제2장 디지털 큐레이션을 가능하게 하는 도구와 기술

이 장의 키워드:
데이터베이스, 시맨틱 데이터 모델링, 온톨로지, 시맨틱 데이터 아카이브

어떤 주제의 디지털 큐레이션을 통해 유의미한 지식의 디지털 아카이브를 생산해 내는 것은 그 분야의 인문지식에 대한 탐구 노력과 함께 디지털 데이터를 조작하고 체계적으로 집성할 수 있는 방법론에 대한 이해, 그리고 그것을 실천할 수 있는 기술적 능력이 있어야 가능한 일이다. 제2장에서는 인문지식의 디지털 큐레이션을 가능하게 할 기술적 요건들에 대해 설명한다. 그것은 첫째, 큐레이션의 내용을 저장하여 아카이브로 쓰일 수 있게 하는 소프트웨어인 '데이터베이스'에 대한 이해, 둘째, 큐레이션의 대상이 되는 지식요소들과 그것들 사이의 관계를 개념적으로 설계하는 '시맨틱 모델링', 셋째, 시맨틱 모델링을 통해 만들어지는 가상 세계의 설계도인 온톨로지, 그리고 넷째, 온톨로지의 설계 내용에 따라 대상 세계의 사실과 문맥이 디지털 세계에서 재현될 수 있게 하는 '시맨틱 데이터 아카이브'이다.

이 장에서 언급할 기술적 방법론은 기본적으로 '정보 기술(Information Technology)' 분야에서 실제로 운용하는 데이터 처리(Data Processing) 기술이다. 하지만 설명의 주안점은 그 기술 자체가 아니라 그 기술을 인문지식 큐레이션의 도구로 사용하는 방법에 둘 것이다. 경우에 따라서는 기존의 정보 기술을 그대로 준용하기보다는 '인문 지식 큐레이션'의 목적에 맞추어 변용하는 방법도 제시될 것이다.

1. 데이터베이스

전통적인 인문학 공부의 '읽기'에 상응하는 디지털 세계의 행위는 내가 얻고자 하는 지식의 원천 소스가 될 수 있는 데이터를 조사하고 수집하는 것이고, '쓰기'에 상응하는 행위는 내가 의미있다고 판단한 데이터들을 가지고 그 '의미'의 문맥이 드러나도록 재구성하는 것이다. 이때 데이터의 수집이 체계적으로 이루어져서 필요에 따라 잘 쓰일 수 있도록 해 주는 기술적인 도구가 '데이터베이스'이다. 인문학 공부의 관점에서 이 데이터베이스의 기능을 설명한다면 그것은 첫째, '읽기'와 '쓰기'를 연결해 주는 '학습장'의 역할이고, 둘째, 학습장에 쓰여진 것을 체계적으로 정리하여 세상에 보일 수 있게 하는 '저술'의 도구이다. 데이터베이스에 기록된 정보는 기록자뿐 아니라 다른 이용자들에 의해 참조되기도 하고, 그들에 의해 다양한 디지털 저작물(새로운 디지털 큐레이션이나 스토리텔링)의 요소로 활용될 수도 있다. 데이터의 공유를 가능하게 하고, 그렇기 때문에 소통과 협업에 의한 인문지식의 확대재생산을 촉진하는 도구라는 점에서 데이터베이스는 디지털 인문학의 핵심적인 기술요소이다.

사실상 '데이터베이스'는 수많은 종류의 소프트웨어 기술 중에서 가장 고전적인 것이라고 할 만하고, 그만큼 우리 사회의 각 분야에서 널리 활용되고 있는 것이기도 하다. 금융, 산업, 행정, 의료, 교육 등 데이터 처리를 수반하는 모든 곳에서 데이터베이스가 쓰이고 있다. 온라인에서 고객을 상대하는 수많은 인터넷 사이트들도 거의 예외 없이 데이터베이스 위에서 작동한다고 할 수 있다.

인문학 공부도 데이터를 수집하고, 정리하고, 활용하는 일이므로 인문학 연구자와 학생들의 데이터베이스 의존도 역시 작지 않다. 그들이 도서관, 기록관, 박물관 등의 디지털 아카이브에서 공부와 연구를 위한 자료를 찾는 일이 모두 데이터베이스 이용이다. 그렇지만 오늘날 인문학도들에게 있어서 데이터베이스는 남이 저장해 놓은 '읽을거리'를 편리하게 찾아주는 도구일 뿐, 나의 지식을 체계적으로 확장시켜 가는 공부의 도구가 되지 못하는 것도 사실이다. 오늘날의 인문학도 중에 남이 만들어 놓은 데이터베이스를 이용해 보지 않은 사람은 없을 것이다. 그런데 그들 중에 자기가 만든 데이터를 데이터베이스에 기록해서 전자적으로 활용할 수 있는 사람은 거의 찾아보기 힘들다.[1] 하지만 앞에서 언급했듯이, 디지털 인문학 교육의 목표는 '디지털로 읽기'와 '디지털로 쓰기' 능력을 균형적으로 신장시키는 것이다. 이 말은 디지털 원어민 세대의 인문학도들이 데이터베이스를 '읽기'의 도구로뿐 아니라 '쓰기'의 도구로 운용할 수 있게 해야 한다는 말과도 통한다. 더구나 지금 우리가 이야기하고 있는 '디지털 큐레이션' 활동은 데이터베이스 관련 기술의 도움 없이는 아예 성립할 수 없는 일이다.

데이터베이스를 사용한다는 것은 구체적으로 무엇을 어떻게 하는 것인가? 데이터베이스에 대한 기술적인 정의는 "컴퓨터상의 여러 프로그램들이 사용할 수 있도록 체계적으로 편성•저장된 데이터의 집합"이다. 이 데이터베이스가 인간의 다양한 지적 활동을 도울 수 있는 것은 단지 데이터를 '저장'하는 기능 때문이 아니라, 그 저장된 데이터를 다양한 관점과 시각에서 여러 가지 형태로 '편성'할 수 있게 하기 때문이다. 이 '편성'과 '저장'의 기능은 '데이터베이스 관리 시스템'(DBMS)이라고 부르는 소프트웨어를 통해 수행된다. 따라서 데이터베이스를 사용한다는 것은 DBMS 소프웨어를 사용해서 데이터를 저장하고, 정리하고, 조회하며, 갱신하는 일을 하는 것이라고 할 수 있다.[2]

---

1) 데이터베이스가 이처럼 중요한데, 왜 우리는 학생들에게 이것을 적극적으로 가르치지 않는가? 아마도 이것이 기존의 인문학 교과목처럼 책과 강의로만 교육할 수 있는 일이라면 일찍이 시행되었을 수도 있었을 것이다. 하지만 데이터베이스 운용에 기반한 디지털 인문학 교육은 그것을 위한 기술적 환경이 사전에 마련되어 있어야 하는데, 그것은 인문학 교수 개인의 노력만으로는 이루기 어려운 것이 사실이다. 다수의 학생들이 함께 사용할 수 있는 서버가 준비되어야 하고, 거기에 '데이터베이스 관리 시스템'이라는 소프트웨어가 설치되어야 하며, 학생들에게 이 컴퓨팅 환경을 이용할 수 있는 계정이 발급되어야 한다. 물론 지금 우리나라에서는 교육 환경의 디지털 전환에 대한 논의가 활발하고, 여러 대학에서 디지털 인문학 교과과정 도입에 관심을 가지고 있으니 상황은 개선되리라 기대한다. 지금 이 시점에 무엇보다 필요한 것은 디지털 인문학 교육의 필요성에 공감하는 교육자들이 '인문학' 본연의 교육과 연구를 위해서는 '프로그래밍 언어'나 '통계•분석 소프트웨어'의 사용법을 가르치는 것보다 데이터베이스 기반의 협업적 공부 환경을 마련하는 것이 중요함을 인식하고, 이에 대한 이해를 높이는 일이다. 디지털 큐레이션 교육의 환경을 마련하기 위한 현실적인 방법에 대한 제안을 이 책은 제O장 '디지털 큐레이션 교육 환경: 정보 시스템의 구축과 운용'에서 제시하고자 한다.

