시맨틱 웹과 인문지식

명시적인 기계가독형 언어로 기술된 웹 온톨로지가 월드와이드웹의 세계에서 공유되고, 특정 분야의 지식과 정보를 생산하는 사람들이 이것을 참조하여 상호 소통할 수 있는 데이터를 생산한다면, 그 분야 지식의 활용과 교육, 연구의 양상이 크게 바뀌게 될 것임은 자명하다. 문제는 누가 어떠한 방법으로 그 일을 시작할 것이며, 또 누가 어떠한 역할로 그 일을 발전시켜 가겠는가 하는 것이다.

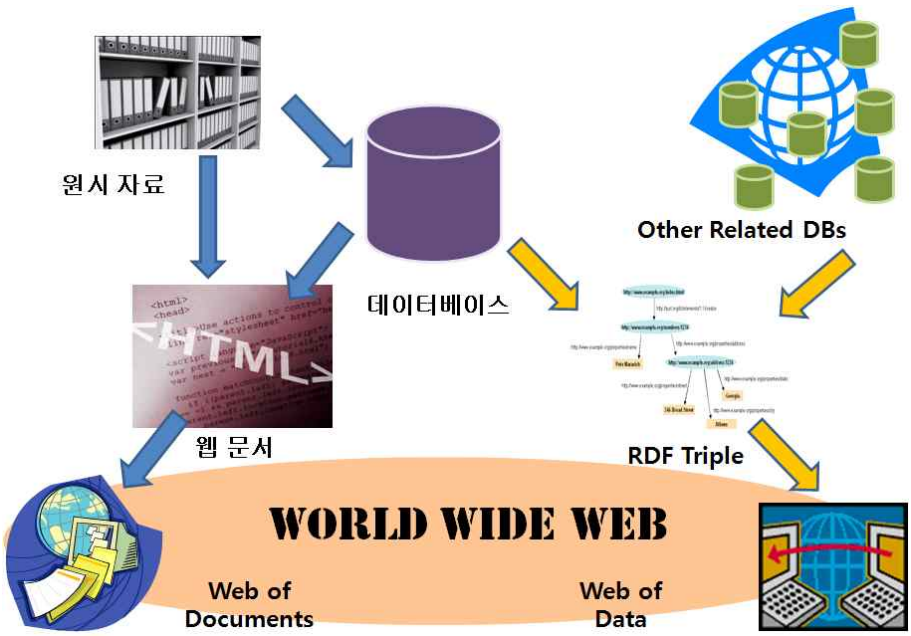
정보화의 대상 세계를 기계적 식별과 분석이 가능하게끔 명시적으로 명세화하자는 것은 시맨틱 웹의 고유한 주장이 아니며, 모든 정보처리 기술의 기본 전제라고 해도 무방하다. 사실상 회사, 공공기관, 특정 목적의 커뮤니티가 운영하는 데이터베이스 안에서는 이미 오래 전부터 시맨틱 웹이 추구하는 것과 다름없는 일들이 일어나고 있었다. 대학에서 운영하고 있는 학사관리 시스템을 예로 들어 보자. 학생들은 누구 하나 예외 없이 '학번'이라고 하는 고유번호를 갖고 있으며, 강의를 담당하는 교수들도 모두 교원번호를 가지고 있다. 모든 개설 교과목은 강좌코드로 식별되고, 강의실과 강의시간도 다른 장소, 다른 시간대와 구별되는 분명한 이름을 가지고 있다. 그 틀 안에서 어느 학생이 어느 과목을 수강하는지, 그 과목은 어느 교수가 담당하고 언제 어디에서 강의하는지 분명하게 알 수 있게 되고, 컴퓨터는 기계적으로 개설 강의 목록과 과목별 출석부, 학생들의 수강 이력과 성적표를 산출한다. 누구든 이 데이터베이스에 접근할 수만 있다면 간단한 질의어 한 문장을 데이터베이스 관리시스템에 입력함으로써 지난 3년간 인문정보학 강의 수강생 가운데 역사, 문학, 철학, 예술 전공 학생의 비율을 조회할 수도 있을 것이다.

