# 3. 『프랑켄슈타인』 디지털 에디션 변환 과정과 인문 데이터 포맷

## 3.1. 서론

메리 셸리의 *Frankenstein; or, the Modern Prometheus*(『프랑켄슈타인』)은 현대 공상과학소설(SF)과 고딕소설의 시작점으로 여겨지는 중요한 작품이다. 이 소설은 시대와 판본에 따라 내용이 여러 차례 수정되어 왔다. 본 문서에서는 이 작품이 종이책이나 필사본과 같은 물리적 자료에서 **디지털 판본**으로 변환되는 과정을 설명한다. 또한 **디지털 인문학**(Digital Humanities) 연구에서 자주 활용되는, 컴퓨터가 읽고 처리할 수 있는 **데이터 포맷**(CSV/TSV, JSON, XML/TEI)의 특징과 활용 방안을 소개한다. 나아가 여러 판본을 비교할 수 있는 디지털 바리오럼(variorum) 에디션을 만드는 과정과, 이런 **텍스트 구조화** 작업이 전통적인 문헌 연구에 어떤 도움을 주는지 살펴본다.

『프랑켄슈타인』을 디지털 판본으로 만들면 큰 장점이 있다. 1818년 초판과 1831년 개정판을 비롯한 여러 판본을 한 곳에 모아 **동시에 비교**하고 **검색**하며 **주석 정보**를 확인할 수 있기 때문이다. 이는 기존에 종이책으로 하던 대조 작업을 훨씬 체계적이고 효율적으로 만들어준다. 따라서 고전 문학을 디지털 판본으로 만드는 작업은 학술적·역사적·교육적으로 모두 중요한 의미를 지니며, *프랑켄슈타인*은 이러한 변환 작업의 대표적인 사례로 꼽힌다.

## 3.2. 종이에서 디지털로: 『프랑켄슈타인』의 디지털 에디션 변환 과정

이 절에서는 전통적인 판본이 디지털 판본으로 변환되는 과정을 **단계별**로 자세히 살펴본다.

#### (1) OCR 및 텍스트 추출

첫 번째 단계는 종이 자료나 스캔 이미지에서 텍스트를 추출하는 것이다. **광학 문자 인식**(OCR)이라는 기술을 사용하는데, 이는 책이나 문서를 스캔한 이미지에서 글자를 자동으로 인식하는 기술이다. 예를 들어 1823년의 드문 중간판본은 이 기술로 디지털 텍스트로 변환했다.

구글 북스(Google Books)에 스캔된 인쇄본 이미지를 OCR로 읽고, 오탈자를 수정하는 과정을 거쳐 1818년과 1831년 판본 모두 양식이 통일된 텍스트 파일을 얻을 수 있다. 이 두 판본은 1990년대에 이미 전자 텍스트(펜실베니아 대학의 HTML본)로 만들어진 것이 있어서, 이를 재활용했다. **텍스트 패턴을 찾고 처리하는 방법**인 정규표현식(Regular Expression)으로 불필요한 태그를 정리하고, 원본 인쇄본이나 원본과 동일하게 복제된 사본(파싱밀리)과 대조하여 오류를 바로잡았다.

#### (2) TEI/XML 인코딩과 텍스트 정교화

OCR로 얻은 텍스트는 **학술적 활용**에 알맞게 정리하는 과정이 필요하다. *프랑켄슈타인* 프로젝트 팀은 모든 판본을 **TEI P5**라는 표준에 따른 **XML 형식**으로 변환했다. TEI는 문단(<p>), 장(<div>), 등장인물(<person>), 삭제(<del>), 추가(<add>) 등 **인문학 연구에 필요한 세부 정보를 태그**로 표시할 수 있는 국제 표준이다.

1818년판(일명 *Thomas 사본*), 1823년 판본, 1831년판(및 기타 주석판)을 모두 같은 XML 구조로 정리했다. 이렇게 하면 **단락**·**장**·**편지글** 등의 구조가 일관되게 유지된다. 또한 **Shelley-Godwin 아카이브**에서 제공한 필사본 TEI 파일을 합쳐서, 각 페이지 끝부분의 여백에 쓰인 글을 자연스럽게 본문에 포함시켰다.