내가 학생들에게 여러해 동안 데이터베이스를 가르쳐 본 경험으로 이야기하자면, DBMS의 사용법 자체를 배우는 것은 누구에게도 어려운 일이 아니며 그렇게 많은 시간을 필요로 하지 않는다. 데이터베이스 수업에서 학생들이 더 어려워했고, 더 많은 시간을 필요로 했던 일은 데이터베이스에 담고자 하는 인문지식의 세계에서 유의미한 이야기거리를 찾아 체계적으로 정리하는 일, 즉 인문학 공부였다. 데이터베이스를 배우는 데 웬 인문학 공부냐고 물을 수도 있겠지만, 데이터베이스에 데이터를 기록하는 것은 논리적이 글쓰기와 다름없는 것임을 상기하면 '인문학 주제의 데이터베이스 구축'은 곧 인문학 공부일 수밖에 없음을 이해할 것이다. 데이터베이스를 이용한 디지털 큐레이션은 그 인문학 공부를 더 풍부하고 재미있게 만드는 역할을 한다. 그래서 그 공부에 빠져드는 동안 데이터베이스 관리 도구인 DBMS에 대해서도 그것을 어떤 용도로 어떻게 써야 할지 알게 되고, 그 소프트웨어를 기술적으로 다루는 능력도 크게 향상된다. 이것이 디지털 인문학 교육의 일환으로 수행하는 데이터베이스 수업의 내용이다.

2. 시맨틱 데이터 모델링

디지털 인문학을 위한 데이터베이스 교육에서 특별히 중점을 두는 것은 첫째, 우리가 다루고자 하는 인문지식의 이야기(또는 이론이라고 하는 것) 속에서 그 이야기는 '무엇'에 관한 것이고 그 '무엇'은 어떠한 구성 요소를 가지고 있는지를 파악하는 것, 둘째, 그 요소들에 대해 우리가 파악하고 이해한 것을 체계적이고 명시적인 정보로 표현하는 방법을 강구하는 것이다. 데이터베이스 이론에서는 이런 일을 '데이터 모델링'이라고 한다. 나의 관점에서는 '인문학 공부를 제대로 하는' 방법이다.

디지털 큐레이션에 관한 이 책에서 데이터베이스의 중요성을 강조하는 첫 번째 이유는 데이터베이스의 설계 과정에서 수행하는 '데이터 모델링'의 방법이 바로 큐레이션의 대상과 범위를 명확하게 정의하는 방법으로 쓰일 수 있기 때문이다. 데이터베이스가 큐레이션의 결과를 저장하여 디지털 아카이브로 남게 하는 기술적 도구라는 것도 중요하지만, 그것은 오히려 두 번째 이유라고 해도 무방하다. 학습과 교육 차원에서의 디지털 큐레이션은 본격적인 데이베이스 관리 시스템(DBMS)의 도입 없이도 시도해 볼 수 있다. 그러나 현실 세계의 어느 것을 디지털 세계에서 재현할 것인지에 대한 구체적인 구상 없이는 어떠한 수준의 큐레이션 활동도 불가능하다.

'데이터 모델링(Data Modeling)'이라고 부르는 과업은 현실 세계의 일이 디지털 데이터로 표현될 수 있도록 그 일의 구성 요소와 그 요소들 사이의 관계를 정의하는 것

2) DBMS (Database Management System): 데이터베이스 관리 시스템이라고 한다. 데이터베이스를 새로 만들거나 기존의 데이터베이스를 갱신하고 데이터베이스 속의 정보를 다양한 기준과 형식으로 조회할 수 있는 기능을 제공하는 소프트웨어이다. 여러 종류의 DBMS 소프트웨어가 있는데, 그 가운데 많이 사용되고 있는 것은 Oracle, MySQL, MS SQL Server 등의 관계형 데이터베이스 관리시스템, MongoDB 와 같은 문서지향형 데이터베이스 관리 시스템, Neo4J 와 같은 네트워크형 데이터베이스 관리 시스템을 들 수 있다. DBMS의 운용에 관해서는 제O장 '디지털 큐레이션을 위한 데이터베이스 운용'에서 설명한다.

이다. 이 때 모델링(modeling)이라는 말은 '추상화(抽象化, abstraction)'의 뜻인데, 복잡다단한 대상 세계를 그대로 베끼는 것이 아니라 그 일의 핵심적인 요소와 기능을 파악하여 디지털로 재현한다는 전제에서, 그것을 데이터로 담아내는 데 적합한 틀을 디자인하는 것이다.

디지털 큐레이션 활동을 설계하기 위해서는 '데이터 모델링(Data Modeling)'과 함께 '시맨틱 모델링(Semantic Modeling)'의 개념도 이해할 필요가 있다. '시맨틱 모델링'은 대상 세계의 구성 요소들 사이의 다양한 관계에 대해서 그 성격과 내용이 명시적으로 표시될 수 있도록 관계성 정의의 틀을 설계하는 것을 말한다. '시맨틱 모델링'은 관계성의 데이터화에 역점을 두는 '데이터 모델링'의 한 방법이라고도 할 수 있겠지만, 일반적으로 '데이터 모델링'이라고 할 때에는 정보화의 대상 세계에 대한 정보를 담기에 적합하도록 데이터베이스의 구조를 설계하는 일을 뜻하는 반면, '시맨틱 모델링'은 정보화의 대상 세계가 어떻게 존재하는지 드러내기 위해 그 구성 요소 사이의 의미론적 맥락(Semantic Context)을 정의하는 일을 뜻한다. 양자의 주안점을 좇아서 이야기한다면, 우리가 하려는 일 가운데 '큐레이션'을 위해서는 '시맨틱 모델링'이 필요하고, 그 큐레이션의 내용물이 아카이브에 남게 하기 위해서는 '데이터 모델링'을 해야 한다고 할 수 있다. 그런데 우리의 디지털 큐레이션과 디지털 아카이브는 별개의 독립된 세계가 아니라, 큐레이션의 결과물이 아카이브에 남고, 그 아카이브가 다시 새로운 큐레이션의 환경이 되는 하나의 연속적 융합체이다. 이 융합적인 일을 위한 디지털 모델 설계를 설명하기 위해 '시맨틱 모델링'과 '데이터 모델링'의 개념을 하나로 접목한 '시맨틱 데이터 모델링'이라는 개념을 사용하도록 하겠다.

'시맨틱 데이터 모델링'이란 데이터의 성격과 용도에 적합한 데이터 저장소(데이터베이스)의 구조뿐 아니라, 그 데이터의 의미와 맥락을 명시적으로 알 수 있도록 하는 데이터 구성 방법을 정의하는 것이다. 독자의 이해를 돕기 위해 예시로서 이를 설명하겠다.

※ 인문지식 큐레이션의 과제 예시

조선시대의 왕실문화를 이해하기 위해 17세기에서 20세기초까지 궁정에서 열렸던 왕실 잔치에 대해 조사하고, 그 조사연구에서 파악한 정보를 데이터로 기록하고자 한다.

이 조사연구는 실제로는 매우 다양하고 많은 사건과 인물, 사물을 대상으로 이루어지겠지만, 여기서는 모델링의 예시를 보이는 것이므로 큐레이션의 요소를 (1) 사건(왕실잔치), (2) 장소(잔치가 일어난 곳), (3) 기록물(잔치에 관한 기록) 등 세 가지 영역에 한정하겠다.

먼저 데이터 모델링의 예시이다. 데이터 모델링은 앞에서 설명했듯이 데이터를 저장할 데이터베이스 구조의 정의가 주안점이다. 일반적인 관계형 데이터베이스에 적용할 수 있는 데이터 모델링의 개요는 다음과 같다.

※ 데이터 모델링에 의해 만들어진 엔티티별 데이터

§ 사건 테이블

id	유형	발생연도	발생장소
무신년진찬	사건>궁중연향	1848년	창경궁 통명전
정해년진찬	사건>궁중연향	1887년	경복궁 만경전
임인년진연	사건>궁중연향	1902년	덕구궁 중화전
.....			

§ 장소 테이블

id	유형	위도	경도
창경궁 통명전	장소>궁궐전각	37.5797	126.994
경복궁 만경전	장소>궁궐전각	37.5811	126.978
덕수궁 중화전	장소>궁궐전각	37.5652	126.975
.....			

