Asia-pacific Journal of Convergent Research Interchange
Vol.9, No.5, May 31 (2023), pp.117-128
http://dx.doi.org/10.47116/apjcri.2023.05.10

# Cases of Public Data Purification Needed to Build Learning Data

# 학습 데이터 구축에 필요한 공공 데이터 정제 사례

So Young Park[1], Dong Yub Lee[2], Jeong ah Kim[3]

박소영[1], 이동엽[2], 김정아[3]

*[1] Student, Computer Education, Catholic Kwandong University, South Korea, so_1022@naver.com*
*[2] Student, Computer Education, Catholic Kwandong University, South Korea, 9754745@naver.com*
*[3] Professor, Computer Education, Catholic Kwandong University, South Korea, clara@cku.ac.kr*

*Corresponding author: Jeong ah Kim*

**Abstract**: Although the quality of AI learning data portal data portal is important, in the case of public data, it cannot guarantee the quality of data portal since the purpose of the public data provide portal was not planned for AI learning .There were many fields that do not exist, and there were many cases that the scope of values allowed to data items that allow data items that allow data items are not clear.Also, there was cases where logical error exists in storage value.In the case of public data portal, it is conducted in a way to register the same institution, rather than upload it.Therefore, there is no unifiedity of data.In this paper, consistency properties of public data was defined as accuracy, consistency, general, publicability, publicability, publicability, publicity, and reliability. Generic functions that can simply exclude missing value errors or solve the case of general-purpose data have been developed, and functions suitable for the domain-specific can be defined and called. When different values are stored, a standard glossary is defined and an algorithm that can be changed to a representative terminology is provided. In this paper, it proposed several general processing algorithms to solve the errors that shows frequently appear in public data. If these algorithms are applied, it is expected to efficiently performing processing processing processing process of public data.

**Keywords**: Public Data, Data Quality, Data Preprocessing, Learning Data

**요약**: 인공지능 학습용 데이터의 품질이 중요하다는 인식을 갖고 있지만, 공공데이터 포털에서 제공하는 데이타의 경우는 인공지능 학습용 데이터 활용을 목적으로 만들어지지 않았기 때문에 데이터의 품질을 보장할 수 없다. 데이터가 존재하지 않는 필드들이 다수 존재함을 발견할 수 있었으며, 데이터 항목에 허용하는 값의 범위가 명확하지 않아서 동일한 값을 서로 다르게 표현하는 경우가 많았다. 또한, 저장 값이 논리적 오류를 갖는 경우도 존재한다. 공공데이터 포털의 경우 한 기관에서 취합하여 업로드하는 것이 아닌 여러 기관에서 보낸 것을 그대로 등록하는 방식으로 진행된다. 따라서 데이터의 통일성이 존재하지 않았다. 본 논문에서는 공공데이터의 품질 속성을 정확성, 완전성, 일관성, 공공성, 활용성, 신뢰성 6가지로 정의하였다. 또한, 본 논문에서는 공공데이터에서 빈번하게 나타나는 오류를 분석하여 이를 처리하는 전처리 알고리즘을 제안하였다. 결측치 오류를 단순하게 배제하거나, 범용적 데이터의 경우를 해결할 수 있는 기본 함수를 개발하였고, 분석 영역에 적합한 함수를

Received: January 06, 2023; 1st Review Result: February 20, 2023; 2nd Review Result: March 17, 2023
Accepted: April 30, 2023

정의하고 호출할 수 있게 하였다. 서로 다른 값이 저장된 경우는 표준 용어집을 정의하여 대표어로 변경할 수 있는 알고리즘을 제공하였다. 본 논문 제공한 일반적 처리 알고리즘을 적용한다면 공공데이터 오류수정을 위한 전 처리 과정을 효율적으로 수행할 수 있을 것으로 기대한다.

**핵심어**: 공공데이터, 데이터품질, 데이터전처리, 학습데이터

## 1. 서론

우리는 정보의 홍수 속에 살고 있다. 인공지능은 예술의 영역에 들어올 수 없다는 과거의 생각과는 다르게, 현재 인공지능은 작곡을 하며, 그림을 그려 많은 논란이 되고 있다. 여기서 우리는 인공지능과 데이터의 관계를 알 필요가 있다.