#### (3) 판본 대조 및 변이 분석

마지막으로, 컴퓨터 프로그램(예: CollateX)을 이용해 여러 판본 사이의 **차이점**을 찾아낸다. 텍스트를 일정한 단락 단위로 나누고 시작·끝 지점을 맞춘 다음, 중요한 차이점만 남기고 사소한 차이는 무시하는 방식으로 비교한다. 이렇게 하면 한 판본에만 있거나 수정된 문장을 **,**  태그로 함께 저장할 수 있고, 결과적으로 **비교 정보가 담긴 XML 데이터**가 만들어진다. 아래는 TEI XML에서 한 구절의 두 판본이 어떻게 표시되는지 보여주는 간단한 예시다.

<app xml:id="C07a\_app23">  
 <rdg wit="f1818">  
 You may easily conceive that I was much gratified by the offered communication;   
 yet I could not endure that he should renew his grief by a recital of his misfortunes.  
 </rdg>  
 <rdg wit="f1831">  
 You may easily imagine that I was much gratified by the offered communication;   
 yet I could not endure that he should renew his grief by a recital of his misfortunes.  
 </rdg>  
</app>

이처럼 판본 간 **문장 단위 차이**를 체계적으로 표시하면, 연구자들은 어느 판본에서 어떤 내용이 달라졌는지 쉽게 파악할 수 있다.

## 3.3. 인문 데이터 포맷의 특징과 활용

*프랑켄슈타인* 디지털 판본 프로젝트에서 만들어진 텍스트와 주석 데이터는 **여러 가지 컴퓨터 처리용 형식**(CSV/TSV, JSON, XML/TEI)으로 저장하고 활용할 수 있다. 원본 자료를 구조화하면 **검색**, **통계**, **시각화**, **관계망 분석** 등 다양한 분석이 가능해지며, 이는 **디지털 인문학**에서 새로운 발견을 하는 핵심 단계가 된다.

아래 표는 세 가지 주요 형식의 **구조**와 **장단점**을 쉽게 비교한 것이다. 각 형식의 구체적인 예시는 표 다음에 자세히 설명한다.

| 데이터 형식 | 구조 | 장점 | 단점 |
| --- | --- | --- | --- |
| **CSV** | 행과 열로 이루어진 **단순한 표** 구조. 쉼표(,)나 탭(으로 각 항목을 구분 | 구조가 단순하고 엑셀·Pandas 등에서 바로 열 수 있음 작은 크기로 관리·공유가 쉬움 대부분의 프로그램이 지원함 | 계층 구조나 자세한 정보 추가가 어려움 본문에 쉼표나 줄바꿈이 있으면 처리가 까다로움 |
| **JSON** | 이름과 값의 쌍, 그리고 목록([])을 포함하는 **계층적** 구조 웹·앱 개발에서 널리 사용됨 | 계층적인 데이터 표현이 가능 웹·앱에서 활용도가 높음 비교적 읽기 쉬운 문법 | 문헌학 표준이 아니어서 장·절·문단 같은 복잡한 구조는 따로 설계해야 함 구조가 복잡해질 수 있음 |
| **XML/TEI** | XML 태그로 문헌 구조(장, 절, 문단), 인물, 주석 등을 자세히 기록 TEI는 인문학 연구를 위한 국제 표준 | 문헌의 구조와 의미(등장인물, 변화 정보 등)를 자세히 태그로 표시 판본 비교, 주석, 부가 정보를 모두 담을 수 있음 학술 연구에 가장 적합 | 태그·규칙이 많아서 배우기 어려움 파일이 커질 수 있고 특별한 편집 도구가 필요 처음 접하기 어려울 수 있음 |

#### (1) CSV 예시

아래 예시는 *프랑켄슈타인*의 1818년판과 1831년판에서 달라진 문장과 변화 유형을 **CSV** 형태로 기록한 것이다.

