

현대문학비평 XML 데이터 구축과 활용 방안*

— 디지털 인문학 연구를 위한 방법론 모색

이병주**

차례

- I. 서론
- II. 현대문학비평 XML 데이터 구축을 위한 마크업 규칙
- III. 현대문학비평 XML 데이터의 활용 방안
- IV. 결론

● 국문초록

본 연구는 전산언어학의 발달과 생성형 AI의 확산이라는 시대적 흐름 속에서, 상대적으로 디지털 인문학적 접근이 미진했던 한국 현대문학비평 텍스트를 대상으로 XML 기반의 데이터 구축 방법론을 제안하고 그 활용 방안을 모색하는 데 목적이 있다. 기존의 현대문학 연구는 문학의 ‘해석’과 ‘미

* 이 논문은 2023년 대한민국 교육부와 한국연구재단의 인문사회분야 신진연구자 지원사업의 지원을 받아 수행된 연구임(NRF-2023S1A5A8079304).

본 연구는 2025년 11월 21일 개최된 한국시학회 제56회 전국학술대회에서 발표한 원고를 수정 및 보완한 것이다. 귀중한 의견을 주신 토론자 김영임 선생님, 그리고 익명의 심사위원 세 분께 깊이 감사드린다.

** 공군사관학교 어문학과 조교수.

적 가치’를 중시하여 데이터 기반 분석에 회의적이었으나, 본 연구는 비평 텍스트의 특징인 ‘평가적 언술’과 ‘인용’ 자체를 XML 구조에 반영함으로써 이러한 이분법적 한계를 극복하고자 하였다.

이를 위해 본고는 국제 표준인 TEI(Text Encoding Initiative) P5 가이드라인을 준용하되, 문학비평의 특수성을 반영한 맞춤형 스키마(XSD)를 설계하였다. 구체적으로 비평가의 가치 판단(긍정·중립·부정)을 속성값으로 기록하는 <interp> 요소와 타 사상가 혹은 작품의 수용 양상을 보여주는 <quote> 요소를 마크업 규칙으로 추가하였으며, ISNI, Wiki, 한국민족문화대백과 등 외부 데이터와의 연계성을 확보하고자 하였다.

이렇게 구축된 XML 데이터는 단순한 키워드 검색을 넘어 XPath와 XQuery를 활용한 문맥적 탐색, 비평가와 작가 간의 영향 관계를 규명하는 네트워크 분석, 비평 용어의 개념사적 추적 등 다각적인 연구에 활용될 수 있다. 본 연구는 개별 연구자의 텍스트 해석 과정을 공유 가능한 기계 가독형 데이터 구조로 전환함으로써, 현대문학 연구의 폐쇄성을 극복하고 지속 가능한 디지털 지식 생태계를 구축하는 실천적 토대를 마련했다는 데 의의가 있다. 다만, 본 연구가 제시한 마크업 규칙에 따라 구축된 현대문학비평 XML 데이터를 현행 저작권법의 제약으로 인해 원문 데이터 전체를 연구자들에게 공개하지 못하는 점은 본 연구의 한계로 남는다.

I. 서론

전산언어학의 발달과 더불어 ChatGPT로 대표되는 생성형 AI의 발달은 새로운 기술의 등장을 넘어, 교육 및 연구 방식에 대한 전반적인 변화를 불러일으켰다. 이러한 변화의 흐름은 국문학 분야도 피할 수 없었으며, 그로 인해 최근 국문학계는 생성형 AI를 활용한 연구방법론을 논하기 시작했다. 이를 더 큰 범주로 보면, 국문학계에 ‘디지털 인문학’의 방법론이 도

www.kci.go.kr

입되고 있다고 볼 수 있을 것이다. ‘디지털 인문학’이라는 용어가 주는 새 로움은 있지만, 데이터를 구축하고 데이터를 바탕으로 텍스트를 분석하는 방식은 ‘국어국문학’이라는 학제 단위로 볼 때, 그 하위 계열인 국어학과 고전문학 분야에서는 그리 낯설지 않은 연구 방법이다. 하지만 대부분의 현대문학 연구자는 ‘디지털 인문학’의 연구방법론 도입에 회의적이다. 문학작품의 내재적 분석을 중시하는 연구방법을 전통적인 연구라고 부른다면, 전통적인 방식으로 현대문학을 연구하는 연구자들은 디지털 기술이 문학 텍스트의 미학적 가치를 제대로 분석할 수 있을지 의문을 품기 때문이다. 이러한 회의는 기술에 대한 거부가 아니라 문학 연구의 본질에 대한 근본적인 질문에서 비롯된다. 문학 연구의 핵심은 ‘해석’이다. 같은 작품을 읽더라도 연구자마다 다른 해석을 내놓을 수 있으며, 그 다양한 해석이 충돌하고 당위성을 확보하면서 문학을 해석하는 담론이 형성된다. 그런데 디지털 인문학이 중요하게 생각하는 ‘데이터 기반’의 분석은 이러한 해석의 다양성을 보장하기 어렵게 한다. 단어의 빈도나 등장인물의 네트워크 분석이 수치로 제시될 때, 기존의 미학적 해석은 객관적이지 못한 것으로 평가절하될 위험이 있고, 데이터에 근거한 절대적인 ‘해석’만 주장할 수 있기 때문이다. 더불어 이러한 방법론이 문학작품을 분석하는 데 항상 올바른 결론에 도달하는 것도 아니다. 예를 들어 문학작품에 등장하는 특정 단어의 빈도가 높다고 하여 반드시 해당 단어가 작품에 핵심 주제를 규명할 수 있는 것은 아니기 때문이다.

이러한 이유 때문인지는 알 수 없으나, 현대문학을 주제로 한 디지털 인문학의 연구성과들은 국어학이나 고전문학에 비해 미비한 수준이다. 2000년대 중반 이후 다양한 연구들이 앞다투어 나오고 있지만, 아직 그 질적 성취는 국어학이나 고전문학에 비해 미흡하다고 할 수 있다. 특히 현대문학을 대상으로 데이터 구축 방안을 논의한 연구는 거의 없다고 할 수 있다. 어쩔 수 없이, 국사학계나 고전문학 또는 외국의 사례를 통해 현대문

학 데이터를 구축하는 방안을 모색할 수밖에 없는 현실이다.

한국의 고문헌 자료를 데이터화하는 분야에 선구적인 업적을 쌓은 김현¹⁾의 경우, 1990년대부터 XML 기반 고문헌 데이터 베이스를 구축하기 위한 방법을 제시하였다. 『조선왕조실록』, 『승정원일기』, 『한국문집총간』 등의 데이터 구조를 설계하였다. 국학중앙연구원이 소장한 데이터를 디지털화하여 2010년부터 ‘한국고문서자료관’에 공개하였다. 김현의 연구를 통해 파악할 수 있는 것은 마크업의 역할은 형태를 표현하는 것이 아니라 자료의 구조를 표현해야 한다는 것이다. 이는 텍스트를 평면적으로 생각하지 않고, 관계의 집합으로 이해했다는 것을 말한다. 예를 들어 『조선왕조실록』은 왕대→연도→월→일→기사라는 계층적 구조로 이루어져 있으며 이는 상하관계(계층적 포함관계), 전후관계(시간적 순서 또는 인접관계)를 의미한다. 이 개념은 2016년 연구²⁾에서 ‘시맨틱 온톨로지 모델’로 확장된다. 김현은 단순히 데이터를 제공한 것에 그치지 않고, 고문서와 족보에 기록된 사실들의 관계를 분석하여 21개의 의미적 관계망을 분석하였다. 김서윤³⁾은 김현의 방식을 참고하여, 『논어』 주석의 XML 데이터를 설계하면서, 관계망에 주목하여, 경문-주석-개념-인용 관계를 TEI 구조와 RDF⁴⁾ 온톨로지 모델로 통합하였다. 이처럼 디지털 인문학 연구는 데

-
- 1) 김현, 「조선왕조실록 전산화를 위한 마크업(MarkUp) 규칙」, 『古文研究』 제9집, 한국고문헌연구회, 1996, 244-262쪽; 김현, 「XML 기반 고문헌 데이터베이스 구축 연구」, 『한국비교문학』 제29집, 한국비교문학회, 183-230쪽.
 - 2) 김현, 「디지털 인문학과 고문헌 자료 연구」, 『열상고전연구』 제50집, 열상고전연구회, 2016, 13-28쪽.
 - 3) 김서윤, 「시맨틱 DB 구축을 위한 『논어』 주석 XML 데이터 설계」, 『한문학논집』 제71집, 근역한문학회, 2025, 211-242쪽.
 - 4) 시맨틱 데이터는 일반적으로 주어(Subject), 술어(Predicate), 목적어(Object)의 트리플 구조(triple)로 이루어진 RDF를 활용해 관계를 표현한다. RDF는 데이터를 문장처럼 표현하는 방식으로, 각 문헌이나 주석 속 ‘개념 간의 관계’를 기계가 이해할 수 있도록 구조화한다. 김서윤은 『논어』의 “朋同類也(벗은 같은 부류이다)”라는 구절을 예로 든다. 주어(Subject): “同類也”(정의 내용) 술어(Predicate): “hasTarget”(정의가 가리키는 관계) 이러한 구조는 단일 구절의 해석에 국한되지

이터를 단순히 작성하는 게 아니라, 연구자의 연구목적에 따라 데이터를 ‘설계’하고 ‘구조화’하는 일이다. 이는 김현·안승준·류인태⁵⁾의 연구에서도 알 수 있는데, 이들은 ‘인문학 자료의 디지털화’와 ‘디지털 인문학’의 차이를 “누가 하느냐”에 있다고 주장하며, ‘인문학 자료의 디지털화’는 자료 이용의 편의성 증진을 목표로 하지만, ‘디지털 인문학’은 디지털 환경에서 자기주도적으로 연구와 교육 활동을 수행하는 것이라고 정의하였다. 장문석·류인태⁶⁾의 연구는 이와 같은 데이터의 구축과 활용을 보여주기 위해 ‘운동주’를 구체적인 예로 제시한다. 장문석·류인태는 운동주와 관련된 범주를 17개의 클래스(시간, 공간, 작품, 기록 등)로 나눈 뒤, 다시 형식-내용-의미의 관계를 중심으로 시맨틱 데이터를 구축한다. 본고와 가장 유사한 목적을 달성한 연구로, 정송이·김바로⁷⁾의 연구가 있다. 조선총독부박물관 문서를 XML 기반으로 설계한 해당 연구는 연구자가 자신만의 분류 체계를 적용하여 데이터 분석을 진행했으며, 국내외 연구기관과의 데이터 연계를 극대화하여 디지털 아카이브의 활용 가치를 증진 가능성을 도모하였다.