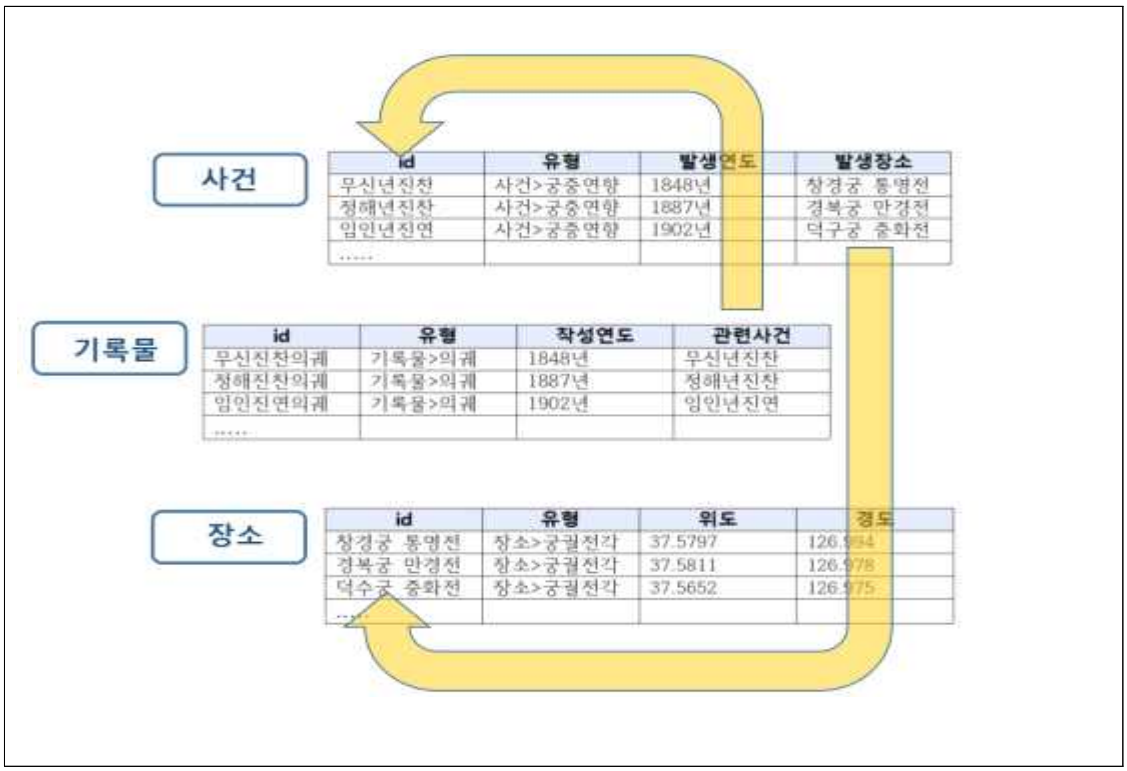
§ 기록물 테이블

id	유형	작성연도	관련사건
무신진찬의궤	기록물>의궤	1848년	무신년진찬
정해진찬의궤	기록물>의궤	1887년	정해년진찬
임인진연의궤	기록물>의궤	1902년	임인년진연
.....			

위와 같은 데이터 모델링은 대상 세계(조선왕실의 왕실잔치)를 디지털 세계에서 재현할 수 있게 하는 3 가지 유형(사건, 장소, 기록물)의 요소를 유형에 따라 어떠한 구조의 데이터로 저장할 것인지를 정의한 것이다. 데이터 유형에 따라 다른 구조(이것을 스키마라고 한다)를 갖는 데이터 세트를 엔티티(entity)라고 하며 한 종류의 엔티티를 담는 데이터의 틀을 테이블(table)이라고 부른다. 이 2차원 테이블 각 행(row)은 해당 엔티티에 속하는 개체이다. 이것을 레코드(record)라고 부른다. 이 2차원 테이블 각 열(column)은 이 엔티티에 속하는 각각의 개체들의 부수적인 특성이다. 이것을 속성(attribute)이라고 부른다.[3] 그런데 각 엔티티의 속성을 자세히 보면, 그 중에는 다른

엔티티에 속하는 개체와의 관계도 정보화 되어 있음을 알 수 있다. 사건 엔티티의 개체는 '발생장소' 속성을 매개로 장소 엔티티의 개체와 연결될 수 있고, 기록물 엔티티의 개체는 '관련사건' 속성을 매개로 사건 엔티티와 연결될 수 있다.

※ 관계형 데이터베이스에서의 엔티티간 연결



여기에서 보듯이 관계형 데이터베이스도 개체 사이의 관계를 정보화하는 방법이 있지만[4] 그것은 다음에 설명할 시맨틱 모델링에 비하면 매우 제한적이고 유연하지 못하다.

시맨틱 모델링도 데이터 모델링과 마찬가지로 사실 세계에서 포착한 정보를 데이터로 기록하는 한 방법이지만, 대상 세계 속 정보 요소들 사이의 관계를 파악하고 표현하는 데 역점을 두는 차이가 있다.

앞에서 우리는 큐레이션 과제 속에 세 가지 범주로 묶을 수 있는 요소들이 있음을 살펴 보았다. 즉 (1) 사건(왕실잔치), (2) 장소(잔치가 일어난 곳), (3) 기록물(잔치에

3) 이 간략한 데이터 모델링 예시에서는 각 엔티티(entity)의 속성(attribute) 이름만 보이고 있으나 실제로 데이터베이스 상에서 새로운 엔티티를 만들 때에는 각각의 속성에 대해 데이터 타입(문자열, 날짜, 정수, 실수, text, xml 등 여러 가지 형태의 데이터 중 어디에 해당하는지를 알리는 것)과 크기 등 데이터 저장을 위해 필요한 요건을 함께 지정하게 된다.

4) 관계형 데이터베이스에서 개체들 사이의 관계를 정보하는 방법은 이처럼 어느 개체의 식별자(identifier, primary key)를 그와 관련 있는 다른 개체가 자신의 속성 정보의 하나로 쓰는 것이다. 하나의 개체가 자기의 엔티티에 안에서 갖는 식별자를 주 키(Primary key)라고 하고, 그것이 다른 엔티티에 속하는 개체의 속성으로 쓰였을 때에는 그것을 외래 키(Foreign key)라고 한다.

관한 기록) 세 범주이다. 그러면 이 세 가지 범주에 속에는 정보화의 대상 요소들은 서로에 대해 어떤 관계를 맺고 있다고 보아야 할까? 우리가 이해하는 대상의 성격에 따라 자연스럽게 도출되는 대상 요소들 사이의 관계성은 다음과 같을 것이다.

이렇게 대상 세계의 요소들을 어떤 범주로 묶을 수 있고, 그 요소들 사이의 관계는 어떻게 표현하면 좋을지 판단하는 것이 '시맨틱 모델링'이다. 한 가지 더 부수적으로 해야 할 일은 각각의 범주별로 그 속의 요소들이 가질 수 있는 부가적인 속성 데이터는 무엇인지 정의하는 것이다. 다음은 앞에서 데이터 모델링의 예시로 보인 것과 동일한 사실을 시맨틱 모델링의 방법으로 데이터화한 것이다.

※ 시맨틱 모델링

§ 클래스 정의

클래스 이름	설명
사건	사건, 행사 등 시간성을 갖는 개체의 범주,
장소	지명, 건물명 등 공간성을 갖는 개체의 범주
기록물	문헌, 금석문 등 지식의 원천 자료가 되는 기록물
개념	정보화 대상을 설명하는 데 필요한 용어와 개념
.....	

§ 속성(Datatype Property) 정의

속성 이름	적용 클래스	데이터 형식	설명
발생연도	사건	문자열	사건의 발생 연도
작성연도	기록물	문자열	기록물의 작성 연도
위도	장소	실수	지리좌표 위도 값
경도	장소	실수	지리좌표 경도 값
....			

§ 관계성(Object Property) 정의

관계성 이름	적용 클래스(소스)	적용 클래스(타겟)	설명
documents	기록물	사건	A는 B를 문서화하다
isHeldAt	사건	장소	A는 B에서 열리다
type	ALL	개념	A는 B의 일종이다
.....			

이 예시에서 다시 확인할 수 있듯이, '시맨틱 모델링'의 이름으로 하는 일은 대체로 1) 대상 세계의 사실들을 소속시킬 범주(클래스, Class)를 정하고, 2) 각각의 범주에 속하는 개체(Individual Object, Node)들이 가질 수 있는 속성(Attribute)은 무엇인지, 그리고 3) 개체들이 서로에 대해 갖는 관계(Relation, Link)를 어떻게 기술할 것인지를 정하는 것이다.

시맨틱 모델링에 의해 데이터화의 틀을 설계하면, 그 틀에 따라 시맨틱 데이터를 만들 수 있게 된다. 이것은 대상 세계에서 의미있는 구성 요소들을 찾아 데이터 세계의 개체(Individual Object)로 정의하고 각 개체의 속성 및 개체간 관계성을 기술하는 일이다.

※ 시맨틱 모델링의 의해 만들어진 시맨틱 데이터

§ 개체(노드)

id	클래스	레이블
궁중연향	개념	궁중연향(宮中宴享)
무신년진찬	사건	무신년 진찬(戊申年進饌, 1848)
정해년진찬	사건	정해년 진찬(丁亥年進饌, 1887)
임인년진연	사건	임인년 진연(壬寅年進宴, 1902)
창경궁_통명전	장소	창경궁 통명전(通明殿)
경복궁_만경전	장소	경복궁 만경전(萬慶殿)
덕수궁_중화전	장소	덕수궁 중화전(中和殿)
무신진찬의궤	기록물	무신진찬의궤(戊申年進饌儀軌)
정해진찬의궤	기록물	정해진찬의궤(丁亥年進饌儀軌)
임인진연의궤	기록물	임인진연의궤(壬寅年進宴儀軌)
....		