시맨틱 웹은 어느 한 기관, 조직, 업무의 폐쇄된 영역이 아니라 월드와이드웹의 무한히 개방된 영역에서 이와 유사한 데이터의 명시적 연계를 이루어내고자 하는 구상이다. 그것은 과연 가능한 일일까?

학교나 회사 내의 사무적인 일이 요구하는 정보와 달리 월드와이드웹의 세계에서는 형식과 내용을 한정할 수 없는 다양한 정보가 무한히 넘쳐나고 있다. 이러한 것을 빠짐없이 취급하는 데이터베이스를 만드는 것은 가능하지 않을 뿐 아니라 필요하지도 않다. 시맨틱 웹에 대해 우리가 오해하지 말아야 할 것은 그것이 전 세계의 모든 데이터를 망라적으로 수용하려는 것은 아니라고 하는 점이다.

시맨틱 웹은 디지털 세계에 존재하는 수많은 층위의 다양한 정보 가운데 그것의 문맥 요소를 명시적으로 식별하고, 그 관계성을 정밀하게 탐색할 필요성

이 있는 것만을 대상으로 삼는다. 그리고 이 '필요성'이라는 것에 정해진 기준이 있는 것은 아니기 때문에, 어떤 영역에서든 지식의 소통과 이를 위한 협업이 필요한 곳에서는 그 분야의 필요성에 부응하는 온톨로지를 설계하고 이를 공유하는 방법으로 기존의 웹의 한 부분을 시맨틱 웹으로 만들어 갈 수 있다.



미래의 월드와이드웹: 기존의 웹과 시맨틱 웹의 공존

특정 조직 내의 데이터베이스 시스템과 달리 시맨틱 웹은 월드와이드웹이라는 개방된 세계에서 만들어지는 것이며, 그 개방의 장점을 최대한 이끌려는 의도를 갖고 있기 때문에 어떠한 주제의 시맨틱 웹이든 그것을 이용하거나 데이터의 생산에 참여할 수 있는 범위는 항상 열려 있다. 하지만 시맨틱 웹이 일종의 '약속'이라는 점을 주목하면, 그 약속을 만들고, 그 약속에 따라 데이터를 만들고, 그렇게 만들어진 데이터를 이용하는 사람들이 존재할 때만 그 세계가 의미를 갖게 된다는 점 또한 부인할 수 없다. '시맨틱 웹'이라는 개념으로 설명하는 '개방적인 데이터베이스'는 특정 기관이나 조직의 소유물은 아니되, 관심사를 공유하는 사람들의 조직적인 노력 없이는 만들어질 수 없는 것이다.

인문학 분야의 학술적인 일에 종사하는 사람들이 '인문지식 시맨틱 웹'을 만들고자 한다면, 그 필요성에 공감하는 사람들이 공동의 약속을 정하고, 그들이 이제부터 만들어낼 '인문지식 데이터' 속에 그 약속에 의한 메타데이터를 첨가하는 노력을 기울여 가야 한다.

그 노력은 한꺼번에 전체를 만들고자 하는 노력이 아니다. 관심이 모여지는 곳에서부터 부분적인 것을 만들되, 그 범위와 경계를 한정하지 않음으로써 향

후에 더 확장될 수 있게 하고, 경우에 따라서는 전혀 다른 영역에서 만들어진 시맨틱 웹 데이터와도 소통할 수 있게 하는 것이다. 예를 들어, 우리나라 여러 지방의 마을 또는 가정에서 전승되고 있는 민간 의학에 관한 정보를 시맨틱 웹 데이터로 생산하는 일을 생각해 보자. 치료하고자 하는 병증, 약재로 쓰이는 식물, 약재를 가공하는 방법, 치료 효과에 대한 주장 등을 일정하게 기술할 수 있는 방법을 정하고 이를 사람들에게 알리면, 대학이나 공공기관의 연구팀이 아닌 개인 블로거들도 자신의 웹 페이지에 쓴 자기 마을의 민간 한방요법 이야기가 시맨틱 웹의 한 노드로 기능하게 할 수 있다. 그리고 시맨틱 웹 상의 다른 정보에 의해 그 블로그에 지방 사투리로 쓰인 약용식물의 이름에 상응하는 학명이 무엇인지 알려지면, 식물학, 화학, 제약학 영역에서 제공되는 정보를 블로거의 글과 연결해서 살펴보고, 그 신빙성을 검증하는 일도 가능해진다.

