A Comparative Study of Data Schemas for Korean and Japanese Historical Text Collections

한/일 고문헌 텍스트 데이터 스키마 비교

조하영 (Hayoung Cho), 한국학중앙연구원

정경진 (Kyungjin Jeong), 이바라키 그리스도 대학

본고는 마크업 언어를 이용한 고문헌 자료의 디지털 판본 구축을 위해, 한국의 『한국문집총간』과 일본의 다이쇼 신수 대장경 데이터베이스(大正新脩大藏經, The SAT Daizōkyō Text Database)의 데이터 설계를 비교-검토하고자 한다. 이를 통해 기존 데이터 설계를 보완한 스키마 설계를 제안하여, 표준 동아시아 고전 텍스트 데이터 설계의 초석이 되기를 기대한다.

인문학 분야에서는 1990년대부터 고문헌 자료의 디지털 판본화 작업을 진행하고 있는데, 한국에서는 XML(eXtensible Markup Language)을, 일본에서는 TEI(Text Encoding Initiative)를 기반으로 이루어지고 있다. 이에 본고는 고문헌 자료의 디지털 판본 구축을 위한 목적으로 마크업 언어를 이용하는 한국고전번역원의 『한국문집총간』 XML과 일본 다이쇼 신수 대장경 데이터베이스(大正新脩大藏經, The SAT Daizōkyō Text Database)의 TEI를 검토하여 각각의 장단점을 파악하고, 두 마크업 언어의 절충안을 제안하고자 한다.

국내에서는 1995년 조선왕조실록 CD-ROM을 시작으로 편년체에 최적화된 데이터 구조의 마크업 언어가 구현된 이후(김바로, 2017), 현재까지도 한국고전종합DB, 디지털장서각, 한국경학자료시스템 등에서 활발하게 이용되고 있다. 우리나라의 역대 문집을 총망라한 『한국문집총간』 XML은 제목 정보, 간행정보, 소장정보, 본문내용을 나타낼 수 있는데 본문내용은 원목차정보와 권차정보로 나눠지고 권차정보는 제목, 최종정보, 문체, 내부록 등의 정보로 이루어져 있다 (그림1). 한국문집총간 XML은 한국 문집의 형식을 충실히 따르며 기존 내용을 왜곡하지 않고 최대한 그대로 서술할 수 있게끔 설계되었다. 이는 사용자의 필요에 의한 자유로운 변형이 가능한 XML의 장점을 잘 보여주지만, 동시에 서로 다른 사용자가 각기 다른 스키마를 설계할 경우에 데이터 호환성에 문제가 발생할 가능성을 내포하고 있다.