§ 개체별 속성

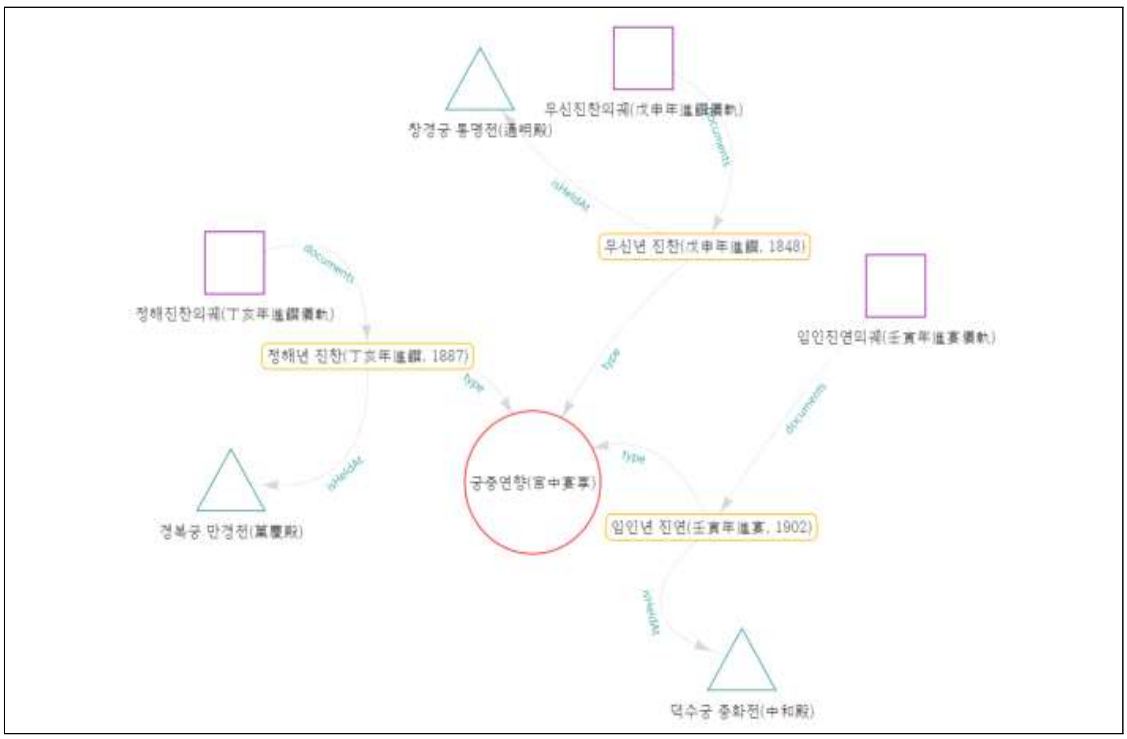
개체 id	속성 이름	속성 값
무신년진찬	발생연도	1848년
정해년진찬	발생연도	1887년
임인년진연	발생연도	1902년
창경궁 통명전	위도	37.5797
창경궁 통명전	경도	126.994
경복궁 만경전	위도	37.5811
경복궁 만경전	경도	126.978
덕수궁 중화전	위도	37.5652
덕수궁 중화전	경도	126.975
무신진찬의궤	작성연도	1848년
정해진찬의궤	작성연도	1887년
임인진연의궤	작성연도	1902년
....		

§ 개체간 관계

소스	타겟	관계어
무신년진찬	궁중연향	type
정해년진찬	궁중연향	type
임인년진연	궁중연향	type
무신년진찬	창경궁_통명전	isHeldAt
정해년진찬	경복궁_만경전	isHeldAt
임인년진연	덕수궁_중화전	isHeldAt
무신진찬의궤	무신년진찬	documents
정해진찬의궤	정해년진찬	documents
임인진연의궤	임인년진연	documents
....		

앞에서 보았던 관계형 데이터베이스의 데이터 모델링과 시맨틱 모델링 사이에서 가장 두드러지게 보이는 차이점은 정보 요소들 사이의 관계를 데이터화하는 방식이다. 전자는 각기 다른 엔티티가 공통의 속성을 갖게 하는 방법, 즉 개체들의 집합인 엔티티와 엔티티 사이의 관계를 정의하는 방식을 주로 이용하지만, 후자는 각각의 개체가 직접 다른 개체와 다양한 관계를 맺을 수 있도록 한다. 시맨틱 데이터로 기술된 개체간 관계는 다음의 예시와 같이 네트워크의 형태로 표현될 수 있다.

※ 네트워크 그래프로 표현한 시맨틱 데이터



시맨틱 모델링의 결과를 보면 대상 세계의 구성 요소들이 서로에 대해 어떤 관계를 맺고 있는지 잘 파악할 수 있다. 시맨틱 모델링이 주안점을 두는 부분이 대상 세계의 의미와 문맥을 파악하는 것이기 때문에 잘 디자인된 모델에 따라 생성된 시맨틱 데이터는 정보화의 대상 세계에서 일어났던 일을 디지털 세계에서 유의미하게 표현할 수 있는 것이다.

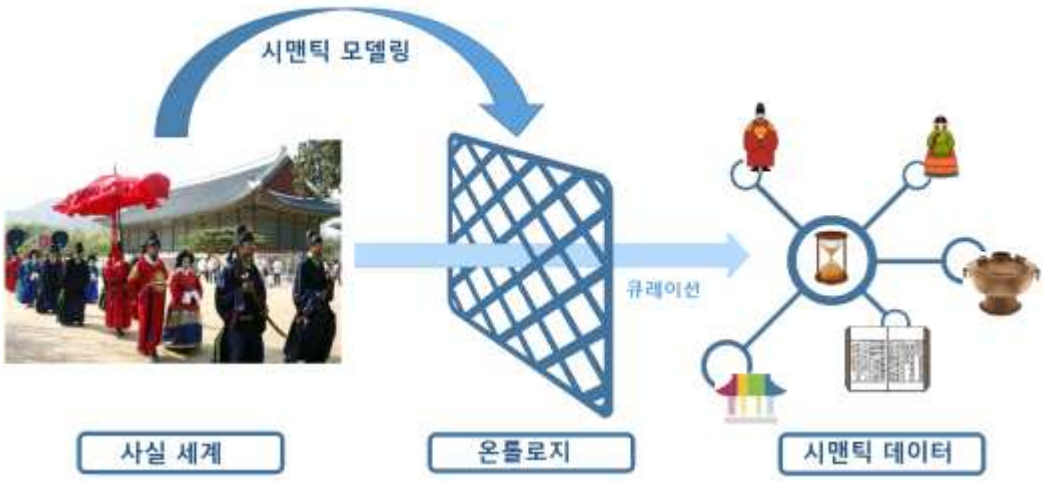
이야기가 될 수 있는 '사실과 문맥'을 데이터화할 수 있다는 점에서 '시맨틱' 모델링은 일반적인 데이터 모델링보다 '디지털 큐레이션'의 방법론으로 더 적합하다. 그러나 '시맨틱 모델링'의 방법만으로 '디지털 큐레이션'의 모든 프로세스를 다 진행시킬 수 있는 것은 아니다. 거듭 말하지만 큐레이션의 결과가 (공유와 재이용이 가능한) 디지털 아카이브로 남기 위해서는 데이터베이스로 만들어져야 한다. 의미를 좇아서 파악한 대상 세계의 '사실과 문맥' 데이터가 장점을 유지하면서 데이터베이스에 저장할 수 있는 형식의 데이터로 전환되어야 할 필요가 있는 것이다.

'시맨틱 데이터 모델링'은 지금까지 차이점을 비교한 '시맨틱 모델링'과 '데이터 모델링'의 함의를 모두 포함하는 개념이다. 앞으로 이 책에서 소개할 여러 가지 유형의 디지털 큐레이션은, 디지털 아카이브 구축까지 포함하는 본격적인 큐레이션일 경우 모두 다 '시맨틱 데이터 모델링'의 과업을 수반한다고 보아도 무방하다.