문장ID,1818년판 문장,1831년판 문장,변화 유형  
S1,"You may easily conceive that I was much gratified by the offered communication; yet I could not endure that he should renew his grief by a recital of his misfortunes.","You may easily imagine that I was much gratified by the offered communication; yet I could not endure that he should renew his grief by a recital of his misfortunes.","단어 변경 (conceive→imagine)"  
S2,"The infant Elizabeth, the only child of his deceased sister, was cared for in his home.","They were fond of the sweet orphan.","인물 설정 변경 (사촌→고아)"

* **활용 방법**: 엑셀(Excel)로 열어서 빠르게 찾아보거나, 파이썬 Pandas로 통계를 내고 그래프로 그릴 수 있다.
* **장점**: 데이터 구조가 단순해서 **다루기 쉽고 공유하기 편하다**.
* **단점**: 문단이나 주석처럼 복잡한 구조나 자세한 추가 정보를 담기는 어렵다.

#### (2) JSON 예시

아래 예시는 판본별 문장 변화와 함께 등장인물 정보, 주석을 **JSON** 구조로 표현한 것이다.

{  
 "문장ID": "S2",  
 "텍스트": {  
 "1818년판": "The infant Elizabeth, the only child of his deceased sister, was cared for in his home.",  
 "1831년판": "They were fond of the sweet orphan."  
 },  
 "변화유형": "인물 설정 변경 (사촌→고아)",  
 "주석": "1818년판은 가족 관계를 자세히 설명하지만, 1831년판에서는 'orphan'이라는 표현으로 인물 설정이 바뀐다.",  
 "인물정보": {  
 "인물명": "Elizabeth",  
 "역할": "주요 등장인물"  
 }  
}

* **활용 방법**: 웹 사이트에서 AJAX로 불러와 화면에 보여주거나, NoSQL 데이터베이스에 저장해 검색할 때 사용한다.
* **장점**: 여러 층위의 정보를 체계적으로 담을 수 있고, 자바스크립트 등으로 처리하기 쉽다.
* **단점**: **문헌학 전용 표준**이 아니어서, 장·절·주석 등 복잡한 문헌 구조는 연구자가 직접 설계해야 한다.

#### (3) XML/TEI 예시

아래 예시는 *프랑켄슈타인*에서 빅터 프랑켄슈타인의 대사와 판본 차이를 **TEI**로 표시한 간단한 XML 예시다.

<TEI xmlns="http://www.tei-c.org/ns/1.0">  
 <teiHeader>  
 <fileDesc>  
 <titleStmt>  
 <title>프랑켄슈타인 (판본 비교 및 인물 정보)</title>  
 <author>Mary Shelley</author>  
 </titleStmt>  
 <publicationStmt>  
 <publisher>Digital Edition Project</publisher>  
 <date>2024</date>  
 </publicationStmt>  
 </fileDesc>  
 </teiHeader>  
 <text>  
 <body>  
 <div type="scene" n="1">  
 <u who="#Victor">I must create life!</u>  
 <person xml:id="Victor">  
 <persName>Victor Frankenstein</persName>  
 <note type="variant">  
 1818년판: "I must create life!", 1831년판: "I must endeavor to create life."  
 </note>  
 </person>  
 </div>  
 </body>  
 </text>  
</TEI>

* **활용 방법**: 문헌의 구조(장, 단락, 인물), 주석, 판본 비교 정보를 자세한 태그로 표시한다. XSLT라는 도구로 변환해서 웹페이지나 PDF 등 다른 형식으로 만들 수도 있다.
* **장점**: 원문의 **구조**와 **의미**(부가 정보, 변화, 등장인물 등)를 상세히 기록할 수 있어 **학술 연구용 판본**을 만들기에 가장 적합하다.
* **단점**: XML과 TEI 규칙을 배우는 데 시간이 많이 들고, 파일이 길어져서 처음 접하는 사람은 이해하기 어려울 수 있다.