앞서 살펴본 선행연구들은 고전문학이나 사료(史料)를 기본 자료로 사

않고, 개념들 사이의 의미망을 형성하는 시맨틱 그래프(semantic graph)로 확장된다. 요컨대 RDF는 문헌의 언어적 관계를 ‘주어-술어-목적어’ 삼중 구조로 전환하여, 인간의 해석 행위를 기계가 이해할 수 있는 의미적 모델(semantic model)로 전환하는 기술이며, 시맨틱 데이터의 기본 원리라 할 수 있다. 김서윤, 위의 글 참고.

- 5) 김현·안승준·류인태, 「데이터 기반 인문학 연구 방법의 모색 : 문중 고문서 아카이브와 디지털 인문학의 만남」, 『횡단인문학』 창간호, 숙명여자대학교 인문학연구소, 2018, 17-61쪽.
- 6) 장문석·류인태, 「디지털 인문학과 한국문학 연구(1) : 작가 연구를 위한 시맨틱 데이터베이스 설계」, 『민족문화사연구』 제75집, 민족문화사학회, 2021, 347-426쪽.
- 7) 정송이·김바로, 「조선총독부박물관 문서 XML 데이터 설계 및 활용 시론 : XML 기반 다중 분류 체계 및 내·외부 데이터 연결을 중심으로」, 『인문과학』 제96집, 성균관대학교 인문학연구원, 2025, 181-227쪽.

www.kci.go.kr

용했다. 현대문학 데이터가 제대로 구축되지 않은 이유는 고전문학이나 역사 사료와 달리 ‘시간적 거리’가 확보되지 않았기 때문이다. 현대문학을 디지털화하고 디지털 인문학의 방법론을 활용하여 연구할 때 가장 근본적인 문제는 ‘저작권’이다. 한국 저작권법에 따르면 저작권(이하 저작권)은 저작자 사망 후 70년간 보호된다. 2025년을 기준으로 보면, 1955년 이후 사망한 작가의 작품은 모두 저작권의 보호 대상이다. 저작권의 상속인이 전혀 없는 경우⁸⁾에는 저작권이 소멸하는 반면, 월북 문인의 경우 저작권의 존속 여부에 대한 별도의 검토가 필요하다.⁹⁾ 이와 같이 특별한 경우가 아닌 이상 대다수의 현대문학 작가는 아직 저작권이 소멸하지 않았다. 하지만 한국현대문학사에서 큰 비중을 차지하는 이광수가 1917년 『무정』을 발표하고, 1950년에 사망하여 2000년 저작권이 소멸하였다는 것을 고려했을 때, 저작권 문제는 시간이 지남에 따라 해결될 수 있는 문제이다. 현대문학비평의 경우, 1910년대 중후반 ‘월평’과 ‘선후평’ 등으로 명명된 ‘문예비평’의 등장과 더불어 1926년부터 ‘비평 논쟁’이라고 부를 만한 비평들이 등장했다는 것을 고려할 때¹⁰⁾, 1910년대의 비평들부터 현대문학비평으로서 데이터베이스를 구축할 필요가 있다. 게다가 이 시기 활동한 작가들의 저작권이 만료되었거나 곧 만료되는 시기가 도래한다.

그렇다고 하여 이러한 저작권 문제로 인해 현대문학비평에 대한 디지털

8) 이러한 대표적인 작가로 ‘이상’이 있다. 이상 문학작품을 데이터화하여 분석한 선구적인 사례로 지혜인의 연구가 있다. 지혜인, 「디지털 감각·감정 분석을 통한 이상 문학의 에피파니 연구」, 『상허학보』 제72집, 상허학회, 2024, 753-827쪽.

9) 현재 (사)남북경제문화협력재단에서 백석, 이기영, 이응악, 박세영, 윤복진, 송영, 조영출, 최명익, 한설야, 현덕 등의 저작권을 위탁받아 관리하고 있다. 그러나 해당 기관이 모든 월북 문인의 저작권을 관리하는 것은 아니며, 그 외의 경우에는 한국저작권위원회를 통해 법정 허락을 받고 저작권 사용료를 공탁하는 방법이 있다.

10) 강용훈, 『비평적 글쓰기의 계보 — 한국 근대 문예비평의 형성과 분화』, 소명출판, 2013 참고

인문학적 연구가 전혀 불가능한 것은 아니다. 저작권 예외 조항¹¹⁾에 따라 제한적으로 연구 활동이 가능하기 때문이다. 먼저 비공개 연구용 데이터를 구축하여, 연구팀 내부에서만 접근할 수 있도록 하거나 혹은 도서관을 통해서만 접근할 수 있는 데이터로 공개하는 방법이 있다.¹²⁾ 이는 해외의 연구자들이 데이터 관련 논문을 발표할 때 사용하는 방식이다. 다음으로 통계 및 분석 결과를 공개하는 방법이다. 원문 전체를 공개하는 것은 저작권법에 위배되지만, 분석의 결과는 제시할 수 있다. 이와 같은 방식이 현재 디지털 인문학 연구자들이 주로 활용하는 방식이지만, 이는 연구 재현이 불가능하므로 과학적 연구를 표방하는 디지털 인문학에 큰 한계로 작용할 수 있다.

이러한 한계에도 불구하고, 본고가 현대문학비평 XML 데이터 구축이 필요하다고 판단한 이유는 다음과 같다. 첫 번째로 데이터의 파편화와 폐쇄성이다. 이미 디지털화된 자료들조차 기관별, 연구자 별로 상이한 방식으로 구축되어 있다. 이로 인해 데이터 간의 상호 통합이 불가능할 뿐만 아니라, 연구자들 간의 협업도 불가능하다. 특히 원문 확인과 그 원문을 현대적 문법에 맞게 변환하는 일은 국문학 연구에서 필수적이다. 그중에서도 국한문혼용, 세로쓰기, 판본의 상이함, 판독 불가능한 글자 등과 같

11) 「저작권법」 제35조 5(공정 이용) “① 제23조부터 제35조의4까지, 제101조의3부터 제101조의5까지의 경우 외에 저작물의 일반적인 이용 방법과 충돌하지 아니하고 저작자의 정당한 이익을 부당하게 해치지 아니하는 경우에는 저작물을 이용할 수 있다. ② 저작물 이용 행위가 제1항에 해당하는지를 판단할 때에는 다음 각 호의 사항등을 고려하여야 한다. 1. 이용의 목적 및 성격 2. 저작물의 종류 및 용도 3. 이용된 부분이 저작물 전체에서 차지하는 비중과 그 중요성 4. 저작물의 이용이 그 저작물의 현재 시장 또는 가치나 잠재적인 시장 또는 가치에 미치는 영향”

12) 이는 이미 국립중앙도서관의 데이터지원 사업으로 진행되고 있다. 국립중앙도서관은 2025년 5월 『가톨릭청년』, 『문장』, 『조광』에 대한 데이터 수요를 조사하여, 26년 1월 초 제한된 연구자들에게 공개하기로 하였다. [https://nl.go.kr/rink/contents/R10400000000.do?schM=view&id=45062\(2025.10.1 최종검색\)](https://nl.go.kr/rink/contents/R10400000000.do?schM=view&id=45062(2025.10.1 최종검색))

은 문제는 모든 현대문학 연구자가 경험하는 연구 제약이다. 하지만 이러한 연구의 경험은 약 70년 가까이 이루어진 국문학 연구에서 축적되지 않고 연구자 개인의 경험으로 남거나, 일부 연구자들끼리만 폐쇄적으로만 공유하고 있었다. 이와 같은 폐쇄성은 연구자의 기초연구 능력의 토대를 다질 수 있게 해준다는 장점은 있지만, 연구를 제한하는 장벽으로도 작용한다.