3. 온톨로지

어떤 대상 세계에 대한 시맨틱 모델링의 결과, 다시 말해서 대상 세계의 구성요소와 그것들 사이의 문맥이 디지털 세계에서 일정한 형식으로 표현될 수 있도록 만든 틀을 '온톨로지(ontology)'라고 한다.[5] 이 온톨로지는 물리적인 현실 세계(또는 작품 속에 있는 가상의 세계)를 디지털 정보의 세계로 옮겨 짓는 설계도 역할을 하게 된다. 그 설계도에 따라 만들어지는 디지털 데이터는 현실세계의 일들이 가졌던 의미를 데이터의 관계로써 표현하기 때문에 '시맨틱(semantic, 의미론적, 의미 기반의)'이라는 수식어를 붙여서 '시맨틱 데이터'라고 한다. 아래의 그림은 (1)시맨틱 모델링, (2) 시맨틱 모델링의 명세서인 온톨로지, (3) 온톨로지에 따라 만들어진 시맨틱 데이터의 관계를 도시한 것이다.

5) '온톨로지'란 정보화의 대상이 되는 세계를 전자적으로 표현할 수 있도록 구성한 데이터 기술 체계이다. 원래 온톨로지라는 말은 철학에서 '존재론'이라고 번역되는 용어로서 '존재에 대한 이해를 추구하는 학문'의 의미를 갖는 말이었다. 그러한 용어가 정보과학 분야에서 중요한 개념으로 등장하게 된 것은 인간이 세계를 이해하는 틀과 컴퓨터가 정보화 대상(콘텐츠)을 이해하는 틀 사이에 유사성이 있다고 보았기 때문이다.



시맨틱 모델링 / 온톨로지 / 시맨틱 데이터

특정 대상 세계에 대한 시맨틱 모델링의 결과, 즉 그 세계에 대한 온톨로지를 어떤 형식으로 기술하는 것이 좋을까에 대해서는 여러 가지 관점과 대안, 이견이 있을 수 있다. 인문학이나 역사학의 관점에서 새로운 온톨로지의 형식을 만들어내는 것도 불가능한 것은 아니겠으나, 정보과학자들에 의해서 만들어진 어떤 형식이 오늘날 여러 분야에서 널리 쓰이고 있다면, 그것을 도입해서 적용해 보는 것이 더 현명한 방법일 수 있다. 그것은 누구에게도 완전히 만족스럽지 않지만, 일반적인 활용 가치를 인정받은 기술이기 때문이다.

정보기술 분야에서 말하는 '온톨로지(ontology)'에 대한 가장 일반적인 정의는 그루버(Gruber, Thomas. 1959~)가 말한 '명시적 명세화의 방법에 의한 개념화'(explicit specification of a conceptualization)이다.[6] 여기서 '개념화'(conceptualization)라는 것은 정보화하고자 하는 대상 세계를 일정한 체계 속에서 파악하는 것, 예를 들면 그 세계에 무엇이 있고, 그것은 어떤 속성을 품고 있으며, 그것들 사이의 관계는 무엇인가 하는 일정한 질문의 틀 속에서 대상 세계를 이해하는 방식이라고 할 수 있다. '명세화'(specification)란 대상 세계에 존재하는 개체, 속성, 관계 등을 일목요연한 목록으로 정리하는 것, 그리고 '명시적'(explicit)이라는 것은 그 정리된 목록을 사람뿐 아니라 '컴퓨터가 읽을 수 있도록'(machine readable) 한다는 것이다. 이런 취지에 따라 대상 세계의 구성 요소가 되는 모든 개별적인 개체(individuals)들을 '클래스'(class)로 범주화하고, 각각의 클래스에 속하는 개체들이 공통의 '속성'(attribute, datatype property)을 갖도록 하고, 그 개체들이 다른 개체들과 맺는 '관계'(relation, object property)를 명시적으로 기술하는 것이 가장 일반적인 온톨로지 설계 방법이라고 할 수 있다.

---

6) Gruber, Thomas Robert. 'A Translation Approach to Portable Ontology Specifications'. *Knowledge Systems Laboratory Technical Report KSL 92-71.* Stanford University. 1992

※ 온톨로지 구성 요소

온톨로지 구성 요소	용도[3]	Web Ontology Language (OWL)
Class, 클래스	공동의 속성을 가진 개체들을 묶는 범주	owl:Class
Individual Object, 개체	클래스에 속하는 개체	owl:NamedIndividual
Relation, 관계성	(같거나 다른 클래스에 속하는) 개체들 사이의 관계	owl:ObjectProperty
Attribute, 속성	개체가 속성으로 갖는 데이터 값	owl:DatatypeProperty
Attribute of Relationship, 관계 속성	개체 사이의 관계에 부수되는 속성	N/A in OWL

① 클래스 (Class)

온톨로지 설계에서 '클래스'라고 하는 것은 대상 세계의 다양한 구성 요소들을 유형별로 묶는 범주이다. 클래스 설계는 대상 세계를 무차별적인 사실이나 사물의 나열로 보지 않고, 일정한 체계 속에서 이해하려는 노력이라고 할 수 있다. 따라서 어떠한 클래스를 정하느냐 하는 것은 연구자가 대상을 어떠한 시각에서 보고자 하느냐에 따라 달라진다.

앞에서 시맨틱 모델링의 예시로 '조선의 왕실 잔치'를 데이터화한 사례를 보였는데, 여기서 '조선의 왕실 잔치'라는 주제가 정보화의 대상으로 정한 세계이다. 이 세계를 구성하는 수많은 요소들이 있을텐데 예시에는 그것을 '사건', '장소', '기록물' 등 3가지 범주로 파악될 수 있다고 보았다.

'클래스'를 설정하는 이유는 두 가지 관점에서 설명할 수 있다. 첫째, 큐레이션 과정에서 대상 세계의 구성 요소를 균형적으로 파악하고 있는지 확인하는 가이드라인 역할을 한다. 예시에서처럼 대상 세계를 3가지 영역에서 파악하려는 계획을 세웠다면, 큐레이션의 과정에서 이 각각의 범주에 속하는 데이터를 찾고자 노력할 것이다. 그러한 기준이 세워져 있지 않으면 데이터는 어느 한 방향으로 편중되고 대상 세계를 균형적으로 보이는 것은 불가능하게 된다. 클래스 설정의 두 번째 이유는 개체의 속성이나 개체 사이의 관계를 기술할 때 유효성과 정확성을 높이는 데 도움이 되기 때문이다. 예시의 관계어 중 'isHeldAt (~이 ~에서 열렸다)'의 주어가 되는 개체는 '사건' 클래스에 속해야 하고, 목적어가 되는 개체는 '장소' 클래스에 속하게 해야 유의미한 설명이 된다. 클래스를 설정하고 모든 개체가 클래스에 소속되게 하는 것은 이처럼 데이터의 유효성을 제고하는 데 도움이 된다.

앞에서 제시한 데이터 모델링의 예시와 온톨로지 설계를 비교해 보면, 온톨로지의 클래스는 관계형 데이터베이스의 엔티티와 유사한 것임을 바로 알 수 있다. 관계형 데이터베이스에서는 데이터베이스를 만들 때 유형에 따라 데이터 구조를 달리할 수

있도록 여러개의 엔티티를 설정하는 것인데, 그 유형과 구조도 대상의 성격과 의미에 따라 파악되는 것이다. 사실상 데이터베이스의 엔티티와 온톨로지의 클래스는 성격과 기능면에서 매우 유사하다. 그렇기 때문에 데이터베이스를 먼저 공부한 학생들은 온톨로지에 대해 배울 때 클래스 설계를 그다지 어렵게 생각하지 않는다. 반대로 데이터베이스를 배운 경험이 없이 온톨로지 이론을 처음 접하는 문과생들은 클래스 설계를 너무나 어려워하는 듯 했다. 그것은 '클래스'가 '학문적 분류 체계'인 것처럼 오해했기 때문이다. 아무리 작은 일이라도 그것에 대해 누구나 동의할 수 있는 객관적 분류기준을 세우는 것은 쉽지 않은 일이다. 나는 클래스 설계를 어려워하는 학생들에게 '분류의 함정에 빠지지 말라'고 충고한다. 온톨로지의 클래스 설계는 대상 세계의 구성요소를 학술적으로 분류하려는 것이 아니라, 내가 그 가운데 어떤 유형의 것들을 관심있게 보고 데이터의 세계로 옮겨 놓으려 하는지를 밝히는 것이다. 물리적으로 동일한 대상 세계라 할지라도 그것을 바라보는 큐레이터의 관점과 시각이 다르면 그에 따라 클래스 설계의 내용도 달라질 것이다.

