹의 한 가지 접근 방법 중의 하나이지, 그 자체가 시맨틱 웹을 대체하거나 개념적 변화를 의미하는 것은 아니다.

5.2 링크드 데이터에서 RDF 어휘의 역할

사회문화적 데이터와 현상을 의미적으로 연결하기 위해 새로운 기술과 표준화가 필요하지만 실제 세계의 올바른 이해를 바탕으로 진행되어야 한다. 소셜 웹이 폭발적 성장의 배경에 사람들의 참여에 의해 발현된 집단지성이 있듯이, 의미적 연결을 위한 작업도 대상 영역의 전문가, 정보 제공자 그리고 소비자의 협업을 통해 이루어질 필요가 있다. 특히, 사회문화적 데이터는 수 많은 사람들의 상호작용에 의해 만들어지기 때문에 사회현상을 분석하는 과정과 이를 표준화된 틀로 담아내는 노력이 병행되어야 한다.

소셜 웹이 활성화되기 전부터 사회문화적 데이터를 의미적으로 표현하려는 노력은 계속되어 왔다. 예를 들어, FOAF는 '친구의 친구'의 의미로 이름, 이메일, 주소와 같은 개인의 신상정보와 친구들의 목록을 표현할 수 있으며, SIOC 은 온라인 커뮤니티 구조와 컨텐트를 기술하기 위한 어휘를 제공한다. SKOS(Simple Knowledge Organization System)은 택소노미(taxonomy)나 시소러스(thesaurus)와 같은 용어들의 개념과 구조를 정의하기 위한 언어이다. 이와 같은 RDF 어휘들을 연결하여 온라인에서 생성되는 다양한 현상을 의미적 수준에서 표현할 수 있는데, 이것이 링크드 데이터의 첫번째 출발점이기도 하다. 그림 7에서 보듯이 어떤 사용자가 온라인 계정을 갖고 있을 경우 FOAF의 foaf:OnlineAccount를 이용해 표현하고 사용자에 대한 구체적인 정보는 sioc:User 클래스를 사용해 구체화한다. 해당 사용자가 관심있는 주제나 커뮤니티의 관심사는 SKOS(Simple Knowledge Organization System)의 skos:narrower(하위 개념) 또는 skos:broader(상위 개념) 속성으로 주제 사이의 관계를 정의한다. 이와 함께 RDF 어휘 사이의 관계에 대한 관심이 높아지고 있는데, UMBEL(Upper Mapping and Binding

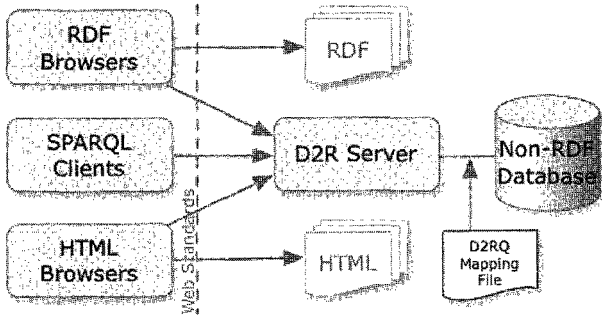


그림 6 D2R 서버 아키텍처

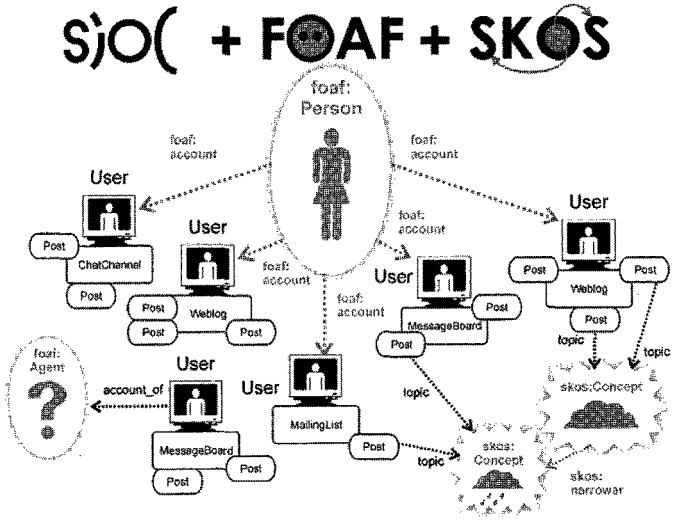


그림 7 SIOC, FOAF, SKOS 어휘의 연계

2) http://www.w3.org/2005/ajar/tab
3) http://www4.wiwiss.fu-berlin.de/bizer/d2r-server/
4) www.dbpedia.org