두 번째로 비구조적 데이터에 머무는 공공데이터의 한계이다. 김바로¹³⁾가 지적한 바와 같이 현재 국립중앙도서관과 같은 기관이 제공하는 인문자료가 여전히 PDF 수준의 데이터에 머물러 있으며, 이는 기계적 접근이 불가능하다는 점에서 분석 가능한 자료의 요건을 충족하지 못한다. 그는 텍스트 분석을 위한 기초 조건으로서, 데이터를 기계가 읽고 해석할 수 있도록 소프트웨어로 내용이나 구조를 확인하고 변환할 수 있는 상태의 정형화된 구조, 즉 CSV, XML, RDF 등의 오픈 포맷으로의 변환이 필수적임을 강조한다. 따라서 공공데이터로서 인문학 자료가 통일된 체계로 작성되어 공유된다면, 현대문학 연구는 더욱 활성화될 수 있을 것으로 기대한다.

세 번째로 문학비평은 문학작품과 다른 특수성을 갖기 때문에 데이터로 구축될 때 효과적으로 분석될 수 있다. 앞서 언급한 것처럼 문학작품을 ‘해석’하는 것은 미적 감상으로서 차이가 발생할 수 있다. 따라서 문학비평에는 비평가만의 ‘평가적 언술’이 존재한다. 비평가는 작품이 뛰어난 이유나 혹은 한계를 서술하는데, 이는 비평가의 미적 평가 방식을 객관적으로 파악할 수 있는 근거가 된다. 즉 이미 비평가가 작품을 ‘해석’한 결과물로서, 이 ‘해석’에 대해 연구자가 평가하고 구조화하는 것이 비평을 연구하는 일(메타비평)의 핵심이다. 따라서 이러한 ‘평가적 언술’을 데이터로

13) 김바로, 「<공공데이터법>과 인문데이터 : 공공기관 보유 인문데이터 공개 신청 사례를 중심으로」, 『한국고전연구』, 제57집, 한국고전연구학회, 2022, 172-173쪽 참고

구축하여 분석할 수 있다면, 더욱 비평가의 비평 원리나 구조를 파악하기 용이할 것이다. 더불어 문학비평은 ‘작가(인물)’와 ‘작품’ 인용이 주를 이룬다. 이러한 인용은 단순한 참고가 아니라 비평가-작가 간의 상호 영향 관계와 더불어 비평 담론의 형성과정을 파악할 수 있는 중요한 자료이다.

본고는 이러한 문제의식을 바탕으로 한국 현대문학비평 텍스트를 XML 기반 데이터로 구축할 수 있는 방법론을 제시하고 이후 디지털 인문학의 방법론에 활용하기 위한 토대 작업을 진행하고자 한다. 구체적으로 한국 문학비평의 디지털아카이브 구축을 위한 마크업 규칙을 제시하고, 나아가 구축한 데이터를 바탕으로 계량적 문학 연구, 문학비평의 용어의 개념사적 연구, 비평에 인용되는 작가들의 네트워크 분석, 문학사를 통한 정전 형성 과정 연구 등 다양한 디지털 인문학의 방법론을 제시하는 것을 목적으로 한다. 이는 추후 시맨틱 데이터베이스의 형태로 확장될 수도 있을 것이다.¹⁴⁾

II. 현대문학비평 XML 데이터 구축을 위한 마크업 규칙

1. 현대문학비평 연구를 위한 마크업 요소 선별

현대문학비평 텍스트를 XML로 구조화한다는 것은 단순히 기존자료를 기계적 접근이 가능한 전자 자료로의 변환을 의미하지 않는다. 연구자는 단순히 문서를 전산화하여 정리하는 것이 아니라, 그 텍스트를 ‘어떤 관점

14) 시맨틱이란 독자 혹은 연구자들의 관심 영역에 따라 다양한 형태로 변환할 수 있는 모델을 말한다. 이는 단순히 키워드 검색의 수준을 넘어 텍스트의 맥락을 식별할 수 있는 단계를 의미한다. 이러한 시맨틱 데이터의 구축은 생성형 AI의 활용을 통해 앞으로 더 폭넓은 활용이 가능할 것이다. 장문석·류인태, 앞의 글, 360쪽.

으로 읽을 것인가’를 고려하여 XML 구조에 반영해야 한다. 예를 들어 박영희가 『조선현대문학사』에서 “『무정』이 발표된 후, 4, 5년 동안은 춘원의 독보시대이었다.”¹⁵⁾라고 서술한 문장을 살펴보자. 이 문장은 연구자가 읽기에는 이광수를 고평하는 간단한 의미의 문장이지만, 기계가 읽기 위해서는 다음과 같은 과정이 필요하다. 먼저 『무정』이라는 표기가 소설 작품임을 표시하여야 하며, ‘춘원’이 작가이며, ‘이광수’와 같은 인물이라는 것을 기계에 알려주어야 한다. 여기서 가장 중요한 것은 이 “독보시대”라는 문장이 긍정적인 평가라는 것을 알려주지 않으면 기계는 판단할 수 없다는 점이다. 추가적으로 연구자의 판단에 따라 “4, 5년 동안”이라는 시간도 마크업할 수 있을 것이다.

그렇다면 이제 현대문학비평을 데이터로 전환하기 이전에, 현대문학비평을 연구함에 있어 어떠한 요소들을 마크업할 필요가 있을지 검토하여야 한다. 하지만 앞서 서론에서 문제제기한 것처럼 본 연구의 목적은 현대문학비평을 XML데이터로 정리하기 위해, 파편화된 데이터를 통합하는 기준을 제시하고, 데이터의 폐쇄성을 극복하여 연구자 간의 지식 축적과 공유 가능성을 확보하는 일이다. 그러기 위해서는 개인 연구자가 임의로 기준을 정하는 것이 아닌 기존에 문학비평을 주제로 진행된 선행연구들이 문학비평의 어느 부분에 주목했는지를 파악해야 한다. 즉, 데이터 구축의 출발점은 새로운 체계의 창안이 아니라, 기존 비평 연구가 형성해온 관심 영역과 분석 단위를 종합하는 것이다. 이때 참조해야 할 것은 바로 문학비평을 대상으로 연구를 진행한 선행연구들이다. 비평에 관한 연구를 흔히 메타비평이라고 부른다. 메타비평은 특정 비평가의 사유 방식이나 비평 담론의 구조를 분석함으로써, 비평이 어떤 층위에서 의미를 생산하는지를 해명한다. 따라서 문학비평 XML 마크업 체계를 설계하는 일은 곧 이들

15) 『사상계』, 1958년 9월호, 438쪽.

메타비평이 제시해온 분석 범주를 데이터 구조의 형식으로 재현하는 과정이라 할 수 있다.

메타비평을 통해 XML 데이터 구축 시에 필요한 마크업을 연구하기 위해 비평에 관한 선행연구들을 선별하였다. 선별은 일관된 기준을 확보하기 위해 KCI 논문 인용 정보의 피인용지수를 활용하였다.¹⁶⁾ 이와 같은 인용정보는 박사학위논문이 제외되어 있어 부분적인 한계가 있으나, 논문을 양적으로 평가할 수 있는 가장 객관적인 방식이라 판단하였다.

〈표 2〉 선행연구에서 작가/작품을 인용한 방식

	인용자	인용 대상	인용 내용
1	하정일	이기영	임화는 특히 이기영의 《고향》에 주목한다. 《고향》에 “생생한 어휘와 조선어 고유의 아름다움”이 담겨 있기 때문이다.
2	조영복	페르낭 레제	김기림은 레제를 인용하며 “‘운동과 생명의 구체화’(페르낭 레제)로서의 기계의 미를 인정하는 것”이라고 주장했다.
3	이미순	조르주 바타유	김수영은 “요즘 시론으로는 조르주 바타유의 문학의 악과 모리스 블랑쇼의 『불꽃의 문학』을 일본 번역책으로 읽었는데, 너무 마음에 들어서 읽고 나자마자 즉시 팔아 버렸다.”라고 말했다.
4	황호덕	김소운	김윤식은 수필가 김소운에 대해 “그가 조선시를 일본어로 옮긴 건 큰 업적입니다”라고 언급했다.
5	오형엽	김영랑	박용철은 김영랑의 시에 대해 “그의 사행시는 천하일품이라고 나는 나의 좁은 문견을 가지고 단언”했다.

16) 선별의 과정은 다음과 같다. ① 비평사를 바탕으로 비평가로 분류할 수 있는 작가들을 선별하였다. ② 발표한 비평의 양을 고려하기 위해 전집 출판 여부를 확인하였다. ③ KCI 논문 인용 정보를 통해 해당 작가에 대한 연구 중 피인용지수가 높은 연구를 선정하였다. 피인용지수는 학술적 영향력의 지표로서 참고하였으며, 이를 통해 선별의 일관성을 확보하고자 하였다. ④ 창작과 비평을 겸한 작가의 경우 창작물에 대한 연구를 제외하고 비평 또는 시론에 관한 연구를 선정하였다.