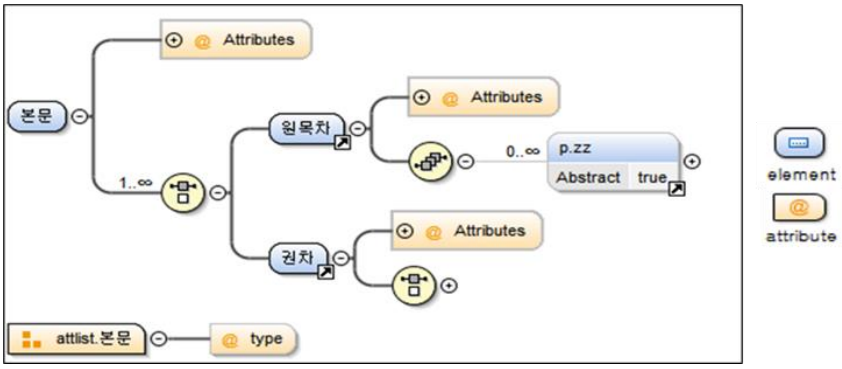

그림1. 한국문집총간 본문정보 일부

반면, TEI는 텍스트 인코딩을 위한 국제적인 규약이기에 TEI 기반의 다이쇼 신수 대장경 데이터베이스는 그러한 규약에 맞게끔 설계되었다. 사전에 정의된 태그를 이용해서 경전별로 고유번호를 부여하고 그 속에 존재하는 방대한 문장도 관리하고 있으며, 글자별 관련 속성까지도 사전에 정의된 태그를 이용했다. 제목정보는 <titleStmt>, 작성자는 <title>, 기타 지적내용은 <respStmt> 등 TEI를 아는 누구나 이해할 수 있는 설계이다 (그림2). 또한 일본의 한문훈독에서 사용되는 가에리텐(返點) 등의 요소는 <metamark> 태그로 처리(허인영, 2022) 하는 등 TEI 표준기술규칙뿐만 아니라 일본 자체만의 요소까지 더하여 DB를 구축하였다. 하지만 TEI는 서구 언어학 분야 중심으로 발전해왔기에 동아시아만의 고유한 고문헌 구조와 다양한 문자를 충분히 지원하지 못하기 때문에, 동아시아 고문헌의 특성을 반영하기에는 다소 제한적이며 다이쇼 신수 대장경 데이터베이스와 같이 자체 요소를 추가하는 순간 국제표준의 성격에서 벗어나게 된다.

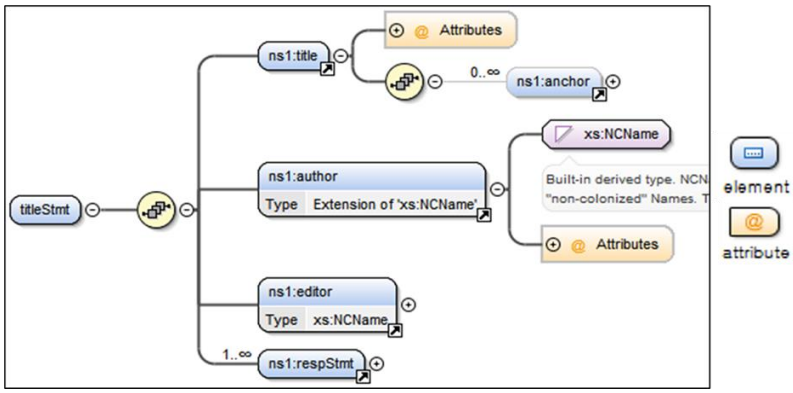

그림2. 다이쇼 신수 대장경 DB 제목정보titleStmt일부

TEI의 새로운 규약을 개발하는 등의 연구가 필요하다. 본고에서는 TEI 또는 XML 중 한 가지 방법만을 이용하여 고문헌을 디지털화 하기보다는, 그림3과 같이 서지적인 메타데이터는 TEI 요소를 기반으로 설계하되, 그 안에 담긴 텍스트는

XML을 이용하여 자유롭게 변형이 가능하도록 서술하는 방법을 제안하고자 한다.

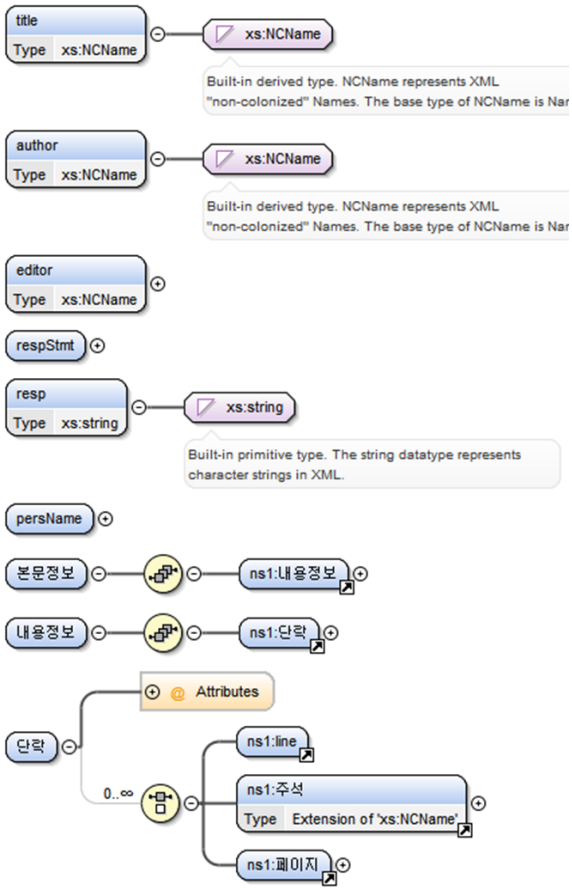

그림3. 본고에서 제안하는 데이터설계 개선안중 일부