Exchange Layer) 온톨로지[5]는 앞서 언급한 RDF 어휘 뿐만 아니라 도메인 온톨로지들의 관계를 정의하는데 목표를 갖고 있다. 그림 8에서 보듯이, UMBEL 온톨로지는 다양한 RDF 어휘들의 관계를 정의하고 있다. 예를 들어, sc:ContextualWork는 foaf:Document의 상위클래스(umbel:superClassOf), sc:Person, sc:Agent_

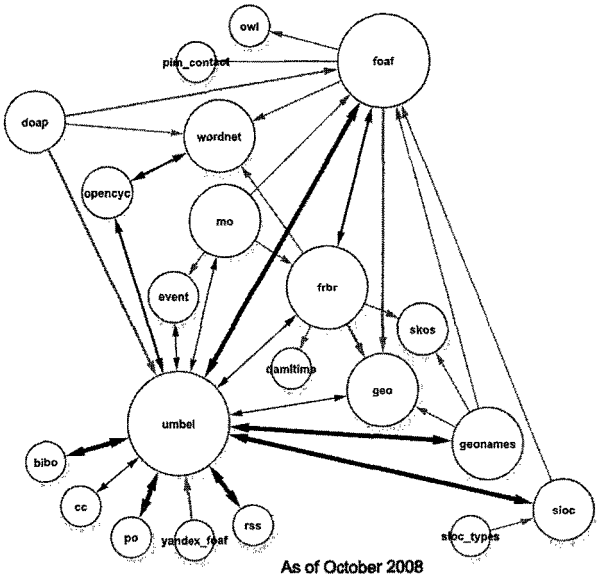


그림 8 RDF (온톨로지) 어휘간 클래스 수준의 관계 구체화

Generic, sc:Organization클래스는 각각 foaf:Person, foaf:Agent, foaf:Organization클래스와 동일한(owl:equivalentClass) 관계로 정의되어 있다. 중요한 점은 RDF 어휘를 재사용할수록 데이터를 연결하기 쉽다는 점이다. 일반적으로 온톨로지 개발을 할 때 보편적으로 많이 활용되는 어휘가 있어도 별도의 어휘로 정의하는 경우가 많은데 이것은 데이터의 연결성을 지원하는데 바람직하지 않다.

5.3 링크드 데이터의 현황

그림 9는 링크드 데이터로 연결된 데이터 집합을 도식화해 보여준다. 그림에서 원의 크기는 웹온톨로지언어로 표현된 데이터의 상대적 크기를 의미하며, 화살표는 다양한 도메인의 정보자원이 연결되어 있음을 표시한다. 2010년 2월 말 기준으로[15,16] 링크드 데이터 클라우드에 13,112,409,691 트리플(triple)이 존재하며 연결되는 데이터 크기는 지속적으로 증가할 것이다(표 1 참고). New York Times[6]와 BBC(BBC 프로그램)에서 링크드 데이터를 지원하는 것은 향후 데이터 공유와 연결의 중요성을 극대화될 것으로 예상할 수 있는 좋은 예가 된다. 이와 함께 최근 각국 정부가 공공정보를 공개하는 움직임과 맞물려 데이터를 의미적으로 연결하기 위한 기술적 접근이 필요하다.