<표 2>의 1번 항목은 하정일의 「임화의 민족문화론과 언어론」에서 발췌한 내용이다. 하정일은 임화가 이기영을 평가한 내용을 근거로, 민족어의 핵심 주체를 부르주아 헤게모니에서 극복한 ‘근대적 생산 인민의 언어’, 즉 민중어에서 찾으려 했던 임화의 언어론을 입증하는 긍정적인 사례로 제시한다. 이는 작품평가와 ‘민족’ 혹은 ‘민중어’라는 개념에 대한 인용이라고 할 수 있다. <표 2>의 2번 항목은 조영복의 「김기림 시론의 기계주의적 관점과 ‘영화시」에서 발췌한 내용이다. 조영복은 레제의 기계주의적 세계관을 김기림이 시론의 이론적 실천적 근거로 삼았음을 증명하기 위해 이를 인용한다. 이는 김기림의 외국 이론 수용을 살펴본 사례라고 할 수 있다. <표 2>의 3번 항목은 이미순의 「김수영 시론과 ‘죽음」에서 발췌한 내용으로, 이미순은 이 인용을 김수영이 불량쇼에 대한 애정을 역설적으로 드러낸 것으로 해석한다. 이는 김수영의 외국 문학 수용과 작가에 대한 평가를 인용한 것이라고 할 수 있다. <표 2>의 4번 항목은 황호덕의 「김윤식 비평과 문학사론, 총체성과 가치중립성 사이」에서 발췌한 내용으로, 황호덕의 연구는 김윤식이 김소운의 친일행위와 업적을 동시에 언급하는 것을 김윤식의 독특한 ‘가치중립성으로서의 근대’ 개념이 적용된 복합적인 판단의 사례로 분석한다. 이 인용은 작가에 대한 평가를 인용한 것으로 볼 수 있다. <표 2>의 5번 항목은 오형엽의 「박용철 시론의 구조와 계보」에서 발췌한 내용으로, 이 연구는 이 인용을 통해 박용철의 비평이 작품의 가치 판단보다 해석에 가까운 인상주의 비평의 한계에 머물렀으며, 작품 구조를 체계적으로 분석할 능력을 보여주지 못했다는 부정적 해석을 도출하는 근거로 활용된다. 이는 작가를 인용/평가하는 박용철의 관점과 더불어 박용철의 비평을 인상주의 비평으로 평가하는 인용을 포함하고 있다.

이 사례를 바탕으로 도출한 XML 마크업 요소들은 다음과 같다. 먼저 ‘작가의 이름과 작가의 역할’이다. 이때 역할은 비평가/소설가/시인 등의 역할을 말하며, 작가의 아호(阿呼)나 필명 등은 동일한 인물로 인식할 수

www.kci.go.kr

있는 작업이 필요하다. 두 번째는 작품명과 장르다. 이때 장르는 문학사의 기준에 따라 구분할 수 있다. 세 번째는 ‘인용’이다. 작가 또는 작품의 인용이다. 그 인용의 형태는 직접인용과 간접인용으로 구분할 수 있다. 네 번째는 ‘개념/용어’이다. ‘개념/용어’를 더욱 세분화한다면 분류의 기준이 모호해질 수 있기 때문에, 데이터를 구축하는 단계에서 세분하는 기준을 마련하기는 어려워 보인다. 마지막으로 가장 중요한 요소는 ‘평가’이다. 대상에 대한 평가는 긍정/부정/중립으로 나누어 볼 수 있다.

예컨대 김수영 연구에서 ‘개념어’ 태깅이 200회로 가장 높게 나타난 것은 김수영 비평 연구가 ‘자유’, ‘참여’ 등 개념적 논의에 집중되어 왔음을 보여준다. 반면 김춘수 연구에서 ‘작품명’ 태깅이 103회로 상대적으로 높은 것은 개별 작품 분석 중심의 연구 경향을 반영한다.

〈표 3〉 Claude를 활용한 태깅 요소 빈도

논문 대상	태깅요소					
	인물명	작품명	기관명	개념어	인용	해석/평가
김기림	81회	19회	0회	28회	5회	23회
임화	47회	15회	8회	156회	48회	67회
박용철	149회	21회	0회	123회	11회	26회
김수영	195회	61회	9회	200회	3회	52회
김춘수	147회	103회	12회	34회	35회	52회
김윤식	214회	72회	34회	223회	35회	47회
김현	46회	23회	0회	59회	13회	37회

위의 표는 Claude의 프로젝트 기능과 Claude Code를 활용하여¹⁷⁾, 마크

17) 본 연구에서는 생성형 AI 모델 중 Anthropic의 Claude를 활용하였다. Claude를

업 규칙을 제시한 뒤에 선행연구들의 서술을 태깅하도록 명령하고 그 결과를 정리한 내용이다.¹⁸⁾ 이를 통해 제시된 마크업 요소의 빈도는 향후 제시할 마크업 규칙이 선행연구의 서술을 통해 충분히 검증 가능한 요소임을 시사한다.

2. 현대문학비평 XML 데이터 구축을 위한 마크업 규칙

현대문학비평 텍스트는 작가, 작품, 인용, 평가, 시대 구분, 서지 정보 등을 포함하기 때문에, 단순히 원문을 디지털화하는 것만으로 분석할 수 없다. 따라서 본 연구는 국제표준을 지향하는 TEI(Text Encoding Initiative) P5 가이드라인¹⁹⁾을 기반으로 XML데이터를 구축하되, 연구자가 추후 연구할 연구의 방향을 마크업에 반영하기 위해 맞춤형 XSD²⁰⁾를 설계하였다. 이

선택한 이유는 다음과 같다. 첫째, Claude는 기본 20만 토큰의 컨텍스트 윈도우를 지원하여 장문의 비평 텍스트 전체를 한 번에 처리할 수 있다. Google Gemini(최대 200만 토큰)는 컨텍스트 윈도우는 더 크지만 지시 수행의 정확도와 텍스트의 뉘앙스 파악에서 Claude가 더 적합하다고 판단하였다. 둘째, Claude는 연구자가 제시한 마크업 규칙을 일관되게 따르는 능력이 우수하여 XML 태깅 작업에 적합하였다.

18) 이와 관련된 데이터는 <https://github.com/ByeongjooLee/xmldata>에 공개되어 있다. KCI/RISS와 같은 학술정보사이트를 통해 접근가능한 자료이므로, 연구의 재현가능성을 위해 분석 데이터를 공개하였으나, 이후 저작권에 관한 문제가 제기될 경우 비공개로 전환될 수 있다. Anthropic, Claude, LLM(대규모 언어 모델), <https://claude.ai> (최종접속일 : 2025. 11. 29.).

19) TEI P5(Guidelines for Electronic Text Encoding and Interchange)는 인문학 및 디지털 텍스트 연구 전반에서 전자문서를 구조화하기 위한 국제 표준 지침으로, 텍스트의 구조적·의미적 특징을 명시적으로 기술(markup) 하여 학문 간 데이터 교환을 가능하게 하는 것을 목표로 한다. https://tei-c.org/release/doc/tei-p5-doc/en/html/index.html?utm_source=chatgpt.com (최종검색일 2025.11.29.); 김바로 외 8명, 『XML(TEI) with 인문학』, 위키독스, 2025, <https://wikidocs.net/book/14569> 참조.

20) 스키마는 데이터 설계도를 의미하며, XSD는 XML문서를 검증할 때 가장 널리 쓰이는 표준화된 스키마 언어이다. 이와 관련된 데이터는 <https://github.com/ByeongjooLee/xmldata>에 ‘tei-criticism-schema.xsd’로 공개되어 있다.

www.kci.go.kr

스키마는 TEI의 표준 구조(<teiHeader>, <text>, <front>, <body>, <back>)를 유지하면서도, 문학비평 텍스트의 핵심 구성요소를 표현하기 위한 세분화된 요소 정의와 속성 체계를 추가하였다. XML 문서 데이터는 크게 서지 정보를 관리하는 <teiHeader> 부분과 비평텍스트를 관리하는 <text> 부분으로 나누어져 있다. 구체적인 마크업 규칙의 원리는 다음과 같다.

1) 서지 정보

〈표 4〉 한국현대문학비평 XML 데이터 마크업 요소 서지 정보 부분

범주	요소(타입명)	기능 및 목적	주요 속성	설명 및 예시
서지정보 (단순)	<bibl> (BibSimpleType)	간단한 출처 표기	type (journal, coterie 등)	<bibl type="journal"> <title> 현대문학 </title> </bibl>
서지정보 (구조)	<biblStruct> (BiblStructType)	구조화된 출처 정보	type (journal, newspaper 등)	내부에 <analytic> + <monogr> 구성
저자 정보	<author> (AuthorMixedType)	저자 표기	—	<author> <persName> 김우창 </persName> <persName> 김윤식 </persName> </author>
발행 정보	<imprint> (ImprintType)	출판지, 출판사, 발행일	pubPlace, publisher, date(when)	<imprint> <pubPlace> 서울 </pubPlace> <publisher> 현대문학사 </publisher> <date when="1959"/> </imprint>

서지정보와 관련된 마크업 요소들은 TEI의 기본구조를 따르고 있다. 이 때 <bibl>은 단순한 서지 사항을 말하며, <biblStruct>는 구조화된 서지 사항을 말한다. 이 둘의 차이는 간단한 인용과 자세한 서지정보의 차이이다. <bibl>은 각주나 참고문헌 리스트에서 간단하게 표기하는 방식이며, <biblStruct>는 개별 요소들(작가, 글제목, 출판사 등)을 마크업하여 더욱

자세한 정보를 포함하고 있다. 이 경우 추후 서지정보를 정리할 때 유용하게 활용될 수 있다.