이러한 일들이 가능성에 대한 담론에 머물지 않고, 이루어질 수 있는 방향으로 나아가기 위해서는 어떠한 실천적인 노력이 필요할까? 나의 웹 문서에 쓰인 정보를 시맨틱 데이터로 기술하는 데 필요한 '이름'을 정하는 일이 우선되어야 한다. 바꿔 말해, 특정 주제의 시맨틱 웹을 위한 온톨로지를 설계하고 그 안에서 주어, 서술어, 목적어로 쓰일 수 있는 어휘들의 표준적인 이름을 제시함으로써 누구든지 그 어휘를 이용하여 '소통할 수 있는' 데이터를 생산하게 한다는 것이다.

이 장의 첫 번째 절에서 살펴보았듯이, '함양 거연정'에 대해 이야기하는 여러 가지 웹 문서들은 '함양 거연정'이라는 건조물, '거연정기'라는 기록물, '임헌회'라는 인물 등 세 가지 개체의 고유한 이름을 매개로 서로 연결될 수 있었다. 두 번째 절에서는 '김현'이라는 인물과 '인문정보학의 모색'이라는 저서가 같은 방식으로 연결되는 구체적인 사례를 보였다. 그리고 세 번째 절에서는 이러한 종류의 개체에 더하여 장소, 연대, 개념어 등에도 고유한 이름을 부여함으로써 문화유산 안내문을 기계가독적인 텍스트로 전환해 보았다. 이 과정에서 다루어 본 몇 가지 범주와 그것에 속하는 개체들, 그리고 그 개체들 사이의 상관관계는 인문학 분야의 지식, 그 중에서도 한국의 전통문화 관련 지식의 정보화를 위한 데이터 모델의 핵심적인 구성 요소이다. 다시 말해, 그와 같은 범주에 속하는 중요 개체들이 표준적인 이름을 갖도록 하는 '한국 문화유산 온톨로지'가 누군가에 의해 만들어지고 많은 사람들이 이를 공유한다면, 우리나라의 디지털 정보세계 전체에서 통용되는 '문화유산 시맨틱 웹'이 만들어질 수 있다는 것이다. 그 일은 누가 해야 할까?

인문지식 시맨틱 웹의 구현이 교육이나 실험의 차원이 아니라 실효성 있는 지식 소통의 기능을 하기 위해서는 이 분야의 학술 진흥에 관한 공익적 책임

을 지고 있는 기관들이 그 역할을 담당해야 한다. 시맨틱 웹의 노드와 링크가 될 요소들을 찾아내고 그것에 고유한 이름을 부여하는 일은 그 분야의 대상 지식을 분석할 수 있는 전문가라면 누구나 할 수 있다. 하지만, 그 일을 연구자들이 저마다 개인적인 차원에서 하려 한다면, 표준을 만드는 일은 점점 더 어려워질 것이다. 우리나라의 인문 분야 학술 연구기관들은 이미 인문지식 시맨틱 웹의 노드로 활용될 수 있는 문맥 요소 데이터를 작지 않은 규모의 데이터베이스로 구축해 놓고 있다.[1] 이 데이터를 기반으로 기본적인 인문지식 시맨틱 웹 온톨로지를 구축하는 것은 그 데이터를 축적해 온 공공기관들이 담당해야 할 과업이다.