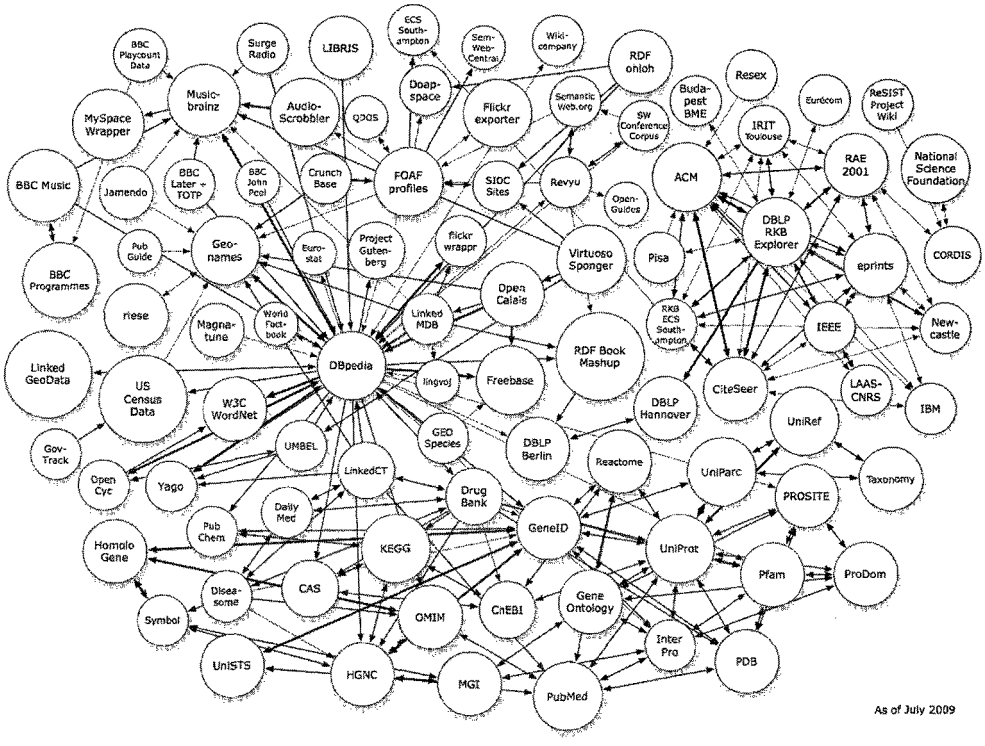


그림 9 링크드 데이터 사이의 연결 관계(그림에서 원은 데이터원을 표시, 원의 크기는 데이터 크기를 의미함)

5) http://www.umbel.org

6) http://data.nytimes.com/

표 1 Top 20링크드 데이터 자료원 및 연결 현황 (연결된 데이터 셋은 가장 많은 자료원만 표시함)

데이터	데이터셋 크기	SPARQL Endpoint	RDF 파일 형식	연결된 데이터셋	링크 수 (range)	링크 수 (actual)
data-gov wiki	5,074,932,510					
LinkedGeoData	3,000,000,000	N	Y	DBpedia	> 10000	53000
US Census Data	1,000,000,000	Y	Y	Geonames	> 100,000	
Bio2RDF:NCBI:PubMed	797,000,000	Y	Y			
AudioScrobbler	600,000,000	Y	N			
Bio2RDF:UniProt:UniParc	490,000,000	Y	Y	Bio2RDF:sgd	> 10000	12806
DBpedia	409,000,000	Y	Y	Freebase	> 1,000,000	2,400,000
Bio2RDF:UniProt:UniRef	338,602,962	Y	Y	Bio2RDF:taxonomy	> 1000000	9833513
Bio2RDF:UniProt:UniProtKB	242,000,000	Y	Y			
Open Archive Initiative(RKB)	216,428,311	Y	Y	dotAC (RKB)	> 100,000	663,467
Bio2RDF:iProClass	191,608,264	Y	Y			
Bio2RDF:NCBI:GeneID	173,132,553	Y	Y			
Freebase	100,000,000	Y	Y	GeoSpecies	> 100,000	
RDF Book Mashup	100,000,000			Revyu	> 100,000	
Geonames	93,900,000			Pub Guide	> 100,000	
FOAFprofiles	60,000,000			AudioScrobbler	> 100,000	
Musicbrainz	60,000,000	Y	Y	Surge Radio	> 100,000	
Bio2RDF:KEGG:Pathway	50,793,314	Y	Y			
Bio2RDF:Affymetrix	45,560,115	Y	Y			
DBLP Hannover	30,000,000			DBLP Berlin	> 1,000	

그러나 일련의 변화가 컴퓨터 과학이나 사회과학이라는 한정된 영역의 예는 아니다. 표 1에서 보듯이, 지금까지 공개되어 연결된 데이터는 음악에서 바이오 분야에 이르는 매우 광범위하다. 시맨틱 웹 커뮤니티를 포함해 컴퓨터 과학 분야에서 의미적으로 링크를 연결하기 위해 자동화된 알고리즘을 개발하고 보급할 수 있지만, 도메인에 따라 일반화시키는 것은 또 다른 차원의 문제가 될 것이다. 여기서 중요한 점은 웹의 사회문화적 역할이 변화함에 따라 웹을 바라보는 관점도 이에 맞춰가야 할 필요가 있다는 것이다. 그동안 웹은 정보활동을 위한 효과적인 공간으로, 사회현상을 발견하기 위한 대안적인 매체로 인식되었지만, 웹 자체를 사회문화적 플랫폼으로 인식하는 발상의 전환이 필요하다. 인류사회가 웹이 없는 현실을 상상하지 못하는 지금 웹이 어떠한 역할을 할 수 있는지에 대한 근본적인 접근과 이를 학문적으로 접근하려는 노력이 필요한 이유이다.