② 개체 (Individual Object)

앞에서 설명한 '클래스'는 대상 세계에 존재하는 다양한 구성요소를 유형별로 묶어서 보려는 범주일 뿐, 그 세계에 실재로 존재하는 실체가 아니다. 그 세계에 실재로 존재하는 것 하나 하나를 '개체'라고 상정한다. 그것은 한 사람의 인물일 수도 있고, 한 권의 책일 수도 있고, 하나의 건물, 어느 한 시점에 일어난 사건일 수도 있다. 온톨로지에서 말하는 개체(Individual Object 또는 Individual)는 그 물리적인 개체에 1:1로 대응하는 데이터이다.

물리적인 세계에 존재하는 개체의 어느 것을 선택해서 그것을 디지털 세계의 개체로 옮겨 놓는 작업, 즉 개체 데이터를 만드는 작업은 '온톨로지 설계'가 아니라 먼저 만들어진 온톨로지에 따라 시맨틱 데이터를 만드는 일에 해당한다. 하지만 앞으로 만들어질 다지털 세계의 개체가 어떤 요건을 갖추어야 하는지, 어떤 속성을 가질 수 있는지를 정하는 것은 온톨로지 '설계'에 속한다.

온톨로지의 개체가 되기 위해서 갖추어야 할 조건을 알아본다.

【식별자】 무엇보다 먼저, 데이터로서의 개체는 그것을 유일하게 식별할 수 있는 고유한 이름(identifier, 식별자)을 가져야 한다. 그래야만 하나의 개체를 다른 것과 연결지을 때 혼란과 애매함이 발생하지 않기 때문이다. 데이터베이스 관리자나 프로그래머들은 이것을 지극히 당연한 일로 여긴다. 하지만 인문학 연구자들의 입장에서 보면 이 일은 시간이 걸리고 판단이 주저되는 일이 되기도 한다. 종묘 공신당 인조묘에 제향된 '이귀(李貴)'라는 인물과 『묵재일기』의 저자 '이귀(李貴)'는 같은 인물인가라는 질문에서부터 1887년 신정왕후의 80세 생일 잔치에 사용된 '표범 가죽 방석'과 1892년 고종의 즉위 30주년 기념 잔치에 사용된 '표범 가죽 방석'을 같은 물건으

로 볼 것인가 하는 등의 물음에 대해서는 사실 확인 절차를 거치거나 판단의 기준을 마련해서 답을 정해야 할 것이다. 그것을 하나의 동일한 개체로 판단한다면 하나의 식별자로 묶어주어야 하고, 다른 것이라고 판단하면 각각 다른 식별자를 갖도록 해야 한다. 인문지식 큐레이션을 위해 온톨로지를 사용하고자 할 때에는 개체에 대해 식별자를 부여하는 기준과 방법에 대해 충분히 숙고해야 한다. 이전의 경험에서 발견한 문제점이 있다면 이번 큐레이션에서는 어떤 대안을 적용할지 미리 생각할 필요가 있다.

시맨틱 데이터가 월드와이드웹의 전역에서 유통될 수 있게 하려는 '시맨틱 웹'의 구상에서는 특정 온톨로지에서 정의된 개체가 범지구적으로 유일하게 식별될 수 있는 방법으로 인터넷 주소 형식의 이름을 부여한다.

```
http://dh.aks.ac.kr/individuals#이귀-1557
http://dh.aks.ac.kr/individuals#표피방석
```

이 형식은 특정 온톨로지 안에서 정의된 개체 식별자 앞에 인터넷 주소 형식의 이름공간(namespace) 식별자(URI, uniform resource identifer) 를 붙이는 것이다. 이름 공간(namespace)이란 하나의 이름이 단 하나의 개체만을 가리키는 범위를 추상적으로 상정한 것이며, 그것이 URI 형태를 취하는 것은 인터넷 상에서 유일성을 보장하는 명명법이기 때문이다. 이름 공간(namespace)의 URI는 그저 이름일 뿐이며, 그 주소에 해당하는 사이트가 인터넷 상에 꼭 있어야 하는 것은 아니다. URI 형식의 개체 식별자는 큐레이션의 결과물이 디지털 아카이브에 축적되었을 때, 그것이 범지구적으로 재이용(reuse)될 수 있게 하기 위해 필요한 장치이다. 큐레이션의 과정에서는 수행중인 프로젝트를 하나의 이름 공간으로 상정하고 모든 개체 데이터가 그 안에서 유일성을 갖게 하면 된다.

【클래스 소속】 온톨로지의 개체는 예외없이 어느 하나의 클래스에 속한다. 필요에 따라 두 개 이상의 클래스 속하게 할 수도 있는데 그것도 기준과 원칙을 세워 두어야 한다. 다음은 데이터 큐레이션 과정에서 자주 고민하게 되는 사례의 하나이다.

source	target	relation
황성신문	남궁억	founder
시일야방성대곡	황성신문	isPostedIn
황성신문_터	황성신문	isSiteOf

위의 그래프에서 하나의 개체로 표시된 '황성신문'은 남궁억 등이 설립한 신문사이고, 을사조약에 항의한 장지연의 논설 '시일야방성대곡'이 실린 '기록물'이며, 현재의

종각역 5번 출구 근처에 자리했던 황성신문사 건물을 뜻하기도 한다. 이처럼 복합적인 성격의 개체는 어떻게 처리해야 할까?

하나의 개체가 '기관/단체', '기록물', '장소' 등의 여러 클래스에 소속하게 할 수도 있지만, 개체를 좀더 엄밀하게 구분하여 각각 하나의 클래스에 속하게 할 수도 있다.

source	target	relation
황성신문(신문)	황성신문(신문사)	publisher
황성신문(건물)	황성신문(신문사)	isOfficeBuildingOf
황성신문(신문사)	남궁억	founder
시일야방성대곡	황성신문(신문)	isPostedIn
황성신문_터	황성신문(건물)	isSiteOf

한편, 이러한 류의 데이터가 많거나 그것이 이번 큐레이션의 주요 관심사라면 '신문'이라는 클래스를 만들고 이 클래스의 개체는 기관/단체, 기록물, 장소에 관계된 속성을 모두 가질 수 있게 하는 방법도 있을 것이다.

이런 여러 가지 방법 중에 어느 것이 더 좋은지를 판단하는 절대적인 기준은 없다. 그러나 어떤 특정 주제를 가지고 하나의 큐레이션 프로젝트를 수행하는 경우라면, 적어도 그 안에서는 개체의 클래스를 정하고 개체 사이의 관계를 기술하는 데 일관성이 있어야 한다.

③ 관계성 (Relation, Object Property)

'관계성'이란 개체가 가질 수 있는 속성 가운데 다른 개체와의 관계로 설명될 수 있는 것을 말한다. 조선시대 화가 김홍도에 대해 설명하는 아래의 두 문장을 비교해 보자.

- 김홍도의 아들은 김양기이다
- 김홍도의 호는 단원이다

여기서 '아들이 김양기'라는 것과 '호가 단원'이라는 것은 모두 '김홍도'라는 개체가 가지고 있는 특성을 이야기하는 것이라고 할 수 있다. 그런데 첫 번째 문장에 나오는 '김양기'는 '김홍도'와 마찬가지로 한 사람의 인물이라는 점에서 독립적인 개체로 인정될 수 있는 존재이다. 이에 반해 '단원'은 굳이 독립적인 개체로 삼을 필요가 없고 '김홍도'라는 인물에 종속된 데이터로만 쓰면 된다. 전자는 '관계성'(Relation)의 표현이고 후자는 '속성'(Attribute)의 표현이다.[7)]

7) W3C에서 제안하는 Web Ontology Language(OWL)에서는 전자를 '객체 속성(Object Property)'이라고 하고, 후자를 '데이터형 속성(Datatype Property)'라고 한다.

‘관계성’을 데이터화 할 수 있다고 하는 것은 온톨로지의 큰 특징이다. 우리가 인문지식의 디지털 큐레이션을 위해 굳이 온톨로지를 공부하고, 그 기술을 도입하려 하는 이유도 그것이 관계성의 표현을 통해 대상 세계의 문맥을 밝히는 데 도움을 주기 때문이라고 할 수 있다.

온톨로지에서 관계성을 정의하는 문법 가운데 대표적인 것이 RDF(Resource Description Framework)인데, 이것은 두 개의 개체를 하나의 관계서술어로 연결하는 구조이다.[8] 몇 가지 예시를 보이면 다음과 같다.

source	target	relation	설명
김홍도	김양기	hasSon	김홍도는 김양기를 아들로 두었다.
단원유묵첩	김홍도	creator	단원유묵첩의 창작자는 김홍도이다.
단원유묵첩	김양기	contributor	단원유묵첩의 기여자는 김양기이다. (편찬)
단원유묵첩	국립중앙박물관	currentLocation	단원유묵첩은 현재 국립중앙박물관에 소장되어 있다.
김양기	호산외사	isMentionedIn	김양기는 호산외사라는 책에 언급되었다.
호산외사	조희룡	creator	호산외사의 저자는 조희룡이다

RDF 문의 구조는 단순하다. 하지만 다양한 관계서술어를 정의해서 사용할 수 있고 이것을 이용하여 서로 관계가 있다고 판단되는 모든 개체들을 연결해 줄 수 있기 때문에 대상 세계의 복잡한 사실 관계를 구조적인 데이터로 전환하는 역할을 할 수 있다.