한국사, 문화유산, 고전문학 등 주제 영역별로 통용될 수 있는 온톨로지가 만들어지고 누구나 이를 참조할 수 있게 되면 개인 연구자의 연구 활동이나 대학의 인문학 교육 과정에서 생산되는 다양한 데이터가 시맨틱 웹의 세계에 진입할 수 있는 길이 열리게 된다. 연구자들이 논문을 발표하거나, 교사가 수업 자료를 만들 때, 학생들이 리포트를 작성할 때, 심지어는 학계의 일원이 아닌 누군가가 문화유적지를 돌아보며 기행문을 남길 때에도 자신이 지금 웹에 남기는 흔적이 #어떤 인물, #어떤 사건, #어느 장소, #어느 문헌, #어떤 개념과 관계가 있는지, 온톨로지 참조를 통해 명시적으로 밝힐 수 있게 되는 것이다.

이렇게 만들어진 시맨틱 웹 데이터는 월드와이드웹 상에서 어떻게 활용될까? 앞의 스파클(SPARQL) 사용 예시에서 보았듯이, 웹상에서 단어를 검색하는 데 머무는 것이 아니라 사실과 사실간의 관계를 추론하여 새로운 지식을 얻는 일이 가능해질 것이다.

시맨틱 웹 데이터를 집적하여 사실의 추론이 가능한 데이터 검색 서비스를 제공하는 역할은 미래의 '시맨틱 웹 포털'이 담당할 역할이다. 시맨틱 웹 포털은 국가가 운영하는 공공 데이터 포털[2]이나 민간 상업 포털이 기능 확장을 통

---

1) 예를 들어 한국학 분야의 정부 출연 연구기관인 한국학중앙연구원이 운영하고 있는 다음과 같은 데이터베이스로부터 '한국학 시맨틱 웹'의 기초자원이 될 수 있는 데이터(고유한 식별자를 가진 개체 목록 및 개체 사이의 관계를 기술하는 RDF 데이터)를 생산할 수 있다.

보유 자원: 데이터베이스	산출 가능한 시맨틱 웹 자원
역대인물정보	우리나라 전근대 인물의 식별자 / 혈연적·사회적 관계 정보
한국사연표	한국사의 주요 사건 식별자 / 사건의 전개와 관련된 시간 정보
한국향토문화전자대전	인물, 작품, 문화유산, 역사적 사건과 지리적 공간 사이의 관련성 정보
한국민족문화대백과사전	한국역사상의 주요 인물, 문화유산, 기록유산, 개념·용어 식별자 / 역사적 인물과 역사적 시공간의 관련성 정보

2) http://www.data.go.kr

해 구현할 수도 있고, 특정 지식 영역의 커뮤니티에서 온톨로지 기반 데이터베이스에 관심 주제의 시맨틱 데이터를 집적하는 방법으로 구현할 수도 있다.

한국 문화유산 시맨틱 웹의 구현을 위한 협력 구도

시맨틱 웹 포털의 기술적인 환경을 구축하는 일은 정보기술 전문가들이 주도적인 역할을 담당할 것이다. 이 영역의 일은 기술 표준이 확립되어 가고 있고, 또 성공적인 선행 모델이 있기 때문에 언제든지 필요하면 만들어질 수 있는 것이라고 해도 무방하다. 반면, 인문학적 내용을 담은 시맨틱 웹 데이터의 생산은 인문학을 중심에 두는 디지털 인문학의 영역에서 만들어져야 한다. 디지털 세계에서 다양한 인문지식 자원들이 의미의 맥락을 좇아 무한히 이어지는, 그 문맥 속에서 새로운 사실을 발견하기도 하고, 알려진 사실의 신뢰성을 검증할 수도 있는 인문지식 시맨틱 웹의 구현은 그 네트워크의 노드와 링크 하나하나를 정밀하게 선택하고 다듬는 노력 위에서 만들어진다. 디지털 인문학의 역할은 인문지식 데이터에 대해 바로 그와 같은 실천적 노력을 투입하는 것이다.