6. 웹 사이언스: 시맨틱 웹의 역할

팀 버너스리가 TED[7]에서 주장했듯이[7], 웹에 가공되지 않은 자료(raw data)가 공유될수록 웹의 연결성은 한층 확대될 것이다. 웹 사이언스(Web Science)는 기존의 컴퓨터 과학 측면의 연구 방법을 넘어 학문의 경계를 넘나들면서 웹의 보편적 가치를 찾는 것을 추구하는 새로운 움직임이다[10]. 과학과 공학은 목적과 대상의 차이가 있는데[1], 과학의 목적은 현상에 대한 설명과 새로운 현상을 예측하는 이론을 만들어 가는 것이고, 공학은 과학적 발견과 이론을 이용하여 삶의 질을 향상시키기 위한 인공물을 만드는 것으로 구분한다. 이러한 관점에서 웹 사이언스는 기술과 사회문화적 중심의 웹뿐만 아니라 인류 사회의 다양한 가치 체계의 접점을 연결해 줄 수 있을 것이다. 김홍기는 웹 사이언스의 중요성을 다음과 같이 강조하고 있다[1].

> '웹 사이언스(web science)'의 탄생은 '인지과학'과 같은 융합학문과 비견될 수 있을 것이다. 인지과학이란 학문은 인간의 인지적 현상에 대해 연구하기 위해 전산학, 심리학, 언어학, 철학, 신경과학과 같은 다양한 학문적 방법론을 통합적으로 수용하면서 탄생했다. 웹 사이언스는 웹 자체와 웹을 매개로 이루어지는 다양한 현상에 대

7) www.ted.com

해 연구하기 위하여 여러 분야의 학문이 학제적으로 결합되는 것을 의미한다. 웹과 관련한 현상이 워낙 광범하고 공학적 연구와도 밀접히 연관되어 있어서 웹 사이언스는 인지과학보다 더 넓은 스펙트럼의 학문적 융합이 요구될 것이다.

기술적 측면에서 웹 사이언스는 소셜 웹과 시맨틱 웹이 융합된 공간에서 혁신적인 기술적 도전을 하게 될 것이다. 웹을 구성하는 데이터 원(source)으로서 사회문화, 비즈니스 관련 데이터는 어느 정도 예측 가능한 수준이지만, 생물학, 법학, 심리학, 의과학 등의 다양한 도메인의 가공되지 않은 데이터가 웹에 공유되어 연결된다면 보다 복잡한 문제에 직면하게 될 것이다. 정보 자원간 또는 사회 네트워크의 의미적 연결성은 전통적인 웹이 갖고 있는 단순한 그래프 연결구조를 넘어 "Giant Global Graph"로 진화할 것이다. 이런 측면에서 시맨틱 웹 커뮤니티가 소셜 웹으로부터 얻은 교훈은 의미적 데이터를 만들어내는데 그치지 않는다. 소셜 웹을 통해 등장한 신뢰, 아이덴티티, 집단지성, 평판 및 윤리 문제 등 윤리철학과 법학과 관련된 이슈들에 대해 심도 깊은 논의는 신뢰의 웹을 구현하는데 중요한 역할을 할 것이다.

7. 결 론

지금까지 시맨틱 웹이라는 새로운 기술을 연결성의 측면에서 바라보며 거대한 정보공간이 어떻게 의미적으로 연결될 수 있는지 살펴보았다. 하이퍼텍스트로 이어지는 연결성은 새로운 가상공간을 인류사회에 제공해 주었고, 사회문화에 혁명적인 영향을 주었다. 웹 2.0으로 대표되는 사회문화적 움직임은 사용자의 자발적 참여와 데이터 공유라는 새로운 패러다임을 만들어 주었고, 웹 중심의 다양한 접근을 가능하게 하였다. 그러나 사람들의 자발적인 참여에 의해 생성된 사회문화적 데이터는 특정한 서비스에 갇혀 있어 데이터를 공유하거나 데이터간의 연결을 통해 창발적 지식을 만들어 내는데 한계가 있었다. 사회적 정보공간에서 연결성의 복잡도가 늘어날수록 이를 해결하기 위해 사회 현상을 분명하게 인식하고 분석해야 하며, 기술의 진보가 함께 이루어져야 한다.