## 3.4. 결론

이 장에서는 *프랑켄슈타인*이 디지털 판본으로 변환되는 과정을 살펴보았다. **OCR**로 텍스트를 뽑아내고, **TEI/XML 인코딩**으로 정리하며, 여러 **판본을 비교 분석**하는 세 단계로 나누어 자세히 설명했다. 또한 CSV/TSV, JSON, XML/TEI 같은 데이터 형식의 특징과 디지털 문헌 연구에서의 활용 방법을 예시와 함께 알아보았다. 이런 **데이터 변환과 구조화** 작업은 방대한 문헌을 컴퓨터로 효율적으로 비교·분석할 수 있게 해주며, 수치 기반 분석과 시각화 도구와 함께 사용하면 **인문학 연구의 새로운 가능성**을 열어준다.

디지털 판본의 가장 큰 장점은 여러 판본의 차이를 한눈에 비교할 수 있다는 점이다. 이를 통해 작가가 텍스트를 수정한 의도와 시대에 따른 변화를 쉽게 파악할 수 있다. 예를 들어 *프랑켄슈타인*의 초판과 개정판은 등장인물 설정, 운명론적 분위기 등 많은 부분이 달라졌는데, 디지털 판본에서는 이런 차이를 **,**  태그로 체계적으로 정리해 제공하므로 독자가 쉽게 원문을 비교할 수 있다. 나아가 웹 프로그램이나 시각화 도구와 연결해서 특정 단어가 얼마나 자주 나오는지, 어떤 식으로 수정되었는지를 그래프나 차트로 확인하는 등 **새로운 방식의 문헌 분석**도 가능해진다.

## 3.5. 연습 문제

1. *프랑켄슈타인*의 OCR 결과와 TEI/XML 파일을 비교해보고, 특정 구절이 어떻게 다르게 표시되었는지 분석해보자.
2. CSV/TSV 파일로 저장된 판본 비교 데이터를 활용해서, 1818년판과 1831년판에서 특정 단어가 어떻게 바뀌었는지 횟수를 세어보고 그래프로 그려보자.
3. JSON 데이터에 있는 주석 정보를 바탕으로, 등장인물(예: 엘리자베스)의 설정이 각 판본에서 어떻게 달라지는지 정리해보자.

## 참고 문헌

* [Balisage: Rebuilding a Digital Frankenstein by 2018](https://www.balisage.net/Proceedings/vol20/html/Beshero-Bondar01/BalisageVol20-Beshero-Bondar01.html#:~:text=In%20the%20fall%20of%202016%2C,24%20the%20Frankenstein%20edition%20on)
* [Hierarchies Made to Be Broken: The Case of the Frankenstein Bicentennial Variorum Edition – DH2018](https://dh2018.adho.org/en/hierarchies-made-to-be-broken-the-case-of-the-frankenstein-bicentennial-variorum-edition/#:~:text=6%20%20We%20are%20also,in%20composing%2C%20amending%2C%20and%20substantially)
* [Method](https://frankensteinvariorum.org/method#:~:text=common%20language,the%201818%20and%201831%20editions)
* [Assembling a Monster: The Frankenstein Variorum – Matthew Lincoln, PhD](https://matthewlincoln.net/2020/01/12/frankenstein-variorum.html#:~:text=%3CrdgGrp%20xml%3Aid%3D,app)
* [DHQ: Digital Humanities Quarterly: Reassessing the locus of normalization in machine-assisted collation](https://www.digitalhumanities.org/dhq/vol/14/3/000489/000489.html#:~:text=CollateX%20supports%20several%20output%20views,with%20just%20the%20n%20property)
* [DHQ: Digital Humanities Quarterly: From Archive to Database: Using Crowdsourcing, TEI, and Collaborative Labor to Construct the Maria Edgeworth Letters Project](https://digitalhumanities.org/dhq/vol/18/2/000424/000424.html#:~:text=from%20different%20sources,to%20meet%20new%20use%20cases)
* [Converting Books to JSON: A Digital Humanities Project – Digital Born](https://digitalborn.org/converting-books-to-json-a-digital-humanities-project/#:~:text=to%20perform%20document%20analysis%20across,This)