2) 글의 구조

〈표 5〉 한국현대문학비평 XML 데이터 태깅 요소 글의 구조 부분

범주	요소(타입명)	기능 및 목적	주요 속성	설명 및 예시
최상위 구조	<text>	비평 텍스트 전체	xml:id	<front> —<body> —<back>로 구성됨.
서두	<front>	서두		본문 시작 전의 초록, 목차, 서문, 일러두기 등을 포함하는 영역.
본문	<body>	비평 내용		실제 비평적 언술을 포함하는 영역
후미	<back>	텍스트의 부수 정보		본 참고문헌, 후기, 부록, 색인 등을 포함하는 영역
구획 단위	<div type= " " > (DivTypeEnum)	장·절·항 등 구분 및 논리적 성격 구분	type = contents / introduction / body / conclusion / section / argument / criticism / review / editorial / column / interview / notes	비평문 내 구조적 단위 구분. 예: <div type="criticism">
단락 단위	<p> (ParagraphType)	문단 표현	—	내부에 <s>문장 요소 포함
문장 단위	<s> (SentenceType)	문장 단위 표현		내부에 인명· 용어·인용·평가 등 포함

글의 구조와 관련된 마크업 요소들도 TEI의 기본구조를 따르고 있다. 문학비평의 내용을 다루고 있는 <text>는 <front>-<body>-<back>으로 3분할 되어 있으며, 각각 머리말-본문-주석 등으로 구성되는데, <front>와 <back>은 부수적인 영역으로, 글 또는 저서의 메타정보를 포함하는 구역이라고 할 수 있다. 초록, 목차, 참고문헌 등이 이에 해당한다.²¹⁾ 실제 비평적 언술은 <body>에 포함된다. <div>는 장·절·항 등 글의 구획을 구분하는 역할을 하며, type 속성을 통해 해당 구획의 논리적 성격을 명시한다. 예컨대 <div type="introduction">은 서론에 해당하는 구획임을 표시한다. 각 <div>는 상·하위 관계를 구성하여 계층 구조를 이룬다. <div>의 내부에는 단락을 나타내는 <p>가 포함되어 있으며, <p>는 다시 문장을 나타내는 <s>의 집합으로 구성된다. <s>를 별도로 구성한 이유는 문학비평은 문장 단위의 분석이 이루어지는 장르이며, 더 작게는 어휘 단위의 분석이 필요하기 때문이다. TEI P5에 따르면 시는 <div>(시전체)-<lg>(연)-<lh>(행)으로 구분하며, 소설은 <div>(장·절구분)-<p>(서사 단락)-<sp>,<speaker>,<stage>(대화, 지문 등)으로 구성된다.

3) 내용 마크업

〈표 6〉 한국현대문학비평 XML 데이터 태깅 요소 내용 부분

범주	요소(타입명)	기능 및 목적	주요 속성	설명 및 예시
인명 표기	<persName> (PersNameType)	작가·비평가 등 인명 표기	role (critic, poet, scholar 등) / xml:id / ref	<persName role="poet critic"> 김수영 </persName>

21) ‘서두’와 ‘후미’는 학술논문의 규범적 형식을 전제한 것이 아니라, 텍스트의 시작부와 종결부를 형식적으로 구분하기 위한 표지이다. 비평 텍스트의 성격에 따라 해당 요소의 적용 여부는 유연하게 판단한다.

작품 · 제목	<title> (TitleInlineType)	작품 · 논문 · 잡지명 등	level (m/a/j), type (critic, poem 등)	<title level="a" type="critic">시와 이론의 각도</title>
용어 정의	<term> (TermType)	개념 · 사조 · 비평어 등	type ("concept", "movement") / xml:id	<term type="concept">모더 니타</term>
인용 표시	<quote> (QuoteType)	비평문 · 작품 인용	type (direct, indirect 등) / genre (critic, poem 등) / source	<quote type="direct" genre="critic"> “시란 자기 존재의 불안이다.”</quote>
주석 처리	<note> (NoteType)	각주, 편집주석, 참고 설명	type / target / ana	<note type= "editorial"> 편집자 주 </note>
참조 연결	<ref> (RefType)	외부 링크, 교차 참조	target (ISNI, URI 등)	<ref target= "https://ko.wikipedia.o rg/wiki/백철"> 백철 </ref>
평가 · 해석	<interp> (InterpType)	비평적 평가 · 태도 명시	value (affirmative, neutral, critical) / ana (분석범주)	<interp value="critical" ana="이론한계">...< /interp>

내용 마크업을 다루고 있는 부분이 본고가 지향하는 연구방향을 포함하는 마크업 규칙이라고 할 수 있다. 먼저 인명 <persName>에 대한 표기는 role을 통해 주체의 속성을 구분하여, 작가-비평가-번역가 등 역할에 따른 관계를 구분할 수 있도록 하였다. 작가-비평가-번역가 등 구분이 모호할 경우, 활동한 영역을 모두 표기하는 것으로 통일하고자 한다. 또한 xml:id라는 개체별로 1개만 부여할 수 있는 고유한 식별값을 통해 다양한 이명(異名)들을 하나로 통일할 수 있다. xml:id는 ISNI 값으로 통일하고자 한다. ISNI값은 전세계 공통으로 사용하고 있는 값이며, 이를 통해 위키와 같은 다른 사이트와 연결성을 이미 확보해놓은 값이기 때문이다.²²⁾

22) 동명이인을 식별하고 작가 정보의 정확성을 기하기 위해, 한국학중앙연구원의

다음으로 제목 요소 <title>은 @level 속성과 @type 속성을 통해 문헌의 서지적 성격과 장르를 분류할 수 있도록 설계하였다. 용어 요소 <term>의 경우, 비평 텍스트에 등장하는 모든 개념을 체계적으로 구분하기는 어려우나, 문예사조나 비평 용어 등 범주화가 가능한 핵심 개념들을 지정할 수 있도록 세부 항목을 구성하였다.

인용 요소 <quote>는 비평문 내부에서 타 텍스트를 수용하고 재맥락화하는 방식을 구조적으로 표현하기 위해 도입하였다. 비평의 언표 행위는 타자의 언표를 매개로 구성되는 경우가 많으며, 이러한 인용의 양상을 명시적으로 기록하는 것은 비평 담론의 형성 과정을 추적하는 데 중요한 의미를 지닌다. 이에 <quote>에는 인용의 형식을 구분하기 위한 @type 속성(직접인용 direct, 간접인용 indirect 등)을 부여하고, 인용 대상의 장르를 명시하기 위한 @genre 속성(critic, poem, novel 등)을 추가하였다. 아울러 인용의 출처를 기록하는 @source 속성을 함께 구성하여, 비평가 간 상호 참조 관계를 분석할 수 있는 토대를 마련하였다.

주석 요소 <note>는 원문 내 각주, 편집자 주석, 연구자의 보조 설명 등을 구분하여 표시할 수 있도록 설계하였다. 참조 요소 <ref>는 외부 데이터베이스 및 식별 체계와의 연결을 담당한다. 이는 본 연구가 지향하는 데이터 간 상호 연결을 구현하는 핵심 요소로서, @target 속성을 활용하여 ISNI, Wikidata, 한국민족문화대백과사전 등 외부 데이터를 참조할 수 있도록 하였다.

〈표 7〉 백철의 『조선신문학사조사』에 인용된 ‘임화’

<p> <s>이 학설의 계통을 이어 <persName

한국민족문화대백과사전과 Wikipedia의 오픈 API를 활용하였다.

role="critic" xml:id="isni-0000000084484348" ref="https://encykorea.aks.ac.kr/Article/E0047762">

임화</persName>는

<title type="critic">신문학사</title><note type="footnote" rget="fn2"(*2)</note> 서론
중에서 한국의 근대적인 특수성을 다음과 같이 말하였다.</s>

</p>

예를 들어 백철 『조선신문학사조사』의 한 문단을 살펴보면, <p>는 문단 구분이며, 그 문단 안에 있는 <s>는 문장을 의미한다. 위의 표에 나온 문장을 XML 태그를 제외하고 살펴보면 “이 학설의 계통을 이어 임화(林和)는 신문학사(新文學史)²⁾ 서론(序論) 중에서 한국의 근세적인 특수성을 다음과 같이 말하였다.”이다. 이를 XML데이터 구조로 살펴보면 <persName role="critic" xml:id="isni-0000000084484348" ref="https://encykorea.aks.ac.kr/Article/E0047762">임화</persName>와 같이 나오는데, 그 상세 내용을 설명하면 <persName>이라는 마크업을 통해 ‘임화’가 사람의 이름이라는 것을 표시하고, 그 <persName> 안의 하위 요소로 role을 통해 ‘임화’가 ‘비평가’임을 지정해준다. 또한 ‘임화’는 ‘임인식, 김철우’ 등과 같은 이칭이 존재하므로, 앞서 언급한 것처럼 고유한 식별 값인 xml:id를 통해 다른 이칭이 나오더라도 이를 ‘임화’와 같은 인물로 인식할 수 있도록 ISNI 값을 사용하여 고유값을 부여하였다. 또한 ref를 통해 링크를 연결하여, 한국학중앙연구원에서 구축해놓은 한국민족문화대백과사전과 연결하여, 인물 데이터를 쉽게 살펴볼 수 있도록 구축하였다. 이와 같은 표기는 해당 인물의 외부 식별자와 XML 내부의 연표를 연결함으로써, TEI 문서가 독립적 텍스트를 넘어 다른 지식 데이터베이스와 연결 가능한 데이터 구조로 기능하도록 한다.