시맨틱 웹 기술은 소셜 웹과 함께 사회문화적 연결성을 지원하는 핵심 기술이다. 시맨틱 웹에 대한 학계·산업계의 관심은 2002년부터 본격화되었지만 뚜렷한 성과를 얻지 못하며 많은 비판에 직면한 것이 사실이다. 팀 버너스리는 시맨틱 웹을 처음 제안했을 때부터 시맨틱 웹이 웹에 적용될 수 있는 단순한 응용 애플리케이션이 아닌 웹 인프라스트럭처를 변환하는 시도로 생각하였다. 웹 2.0이 등장하기 전까지 시맨틱 웹은 표준화와 도메인 온톨로지 중심의 추론이라는 다소 어려운 주제에 연구 역량이 집중되었었다. 그러나 소셜 웹의 출현으로 새로운 기술적 접근을 할 수 있는 기회와 의미적 데이터를 만들어내며 그 필요성을 다시 부각되고 있다. 더불어 여러 학문이 융합되어 발전하는 웹을 공학적 측면이 아닌 새로운 학문 영역으로 인식하려는 시도로 웹 사이언스도 주목받고 있다. 기술적으로 웹 사이언스가 현실화되기 위해 시맨틱 웹은 정보자원의 의미적 연결성을 지원하는데 중요한 요소기술이며, 링크드 데이터는 다양한 데이터 사이에서 의미적 연결성을 만들어줄 수 있는 중요한 접근 방법이 될 것이다.

참고문헌

[1] 김홍기, Semantic Web 2.0: 웹 사이언스의 기술적 기반 (2007). http://www.ibm.com/developerworks/kr/library/dwclm/20070327/ (accessed 27 February 2010).

[2] 김홍기, 태깅(Tagging)의 존재 이유 (2007). http://www.ibm.com/developerworks/kr/library/dwclm/20070710/ (accessed 27 February 2010).

[3] T. Berners-Lee. Information Management: A Proposal (1989). Available at: http://www.w3.org/History/1989/proposal.html (accessed 27 February 2010).

[4] T. Berners-Lee. Semantic Web Road map (1998). Available at: http://www.w3.org/DesignIssues/Semantic.html (accessed 27 February 2010).

[5] T. Berners-Lee. What the Semantic Web can represent (1998). http://www.w3.org/DesignIssues/RDFnot.html (accessed 27 February 2010).

[6] T. Berners-Lee. Linked Data (2006). http://www.w3.org/DesignIssues/LinkedData.html (accessed 27 February 2010).

[7] T. Berners-Lee. The next Web of open, linked data (2009). http://blog.ted.com/2009/03/tim_berners_lee_web.php (accessed 27 February 2010).

[8] T. Gruber. What is an Ontology? (1992). http://www-ksl.stanford.edu/kst/what-is-an-ontology.html (accessed 27 February 2010).

[9] L. Sauermann and R. Cyganiak. Cool URIs for the Semantic Web (2007). W3C Working Draft 17 December 2007. Available at: http://www.w3.org/TR/

2007/WD-cooluris-20071217/(accessed 27 February 2010).

[10] J. Hendler, N. Shadbolt, W. Hall, T. Berners-Lee, and D. Weitzner. Communications of the ACM. 51 (7). p. 60-69. 2008.

[11] J. G. Breslin, A. Harth, U. Bojars, and S. Decker. Towards Semantically-Interlinked Online Communities, in A. Gomez-Perez & J. Euzenat, ed., 'European Semantic Web Conference (ESWC)', Springer, pp. 500-514, 2005.

[12] H.L. Kim, J. G. Breslin and S. Decker. Representing and Sharing Folksonomies with Semantics. Journal of Information Science. 36(1). pp. 57-72. 2010.

[13] D. Brickley and L. Miller, FOAF Vocabulary Specification (2005), available at: http://xmlns.com/foaf/0.1(accessed 27 February 2010).

[14] C. Bizer, R. Cyganiak, and T. Heath, How to Publish Linked Data on the Web (2007), Available at: http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/ (accessed 27 February 2010).

[15] SWEO Community Project: Linking Open Data on the Semantic Web, Statistics on links between Data sets (2009), http://esw.w3.org/topic/TaskForces/CommunityProjects/LinkingOpenData/DataSets/LinkStatistics (accessed 27 February 2010).

[16] SWEO Community Project: Linking Open Data on the Semantic Web, Statistics on Data sets (2009), http://esw.w3.org/topic/TaskForces/CommunityProjects/LinkingOpenData/DataSets/Statistics (accessed 27 February 2010).

김 학 래

2010 아일랜드 국립대학교 전산학 박사
2009~현재 삼성전자 책임 연구원
2006~2009 Digital Enterprise Research Institute, Ireland, 연구원
관심분야 : 소셜 시맨틱 웹, 웹 사이언스, 모바일 웹, 온톨로지 공학
E-mail : haklae.kim@gmail.com