RDF 구문에 사용하는 관계서술어 어휘는 클래스(Class) 정의와 마찬가지로 온톨로지 개발자가 정하는 것이다. 어떤 어휘로 무슨 의미를 표시하는가에 대해 기술적인 표준안이나 권장안이 존재하는 것은 아니지만, 여러 전문분야의 국제적인 기구나 단체에서 해당 분야 데이터의 교환·공유를 위해 제정한 온톨로지의[9] 어휘들을 참조할 수 있다.

디지털 큐레이션의 방법으로 온톨로지를 쓰고자 한다면, 누군가 만들어 놓은 기존 온톨로지의 관계서술어를 그대로 사용할 수도 있고, 큐레이터 자신이 새로운 어휘들

8) RDF는 W3C(World Wide Web Consortium)가 제안하는, 웹상의 데이터 교환을 위한 표준 모델이다. RDF 구문의 형식은 웹의 하이퍼링크 구조를 확장한 것으로, 연결의 출발점과 도착점뿐 아니라 그것들 사이의 관계도 URI로 명명한다. (Resource Description Framework (RDF), https://www.w3.org/RDF/)

9) Dublin Core (DC) ontology (http://dublincore.org/ 문헌정보), Friend Of A Friend (FOAF) ontology (http://www.foaf-project.org/ 인적/사회적관계), Europeana Data Model (https://pro.europeana.eu/share-your-data/metadata 문화유산 메타데이터) 등

을 정의할 수도 있다. 디지털 큐레이션의 현장에서는 이 두 가지 방법을 절충적으로 운용한다. 즉, 내가 쓰고자 하는 것과 동일한 의미를 표현하는 관계서술어가 누군가에 의해서 이미 정의된 적이 있고, 또 여러곳에서 그것을 사용하고 있다면, 내가 굳이 같은 것을 다시 만들기보다는 기존의 것을 이용하는 것이 권장된다. 온톨로지의 관계서술어 중에 기존의 것과 새로 정의한 것이 혼재해 있을 때, 그것이 원래 어디에서 정의된 것인지를 알게 하는 방법은 이름 공간을 표시하는 접두어와 함께 사용하는 것이다.[10)]

디지털 큐레이션을 수행할 때 시맨틱 데이터의 관계서술어를 미리 약속된 어휘로 정의하는 이유는 데이터 사이의 관계가 편찬자에 따라 다르게 정의될 경우, 실효성 있는 데이터 네트워크의 구현이 어렵기 때문이다. 온톨로지의 관계서술어는 인간의 언어에서 가져오는 경우가 대부분이지만, 자연어에서처럼 다의적이거나 모호하게 사용되지 않도록 명확하게 한정된 의미로만 쓰이게 해야 한다. 처음부터 많은 수의 어휘를 정의하기보다는 아카이브 대상 정보의 성격을 살펴서 명확하게 데이터화 할 수 있는 기본적인 관계들을 먼저 정의하고, 데이터를 확장할 때 새로운 관계어들을 추가하는 방법이 권장된다.[11)]

④ 속성 (Attribute, Datatype Property)

앞에서 보았듯이, 어느 개체의 특성을 설명할 때 다른 개체와 어떤 관계를 맺고 있느냐를 알리는 것이 '관계성'의 기술이다. 이에 반해 다른 개체와 무관하게 그 개체에 속한 특성을 문자나 숫자 값으로 서술하는 것을 '속성'(Attribute, Datatype Property)이라고 한다. 화가 김홍도에 대해 그의 호가 '단원(檀園)'이었다거나, 그의

10) 본문의 예시에서 사용한 관계서술어의 이름 공간: creator와 contibutor는 도서관 관련 정보의 표준적인 기술(記述)을 위해서 제시된 '더블린 코어 용어'에서 정의된 것을 빌려온 것이며, 'currentLocation'은 문화유산의 정보화를 위해 만든 '유로피아나 데이터 모델'에서 가져온 것이다. 'hasSon'가 'isMentionedIn'은 한국문화 자원의 디지털 큐레이션을 위해 '한국문화 백과사전적 아카이브 데이터 모델(Envyes of Korean Culture)'의 이름 공간 안에서 정의한 관계서술어이다.

이름 공간 명칭	이름공간 식별자	접두어(prefix)	관계서술어(relation) 예시
Doublin Core Terms	http://purl.org/dc/terms/	dcterms	dcterms:creator dcterms:contributor
Europeana Data Model	http://www.europeana.eu/schemas/edm/	edm	edm:currentLocation
Encyves of Korean Culture Data Model	http://dh.aks.ac.kr/ontologies/	eckc	ekc:hasSon ekc:isMentionedIn

11) 한국학중앙연구원 디지털인문학연구소의 온톨로지 운용 및 확장 프로세스
① 연구팀은 새 프로젝트의 기초 연구자료와 스토리 주제가 선정된 시점에 그 프로젝트의 데이터 큐레이션에 적용할 온톨로지 초안을 제정. (※ 기존의 EKC 데이터 모델을 기반으로 하되, 당해 연도의 기초 연구자료 처리에 필요한 신규 어휘를 추가)
② 이 온톨로지에 입각하여 데이터 큐레이션을 수행하면서, 온톨로지 어휘의 사용 현황을 모니터링.
③ 온톨로지 어휘의 관리를 담당하는 태스크 포스를 구성하여, 새로운 어휘 제정의 요구가 있을 때 이를 판단, 제정, 공시하는 일을 수행.

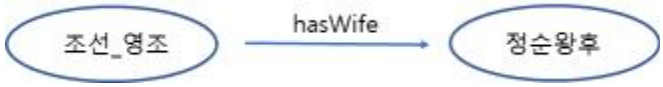
출생년도가 '1745년'이라고 하는 등의 정보는 대체로 '속성' 데이터로 취급되는 경우가 많다.

어느 한 개체에 대해 언급할 수 있는 속성은 무수히 많을 수 있겠지만, 그 가운데 의미가 있다고 판단되는 몇 가지를 클래스 별로 일정하게 기술할 수 있게 하는 것이 온톨로지에서 '속성' 데이터를 다루는 방법이다. '인물' 클래스에 속하는 개체는 모두 '출생연도'와 '사망연도'를 속성으로 갖게 한다든가, '장소' 클래스에 속하는 개체는 모두 '경위도 좌표'를 속성으로 갖게 하는 식이다.

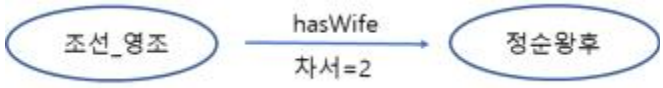
어느 클래스에 속하는 개체의 유의미한 속성이 무엇이냐에 대한 판단은 다른 온톨로지 요소 설계와 마찬가지로 그 데이터를 어떤 목적으로 활용할 것이냐에 따라 달라진다. 따라서 온톨로지의 속성 설계는 관계성 설계와 마찬가지로 처음부터 망라적인 탬플릿을 만들려 하기보다는 현재의 실용적인 관점에서 쓰임새가 있는 요소들을 위주로 하는 것이 권장된다. 향후 대상 자원이나 활용 목적이 확장되었을 때 새로운 속성 요소를 추가하는 것이 가능하기 때문이다.

⑤ 관계 속성 (Attribute of Relationship)

'관계 속성'이란 두 개체 사이의 관계에서 파악되는 특성에 대한 정보이다. 예를 들어 설명하기로 한다. 조선 영조의 왕비가 정순왕후였음을 알리려면 '조선_영조'와 '정순왕후'라는 두 개체를 'hasWife'라는 관계어(Relation, Object Property)로 연결해 주면 된다.



그런데, 정순왕후가 영조의 두 번째 왕비였음을 부가적으로 알리고자 한다면, 그 정보는 어디에 귀속시켜야 할까? 여기서 '두 번째'라는 정보는 '조선_영조' 또는 '정순왕후'라는 개체의 속성이 아니고, 'hasWife'라는 관계어의 속성도 아니다. 그것은 '조선_영조'가 '정순왕후'를 '아내로 삼았다(hasWife)'고 하는, 두 개체 사이에 맺어진 관계(relationship, connection)의 속성이다. 이렇듯 두 개체 사이의 연결이 이루어졌을 때 부수적으로 발생하는 특성을 '관계 속성'이라고 부르기로 한다.