끝으로 비평가의 가치 판단을 마크업하기 위한 요소인 <interp>는 본고에서 특히 중요하게 생각하는 부분이다. 비평가의 판단을 긍정·부정·중립으로 표시할 수 있도록 하였으며, 이는 다른 마크업 방식과는 다르게 연

구자의 ‘해석’이 반영되어야 하는 영역이다. 이는 ‘해석’의 주관을 고려하여 @value 속성을 통해 평가 방향을 표시할 수 있도록하고, @ana 속성을 통해 해석의 근거를 병기하여 다른 연구자들에게 설명할 수 있도록 구성하였다.²³⁾

〈표 8〉 백철의 『조선신문학사조사』에 포함된 가치판단

```
<p>
<s>이렇게 보면 우리나라에
<term type="concept">신문학</term>이
<term type="concept">창가</term>나
<term type="concept">신소설</term>의 형식으로 나오기 전에, 일종의
<term type="concept">번역행위</term>가 앞서서
<interp value="affirmative" ana="번역이 신문학 형성의 준비 과정이었다는 긍정적
평가">그 직접 준비의 과정으로 선행한 것을 짐작할 수 있는 것이다.</interp></s>
</p>
```

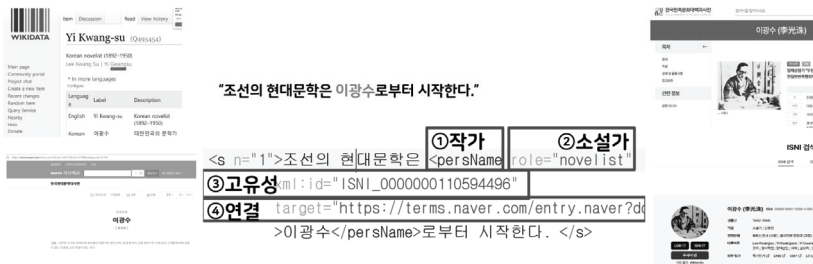
위의 문장은 “이렇게 보면 우리나라에 신문학(新文學)이 창가(唱歌)나 신소설(新小說)의 형식으로 나오기 전에, 일종의 번역행위가 앞서서 그 직접 준비의 과정으로 선행한 것을 짐작할 수 있는 것이다.”라는 문장을 XML구조로 데이터화한 형태이다. <term>이란 개념을 마크업하는 태그를 통해 ‘창가’, ‘신소설’, ‘번역행위’ 등이 정리되어 있으며, 주목할 것은 <interp>라는 마크업이다. 번역을 ‘신문학’의 선행조건으로 판단하는 백철의 평가를 ‘긍정적’평가라고 표시해 둔 것이다. 이때 @value를 긍정적으로 마크하고, @ana를 통해 연구자가 해당 문장을 왜 긍정적으로 평가했는지

23) 이때, 문학비평의 가치 판단을 긍정/부정/중립으로 단순화하여 마크업하는 것이 비평의 다층적인 양상을 포괄하지 못할 수 있다는 문제가 있을 수 있다. 하지만 본 연구에서 제안하는 긍정/부정/중립의 평가 범주는 XML 마크업 단계에서의 기초 태깅에 해당한다. 문학비평이 포함하는 원인-결과 설명, 개념 간 관계 설정, 역사적 맥락화 등 다층적 해석의 세분화는 이 데이터를 기반으로 한 후속 연구, 특히 RDF 구성이나 온톨로지 설계 단계에서 수행될 수 있다. 본 연구는 그러한 확장을 위한 토대를 마련하는 데 목적이 있다.

근거를 제시할 수 있다.

Ⅲ. 현대문학비평 XML 데이터의 활용 방안

데이터를 XML로 구축하는 것은 데이터 보존과 데이터 구조화 그리고 데이터 변환 및 공유를 목적으로 한다. 1차적으로 보존을 목표로 제작된 XML 데이터는 2차적으로는 데이터 구축한 사람의 의도에 따라 구조화된 것이다. 본고는 2장 2절을 통해서 연구자가 의도한 구조화를 설명하였다. 우선 기존의 데이터와의 연결성을 확보하기 위해, ISNI, Wiki, 한국민족문화대백과사전, 네이버지식백과 간의 데이터 연결을 시도하였다.



[그림 1] 이광수 데이터 연결

데이터와 문학비평이라는 텍스트의 특성을 고려하여, ‘인용<quote>’ 및 ‘평가<interp>’ 요소를 설계에 반영하였으며, 추후 개념사 연구를 위해 ‘개념<term>’ 요소도 추가하였다. 이렇게 마크업 된 데이터를 활용하는 방식은 연구자의 연구 의도에 따라 다양하게 활용될 수 있다.

XML 데이터를 분석할 때, 기본적인 방법은 XPath(XML Path Language)를 활용하는 것이다. XPath는 XML 문서에서 원하는 정보를 찾아내기 위한 경로식이다. 이 경로식을 활용하면, 우리가 흔히 사용하는 문서에서

‘비평’과 같은 특정 단어를 찾는 기본 기능을 포함하여, 더 나아가 특정 단어가 등장하는 맥락을 고려하여 찾을 수도 있다. 특정 어휘를 찾을 수 있다는 것은, 해당 어휘의 계열적 관계를 파악할 수 있다는 것이다. 예를 들어 백철의 『조선신문학사조사』에서 「신문학 태동기」에서 ‘긍정적’인 평가를 추출해 볼 수 있을 것이다. 이때 ‘긍정적’ 평가로 추출된 문장은 총 15개이며, 그 일부를 살펴보면 다음과 같다.

〈표 9〉 백철의 『조선신문학사조사』에 나타난 ‘긍정적 평가’ 일부

문장	의미
따라서 그 하나하나의 사조가 역시 시대와 현실의 특색을 띠고서 각각 등장한 것이며, 우리가 신문학(新文學)의 발전을 보는 데서 그 사조의 의미를 구하여 여기에 신문학운동(新文學運動)의 사실(史實)을 파악하려고 하는 것이다.	사조사적 접근의 유효성에 대한 긍정적 평가
여기에 우선 우리는 신시(新詩)의 최초의 작가인 최남선(崔南善)이 동시에 창가(唱歌)의 대표적 작사자(作詞者)였다는 사실에 주목을 해 두기로 하자. 최남선이 창가와 신시의 교차점에 위치한다는 중요한 문학사적 사실 지적	최남선에 대한 평가

이와 같은 ‘긍정적’ 평가에 대한 추출은 ‘개념사’ 또는 ‘비평 어휘’와도 연결될 수 있다. 예를 들어 ‘신소설’ 장르/개념과 ‘개화기’라는 시대적 특징이 어떠한 평가 및 관계를 보여주고 있는지도 확인이 가능할 것이다. 더 나아가 XQuery를 활용하면, XPath의 단일 경로 탐색을 넘어서 여러 개의 문서 간의 비교도 가능하다. 예를 들어 백철, 임화, 박영희의 문학사 데이터를 동일 스키마로 구축했을 경우, 세 비평가가 ‘최남선’을 언급할 때의 평가를 비교할 수 있다. 박영희의 문학의 경우, 동시대적인 증언들이 주를 이루며, 동인지 중심 서술로 구성되어 있는 반면, 백철의 경우 역사적 거리가 확보된 상태에서 문예사조 혹은 외래 사조적 접근을 하고 있다는 것을 확인할 수 있다.²⁴⁾

더 나아가 XML로 이루어진 데이터는 단순한 데이터 추출을 넘어서, 비

이회승 ~	3	30
임석재 ~	1	9
임현도 ~	1	1
장덕순 ~	7	10
전광용 ~	4	29
정래동 ~	1	1
정병욱 ~	5	19
정인승 ~	1	3
정한모 ~	1	10
조연현 ~	1	2
조문계 ~	1	2
조항 ~	3	5
진성기 ~	2	1
최승범 ~	1	5
최현배 ~	2	3
허웅 ~	7	8
홍용선 ~	1	1
홍종인 ~	1	5
	국어국문학_기고횟수	사상계_기고횟수 기고자

[그림 3] 『국어국문학』-『사상계』 기고 횟수 일부

또한 한국문학비평을 XML데이터로 구축하는 것은 연구의 효율성을 확보하는 일이다. 조선왕조실록이 왕과 신하들의 발언을 날짜별로 정리하여 기록한 데이터베이스라면, 문학비평의 XML 데이터베이스는 비평 어휘, 개념어, 평가 등을 기록한 데이터베이스가 될 수 있다. 특히 이미 구축된 한국민족문화대백과사전, 네이버지식백과, 한국사데이터베이스 등과 연동하여 데이터를 구축한다면 이는 ‘개념사’를 효과적으로 연구할 수 있는 계기가 될 것이다. 더불어 이 체계에서는 비평가가 사용하는 특정 비평 어휘나 개념어를 추출하여 ‘단순 사용-문맥-시대 또는 매체’ 등의 관계망을 보여줄 수 있을 것이다.

본고가 XML 데이터 구축 방안을 제시한 후 꿈꾸는 현대문학비평 데이터 구축의 미래는 해외의 디지털 인문 프로젝트인 ‘데리다 프로젝트’의 모습이다. 이 프로젝트는 프린스턴 대학교 도서관 주관으로 자크 데리다의

개인 장서의 메모나 밑줄을 TEI 표준을 기준으로 XML 데이터화하여, 웹 상에서 시각화하였다. 이는 데이터가 단순히 보존되는 것을 넘어, 한 사상가의 지적 탐구와 사유의 과정을 추적하는 도구가 된 사례라고 할 수 있다.²⁶⁾

본고에서 제안한 현대문학비평 XML 데이터 구축 방안을 통해 김윤식, 김현, 김우창, 유종호 등 한국 현대비평을 이끈 거장들의 비평텍스트를 데이터화하고 구조화하는 작업을 진행한다면, 이는 단순한 아카이빙을 넘어 대가(大家)라고 불리는 비평가들의 사상적 흐름과 비평적 전략을 파악할 수 있는 새로운 방법이 될 것으로 기대한다. 이는 계보학적 측면에서 접근하거나 시대적 흐름을 파악할 수도 있는 중요한 데이터가 될 것이다. 이러한 활용 방안의 실제 사례를 김우창의 비평문 분석을 통해 제시하면 다음과 같다.²⁷⁾

〈표 10〉 2회 이상 인용 15인

순위	사상가	인용 횟수	인용된 글
1	데카르트	6	한국 현대소설의 형성, 한용운의 소설, 운동주의의 시, 김수영론, 정현종론, 나와 우리
2	하이데거	5	정현종론, 세계와 문학의 세계, 나와 우리, 물음에 대하여, 하버마스 비판이론
3	사르트르	4	궁핍한 시대의 시인, 나와 우리, 물음에 대하여, 하버마스 비판이론

26) Derrida’s Margins이라는 사이트로, 데리다가 읽은 서적을 XML데이터뿐만 아니라 이미지 파일로 게시하여 연구자의 접근성을 확보하였다. <https://derridas-margins.princeton.edu/library/>(최종검색일 2025.11.29.)

27) 김우창 전집 1권인 『궁핍한 시대의 시인』의 수록된 비평을 XML 데이터로 구축한 뒤, Claude를 통해 분석하였다. 다만 김우창의 비평은 저작권이 있는 글이므로, 분석의 내용만 공개하였다. 해당 분석의 내용은 <https://github.com/ByeongjooLee/korean-critique-schema>에 공유되어 있다.

4	T.S. 엘리엇	4	한국시와 형이상, 금강, 소설가 구보씨, 김종길 시론
5	메를로퐁티	4	한용운의 소설, 말과 현실, 나와 우리, 세계와 문학의 세계
6	루카치	3	한국 현대소설의 형성, 한용운의 소설, 나와 우리
7	골드만	3	궁핍한 시대의 시인, 나와우리, 김윤식 비평사
8	하버마스	2	물음에 대하여, 하버마스 비판이론
9	한나 아렌트	2	나와 우리, 하버마스 비판이론
10	발레리	2	정현종론, 김종길 시론
11	막스 베버	2	한용운의 소설, 진실과 언어(김종길)
12	마르쿠제	2	물음에 대하여, 하버마스 비판이론
13	헤겔	2	김수영론, 하버마스비판이론
14	니체	2	정현종론, 나와 우리
15	플라톤	2	세계와 문학의세계, 나와 우리

김우창의 비평은 개별 작품의 분석에 그치지 않고, 광범위한 서양 지성사를 바탕으로 한국 문학을 보편적인 세계사적 맥락 위에서 해명하려 했다는 점에서 그 의의가 크다. 『궁핍한 시대의 시인』에 수록된 22편의 텍스트에서 식별된 서양 사상가는 총 60여 명에 달하며, 인용 빈도 상위 15인은 <표 1>과 같다. 가장 주목할 점은 데카르트가 최다 인용되었다는 사실이다. 이어 하이데거, 사르트르, 메를로퐁티 등 현상학 및 실존주의 계열의 철학자들이 상위권을 점유하고 있으며, 엘리엇과 같은 모더니즘 비평가와 루카치, 골드만 등의 사회학적 비평가들이 뒤를 잇고 있다.

1970년대 한국 비평계가 순수문학론과 참여문학론으로 양분되어 대립할 때, 김우창은 이 두 흐름을 지양하고자 했던 것으로 해석할 수 있을 것이다.



[그림 4] 비평문과 인용된 사상가의 관계망

[그림 4]는 김우창이 『공핍한 시대의 시인』에 수록된 비평문에서 인용한 사상가들을 네트워크로 시각화한 것이다.²⁸⁾ 데카르트와 하이데거와 같

28) 해당 네트워크 시각화는 Gephi 0.10.1에서 ForceAtlas 2 레이아웃 알고리즘을 적용하여 제작되었다. 주요 파라미터는 Scaling 50.0, Gravity 5.0으로 설정하여 100개 노드와 132개 엣지로 구성된 네트워크의 가독성을 확보하였으며, Dissuade Hubs 옵션을 활성화하여 비평에 활용된 철학자의 인용 관계를 명확히 드러내게 하였다. 노드 크기는 진입 차수(In-Degree)를 기준으로 1~10 범위로 설정하여 인용 빈도가 높은 철학자(데카르트, 하이데거 등)가 시각적으로 강조되도록 하였다.

이 인용빈도가 높은 사상가의 경우 노드를 크게 하여, 김우창 비평에서 해당 사상가가 차지하는 비중을 가늠하게 하였으며, 데이터를 구성할 때, 인용문을 함께 구성하여, 인용문을 확인하여 추후 연구에 활용할 수 있도록 구성하였다.

〈표 12〉 Edges의 데이터 구성 일부

target	source/essay	quote
골드만	궁핍한시대의시인	프랑스의 철학자 루시앵 골드만은, 그의 저서 『숨어 있는 신(神)』에서, 어려운 시대에 사는 인간의 한 유형을 파스칼과 라신의 생애와 저작을 통해서 추출하여 보여 준다.
파스칼	궁핍한시대의시인	파스칼은 ‘진정한 도덕가는 도덕을 싫어한다.’고 했지만, 이 말은 한용운의 경우에도 해당된다.
사르트르	궁핍한시대의시인	욕정은 부재나 마찬가지로 인간 존재의 부정성 — 사르트르 식으로 말하여, 인간 존재의 본질이 ‘결여(manque)’라는 사실에 그 존재론적 근거를 갖는다.

김우창의 사례는 프랑코 모레티가 제안한 ‘멀리서 읽기’의 방법론을 넘어, XML 기반 데이터 구축이 비평가의 지적 지형도를 정밀하게 재구성하는 데 유효한 도구임을 보여준다. 비평 텍스트 내에 산재한 인용 정보와 이론적 개념들을 구조화하고 네트워크로 시각화함으로써, 개별 텍스트의 정독만으로는 포착하기 어려운 비평가의 사유 구조와 지적 참조 체계를 조망할 수 있게 된다.

본 연구에서 제시한 XML 데이터 분석 모델은 다음과 같은 후속 연구로 확장될 수 있다는 점에서 방법론적 함의를 지닌다. 첫째, 비평가 간 영향 관계 및 이론 수용 양상의 규명이다. 구축된 데이터를 통해 특정 이론이나 사상가가 시기별로 어떤 비평가에게 수용되고 변용되었는지를 추적할 수 있으며, 이는 한국 비평사 내부의 상호텍스트성과 지적 계보를 실증적으로 해명하는 데 기여할 것이다. 둘째, 문학 정전 형성 과정의 실증적 분석

이다. 본고의 분석 대상에서는 제외되었으나, 구축된 XML 데이터에는 비평가가 특정 작가나 작품에 대해 내린 평가 정보가 포함되어 있다. 비평가가 어떤 작가를 옹호하고 어떤 작품을 배제했는가에 관한 데이터는 한국 문학사에서 정전이 형성되고 위계화되는 과정을 파악하는 토대가 될 수 있다. 결론적으로, XML 데이터를 활용한 비평 연구는 기존의 인상비평이 지닌 주관성을 보완하고, 현대문학비평의 지형을 데이터에 기반하여 재구성하는 새로운 연구 방법론을 제시한다는 점에서 의의가 있다.

IV. 결론

본 연구는 국문학을 연구하는 연구자로서 텍스트를 면밀히 읽는 과정에서 제기되는 근본적인 의문으로 시작하였다. 국한문혼용체로 작성된 텍스트들은 전집이 출간되기 이전까지 한글이 병기되지 않은 텍스트로 존재한다. 이러한 형태는 연구자마다 한자가 병기된 문장을 다시 읽게 하고, 오식(誤植)의 가능성을 항상 걱정해야 하는 문제를 야기한다. 그 대표적인 예가 『사상계』이다. 『사상계』는 1950-60년대 한국문학과 지성사를 연구하는 연구자들에게 통과의례와도 같은 텍스트이다. 『사상계』 원본은 국립중앙도서관이나 동방미디어와 같은 웹사이트를 통해 PDF 형태로 제공되지만, 세로쓰기와 국한문혼용체 그리고 낮은 해상도로 인해 연구자가 접근하기 어려운 텍스트이기도 하다. 결국 연구자는 스스로 문장을 다시 입력하고, 대조하며 텍스트의 진위를 판단하여야 한다.

이와 같은 과정은 연구자에게 당연하게 여겨지는 과정이지만, 여전히 의문은 해결되지 않았다. 국문학 연구가 시작된 이래로 『사상계』를 읽는 연구자들은 왜 항상 첫발자국부터 시작해야 했을까. 이와 같은 경험에서 비롯한 본고의 문제의식은 다음 세 가지로 수렴된다.

www.kci.go.kr

먼저 “왜 디지털인가”하는 문제이다. 수많은 연구자가 반복적으로 수행하는 텍스트의 복원, 교정, 해석의 과정을 더 이상 반복하지 않을 수 있는 방법은 없을까 하는 물음에서 출발하였다. 원전 텍스트를 직접 입력하고 교정하는 과정은 연구자의 개인 노동에 의존하고 있으며, 힘겹게 만들어진 결과물은 쉽게 공유되기 어렵다. 공유된다고 하더라도 동일한 텍스트가 연구자마다 다른 방식으로 재입력·교정되는 현실은 학문의 양적·질적 축적을 방해하고 있는 요소이다. 따라서 본고는 이러한 비효율적 구조를 극복하기 위해 데이터화하여 연구자들에게 공개하는 것이 필요하다고 판단했다.