'관계 속성'은 '관계어'만으로는 표현하기 어려운 관계의 정황을 데이터화하기 위해서 사용한다. 위의 예에서 보인 관계의 순서를 비롯해, 횟수, 분량, 강도 등에 관한 정보를 부가하거나 관계성을 만들어낸 행위에 부수되는 정보(시점, 역할 등)를 표시하

는 데 사용할 수 있다.

⑥ 온톨로지 기술 언어

이상에서 언급한 클래스, 개체, 관계성, 속성, 관계 속성 등이 온톨로지를 설계할 때 필수적으로 포함시키는 주요 구성 요소들이다. 대상 세계를 분석해서 그 세계를 정보화하는 데 필요한 온톨로지 요소들을 디자인하는 구상이 이루어졌다면 다른 사람들도 그 구상을 이해하고 참조할 수 있도록 정리해 놓을 필요가 있다.

이와 관련하여 가장 권장할 만한 방법은 온톨로지를 설계할 때 '온톨로지 편집기(Ontology Editor)'라는 유형의 소프트웨어를 사용하는 것이다.[12] 온톨로지 편집기는 '온톨로지'의 설계와 검증, 지속적인 확장 및 수정, 보완 기능을 제공하는 소프트웨어이다. 이용자는 이 소프트웨어상에서 클래스, 속성, 관계성, 개체 등 온톨로지 핵심 요소들을 디자인할 수 있고, 데이터의 입력을 통해 어느 정도 규모의 실험적인 데이터베이스를 구축할 수도 있다. 온톨로지 편집기상에서 데이터 클래스의 구조, 개체 상호간의 관계성 등을 시각적인 그래프로 확인하는 것도 가능하다.

온톨로지 편집기는 온톨로지 설계와 그 결과의 확인을 용이하게 하지만, 그것을 쓰는 보다 큰 이유는 이 소프트웨어를 사용함으로써 온톨로지 설계의 결과를 '웹 온톨로지 언어(Web Ontology Language, OWL)'[13]라고 하는 표준적인 온톨로지 기술 언어로 문서화할 수 있기 때문이다.

OWL은 온톨로지 구상을 명시적인 데이터로 표현할 수 있게 하는 기술 체계이다. 그렇기 때문에 OWL로 기술된 온톨로지 명세서는 어느 컴퓨터에서나 기계적으로 읽고 해석될 수 있다. 내가 설계한 온톨로지를 다른 사람들이 참조하거나 공유•활용하는 것을 용이하게 할 수 있는 것이다. OWL에 대해 깊이 있게 알지 못한다 하더라도 그것이 소통의 수단이 될 수 있다는 점에서 이 정도의 활용은 누구나 시도할 만한 일이다.

한편, OWL은 합리적인 온톨로지의 구현을 위한 여러 가지 약속의 체계이기 때문에 이 약속의 체계를 깊이 있게 이해하는 것은 정확성과 상호운용성이 높은 온톨로지를 설계하는 데 큰 도움이 된다. 그러나 나는 주관적으로 이 기계적 언어의 터득이 인문지식의 디지털 큐레이션을 위한 선행과정으로서 반드시 필요한 것은 아니라고 판단한다. OWL은 온톨로지를 기술(記述)하는 방법의 하나일 뿐, 온톨로지 설계의 절대적인 기준은 아니기 때문이다. 실제로 인문학 영역에서 추구하는 지식을 데이터화할 때,

---

12) 인문학도들에게도 진입장벽이 높지 않은 온톨로지 편집기로 Protégé™를 추천한다. Protégé™는 온톨로지 설계 및 시각화 기능을 제공하는 소프트웨어이다. 미국 스탠포드 대학의 의료정보학센터(Stanford Center for Biomedical Informatics Research)에서 개발 연구를 이끌고 있는 오픈 소스 소프트웨어이며 다양한 온톨로지 에디터 가운데 교육 및 연구 분야에서 가장 많이 쓰이고 있는 제품이다. http://protege.stanford.edu/

13) OWL(Web Onlogogy Language): W3C가 제안하는 온톨로지 기술 언어이다. 2004년에 제안되었으며, 이것을 부분적으로 수정·확장한 OWL2는 2009년에 발표되었다.
https://www.w3.org/TR/owl2-overview/

OWL의 규칙을 그대로 따르는 것이 불편하거나 비효율적인 경우도 적지 않다. 실제적인 데이터 아카이브의 구현을 위해서는 온톨로지 설계 기법과 함께 데이터베이스 구축에서 일반적으로 사용하는 엔티티 설계의 방법을 혼용해야 할 필요도 있다. 내가 이 책에서 지향하는 것은 독자들이 온톨로지 기반 데이터 큐레이션의 목적과 취지를 충분히 이해하고 그 방법을 준용할 수 있게 하되, 그것을 위해 마련된 기술적 형식의 세세한 부분에는 구애될 필요 없이 보다 유연한 방법으로 인문지식 큐레이션의 결과를 얻을 수 있게 하려는 것이다.

4. 시맨틱 데이터 아카이브

대상 세계의 구성 요소를 파악하고, 그 구성 요소 사이의 의미론적 맥락(Semantic Context)을 정의하는 일, 즉 '시맨틱 모델링'이나 '온톨로지 설계'라고 하는 과업을 수행했으면, 그 추상화된 얼개에 따라 대상 세계의 한 부분 한 부분을 디지털 세계에 재현하는 일이 이어지게 된다. 이 과정이 실제적인 디지털 큐레이션이다. 그리고 그 노력의 결과로 '디지털 아카이브'가 만들어지게 된다. 이렇게 '시맨틱 모델링'을 기반으로 만들어지는 디지털 아카이브를 '시맨틱 데이터 아카이브'라고 부르기로 하겠다.

다음 페이지에 보이는 네트워크 그래프는 우리가 만들어갈 시맨틱 데이터 아카이브가 어떠한 형태의 디지털 저작물인지 알게 하기 위한 예시이다.

이 예시에서 일차적으로 보이는 것은 대상 세계 구성 요소들의 의미론적인 관계망이다. 즉, 어떤 사건과 장소, 인물, 그리고 문헌 자료가 서로 어떠한 관계가 있는지를 알게 하는 것이다. 그런데, 우리가 만들려고 하는 디지털 콘텐츠는 이러한 식의 문맥 표현만을 목적으로 하는 것이 아니다. 네트워크 상에서 장소를 표시하는 노드를 클릭하면 그 건물의 형상을 3D 모델로 확인하거나 그곳의 지리적 위치와 지형을 3D 지도 상에서 확인할 수 있고, 사진첩 노드를 클릭하면 현장의 경관을 담은 실사 영상을 볼 수 있고, 문헌 자료 노드를 클릭하면 그 자료의 서지사항은 물론 원문 텍스트까지도 열람할 수 있으며, 사건 노드를 클릭하면 그것에 포함된 다양한 이벤트들을 하나 하나 살펴 볼 수 있다.

'시맨틱 데이터 아카이브'는 우리의 인문학적 탐구의 대상이 되는 세계를 데이터로 재현한 디지털 저작물이다. 이것은 대상 세계의 구성 요소 하나 하나에 대응하는 디지털 어셋들이 그 요소들 상호간의 의미론적 맥락까지도 알 수 있게 하는 형태로 집적되어 있는 것이라고 설명할 수 있다.

이러한 성격의 시맨틱 데이터 아카이브는 앞에서 설명한 '시맨틱 데이터 모델링' 즉 대상 세계의 의미론적 맥락에 대한 분석과 함께 그 데이터의 저장구조에 대한 설계의 과정을 거쳐서 만들어지게 된다. 이제 다음 장에서부터는 다양한 '시맨틱 데이터 모델링'의 사례들을 살펴봄으로써 자신의 관심사에 부합하는 인문지식 디지털 큐레이션의 실천 능력을 키워 가기로 한다.

※ 예시: 무신년진찬(1948년) 시맨틱 데이터

(김현, '제2장 디지털 큐레이션을 가능하게 하는 도구와 기술', 『디지털 큐레이션』, 북코리아, 2024)

※ 출간전 온라인 게시물 인용 표기:

김현, '디지털 큐레이션을 가능하게 하는 도구와 기술', 2024년 한국학대학원「인문정보학입문」강의 자료, http://dh.aks.ac.kr/~tutor/Documents/PDF/2024/김현-2024-디지털큐레이션(02).pdf