두 번째로 “어떻게 디지털화할 것인가”에 대한 문제이다. 국문학 연구자들이 대부분 그렇듯이, 연구할 자료를 정리한다는 것은 HWP파일 형식으로 타이핑하는 일이다. 하지만 이와 같은 HWP파일 형식의 저장은 연구자의 컴퓨터에만 머무르는 폐쇄적 문서구조로 기계가 읽거나, 웹사이트에 게재할 수 있는 형태가 아니다. 따라서 HWP파일 형식을 뛰어넘는 기계와 인간이 모두 공유할 수 있는 형식이 필요하다고 판단하였다. 이를 해결하기 위해서는 가장 먼저 텍스트가 XML 데이터의 형태로 구축되어야 한다.

마지막으로 “무엇을 연결할 것인가”에 대한 문제이다. 이렇게 개인 연구자가 구축한 텍스트 데이터들은 기존에 국가 단위로 구축된 데이터베이스와 어떻게 연결될 수 있는가가 중요한 과제로 남아 있다. 지난 20여 년간 국가 차원에서 진행된 인문학 데이터베이스 구축사업은 양적 측면에서 방대하지만, 서로의 호환가능성은 거의 고려되지 않았다.²⁹⁾ 앞으로 활성화될 개별 연구자의 데이터 구축이 이러한 국가 단위 인프라와 단절된다면, 또 다른 데이터 섬을 양산하게 될 뿐이다. 이를 해결하기 위해 본 연구

29) 이는 국사편찬위원회, 한국고전번역원, 한국학중앙연구원, 성균관대학교 등 「고문헌 XML(TEI)설계 검토」를 통해서도 알 수 있다. 김바로 외 8명, 『XML(TEI) with 인문학』, 위키독스, 2025, <https://wikidocs.net/book/14569> 참조.

는 TEI라는 국제표준을 지향하는 문서형태의 형식을 따르면서 연구자가 필요하다고 판단한 마크업 요소들을 추가하였다. 추가적으로 ISNI, Wikidata, 민족문화대백과사전 등의 외부 데이터와의 상호 연결 가능한 구조를 설계하였다.

요컨대 이 세 가지 문제의식은 각각 연구의 필요성, 기술적 실천, 지식 생태계의 연속성을 가리킨다. 본 연구가 구축한 현대문학비평 XML 데이터는 이러한 문제의식이 단계적으로 구체화된 결과이며, 더 나아가 개별 연구자가 수행하는 텍스트 해석의 과정을 공유 가능한 데이터 구조로 기록하려는 시도라 할 수 있다. 다만 본 연구에서 제안한 평가(공정/부정/중립)의 삼분 범주는 XML 마크업 단계의 기초 태깅으로서, 문학비평이 지닌 다층적 해석의 복합성을 온전히 포착하기에는 한계가 있다. 원인-결과 설명, 개념 간 관계 설정, 역사적 맥락화 등 비평 언술의 세분화된 유형 분류는 본 연구의 데이터를 기반으로 한 후속 연구, 특히 RDF 구성이나 온톨로지 설계 단계에서 수행될 과제로 남겨둔다.

• 참고문헌

- 강용훈, 『비평적 글쓰기의 계보 — 한국 근대 문예비평의 형성과 분화』, 소명출판, 2013.
- 국립중앙도서관, 「국립중앙도서관 연구정보서비스」,
<https://nl.go.kr/rink/contents/R10400000000.do?schM=view&id=45062>
(2025.10.01.).
- 김바로, 「<공공데이터법>과 인문데이터 : 공공기관 보유 인문데이터 공개 신청 사례를 중심으로」, 『한국고전연구』 제57집, 한국고전연구학회, 2022, 145-180쪽.
- 김바로 외 8명, 『XML(TEI) with 인문학』, 위키독스, 2025.
<https://wikidocs.net/book/14569>.
- 김서운, 「시맨틱 DB 구축을 위한 『논어』 주석 XML 데이터 설계」, 『한문학논집』 제71집, 근역한문학회, 2025, 211-242쪽.
- 김예리, 「김춘수의 ‘무의미시론’ 비판과 시의 타자성」, 『한국현대문학연구』 제38호, 한국현대문학학회, 2012, 105-136쪽.
- 김현, 「조선왕조실록 전산화를 위한 마크업(MarkUp) 규칙」, 『古文研究』 제9집, 한국고문연구회, 1996, 244-262쪽.
- , 「XML 기반 고문헌 데이터베이스 구축 연구」, 『한국비교문학』 제29호, 한국비교문학학회, 2003, 183-230쪽.
- , 「디지털 인문학과 고문헌 자료 연구」, 『열상고전연구』 제50집, 열상고전연구회, 2016, 13-28쪽.
- 김현·안승준·류인태, 「데이터 기반 인문학 연구 방법의 모색 : 문중 고문서 아카이브와 디지털 인문학의 만남」, 『횡단인문학』 창간호, 숙명여자대학교 인문학연구소, 2018, 17-61쪽.
- 오형엽, 「박용철 시론의 구조와 계보」, 『비평문학』 제18호, 한국비평문학학회, 2004, 7-30쪽.
- 이병주, 「1950-60년대 비평 담론의 형성 과정과 수행성 연구」, 고려대학교 박사학위논문, 2025.
- 이미순, 「김수영 시론과 ‘죽음’ — 블랑쇼의 영향을 중심으로」, 『국어국문학』 제159호, 국어국문학학회, 2011, 323-352쪽.
- 장문석·류인태, 「디지털 인문학과 한국문학 연구(1) : 작가 연구를 위한 시맨틱 데

- 이터베이스 설계, 『민족문화사연구』 제75집, 민족문화사학회, 2021, 347-426쪽.
- 정송이·김바로, 「조선총독부박물관 문서 XML 데이터 설계 및 활용 시론: XML 기반 다중 분류 체계 및 내·외부 데이터 연결을 중심으로」, 『인문과학』 제96집, 성균관대학교 인문학연구원, 2025, 181-227쪽.
- 조영복, 「김기림 시론의 기계주의적 관점과 ‘영화시’(Cinepoetry) — 페르낭 레제 및 아방가르드 예술관과 관련하여」, 『한국현대문학연구』 제26호, 한국현대문학회, 2008, 389-419쪽.
- 지혜인, 「디지털 감각·감정 분석을 통한 이상 문학의 에피파니 연구」, 『상허학보』 제72집, 상허학회, 2024, 753-827쪽.
- 하정일, 「임화의 민족문화론과 언어론」, 『한국근대문학연구』 제23호, 한국근대문학회, 2011, 7-38쪽.
- Anthropic, Claude LLM, <https://claude.ai>(2025.11.29.).
- Derrida's Margins, *Princeton University Library Project*, <https://derridas-margins.princeton.edu/library/> (2025.11.29.).
- TEI Consortium, *TEI P5: Guidelines for Electronic Text Encoding and Interchange*, <https://tei-c.org/release/doc/tei-p5-doc/en/html/index.html> (2025.11.29.).

- **주제어** : 디지털 인문학, 현대문학비평 XML, TEI, 텍스트 마이닝, 데이터 아카이브, 비평 데이터베이스

논 문 접 수 일	2025년 11월 30일
심 사 완 료 일	2025년 12월 10일
게 재 확 정 일	2025년 12월 23일

- Abstract

Constructing and Utilizing XML Data for Korean Modern Literary Criticism

: A Methodological Approach to Digital Humanities Research

Lee, Byeong-Joo
(Korea Air Force Academy)

This study aims to propose a methodology for constructing XML-based data for modern Korean literary criticism and to suggest its utilization strategies in the era of digital transformation and generative AI. While digital humanities have expanded in other fields, modern Korean literature research has been relatively passive in data construction due to copyright issues and the discipline's emphasis on subjective "interpretation." Consequently, most materials remain fragmented as unstructured data (PDF, HWP), limiting collaboration and knowledge accumulation among researchers.

To address these limitations, this study proposes a customized XML schema (XSD) based on the TEI (Text Encoding Initiative) P5 guidelines, designed to transform texts into machine-readable structured data while reflecting the unique characteristics of literary criticism. By analyzing existing meta-criticism studies, this research identified "Quote" and "Interpretation" as core markup elements, alongside standard elements such as author, work, and term.

www.kci.go.kr

Specifically, the schema is designed to structure the critic's value judgments (affirmative, neutral, critical) through the <interp> element and to classify the types and genres of citations using the <quote> element, thereby database—izing the context of critical discourse.

The construction of such XML data enables contextual exploration using XPath and XQuery beyond simple keyword searches. Furthermore, it lays the foundation for quantitative literary research, including the visualization of networks between critics and authors and the tracing of the conceptual history of critical terms. By converting the individual researcher's interpretation process into a shareable data structure, this study contributes to overcoming the insularity of modern literature research and establishing a sustainable digital knowledge ecosystem.

- **Key words** : Digital Humanities, Modern Literary Criticism XML, TEI, Text Mining, Data Archive, Criticism Database