인공지능이 어떻게 만들어지는지는 누구나 알고 있다. 인공지능은 머신 러닝을 통해 학습하고 그 결과를 바탕으로 의사결정한다. 학습을 위해 필수로 필요한 것이 데이터이며, 인공지능의 의사결정에 필요한 요소도 데이터라 볼 수 있다. 여기서 데이터란 이론을 세우는데 기초가 되는 사실, 또는 바탕이 되는 자료를 말한다.

인공지능 학습용 데이터 구축 과정은 목적 정의, 데이터 확보, 데이터 정제, 데이터 라벨링, 데이터 학습의 순서로 이루어진다. '데이터 확보' 단계에서는 인공지능의 기계학습에 필요한 데이터를 현실 세계에서 직접 생성하거나, 이미 보유하고 있는 조직이나 시스템 등으로부터 인공지능 학습에 필요한 데이터를 수집한다. 이 확보 활동 중에 가장 많이 사용할 수 있는 사이트가 공공데이터 포털[1] 및 AI-Hub 포털[2]이다.

인공지능 학습용 데이터의 품질이 중요하다는 인식을 갖고 있지만, 공공데이터 포털의 경우는 인공지능 학습용 데이터로 만들어지지 않았기 때문에 데이터 품질을 보장할 수 없다. AI-Hub 데이터 포털이 제공하는 데이터의 경우, 데이터 구축 사업에 대한 이해가 부족하거나 품질 관리 역량이 낮은 기관이 생성한 데이터의 경우는 품질을 보장할 수 없다.

'공공데이터포털'은 각 지자체, 정부기관에서 다양한 정보들을 정리하여 제공하는 사이트이다. 다양한 데이터를 오픈소스로 제공하여, 많은 연구와 활용이 이루어지고 있지만, 데이터 관리기관이나 메타데이터가 통일되어 있지 않아 데이터 전처리 과정을 필수로 거쳐야 한다.

본 논문에서는 위 '공공데이터포털'이 제공하는 데이터의 문제점을 파악하고 이를 해결할 수 있는 전처리 방식을 구현하여 누구나 활용 가능한 범용 전처리기를 제안하고자 한다.

## 2. 연구배경

### 2.1 데이터 품질 지표

데이터 품질에 대한 정의, 평가 지표와 관해 여러 연구가 이루어졌다. 공공데이터 품질 가이드라인[3]에 의하면 데이터의 최신성, 정확성, 상호 연계성 등을 확보하여 이를 사용자에게 유용한 가치를 줄 수 있는 수준으로 정의하고 있다.

인공지능 학습용 데이터 품질 관리 가이드라인에서는 인공지능 학습용 데이터 품질을

인공지능 학습에 필요한 데이터의 적합성, 정확성, 유효성 등을 확보하여 사용자에게 유용한 가치를 줄 수 있는 수준이라고 정의하였다.

김학래[4]에 따르면 일반적으로, 데이터 품질은 정확성(Accuracy), 완전성(Completeness), 일관성(Consistency)과 같은 다양한 차원으로 평가되며, 공공 개방 데이터가 갖추어야 할 품질로는 공공성 (Publicity), 활용성(Usability), 신뢰성(Reliability), 완전성(Completeness)의 4가지 관점이 제시되었다[5]. 본 논문에서는 다양한 데이터 품질에 대한 연구 결과를 바탕으로 공공데이터의 품질 속성을 정확성, 완전성, 일관성, 공공성, 활용성, 신뢰성 6가지로 정의하였다.

[표 1] 데이터 품질 측정 방법

[Table 1] Data Quality Measurement Method

| 품질 속성 | 설명 |
|---|---|
| 정확성 | 데이터의 바르고 확실한 성질 |
| 완전성 | 필요한 요소를 모두 갖추어 부족함이나 결함이 없는 상태 |
| 일관성 | 방법이나 태도 따위가 한결같은 성질 |
| 공공성 | 한 개인이나 단체가 아닌 일반 사회 구성원 전체에 두루 관련되는 성질 |
| 활용성 | 잘 이용할 수 있는 특성 |
| 신뢰성 | 굳게 믿고 의지할 수 있는 성질 |

정확성이란 바르고 확실한 성질을 말한다[6]. 완전성은 필요한 요소를 모두 갖추어 부족함이나 결함이 없는 상태를 말한다[7]. 일관성이란 방법이나 태도 따위가 한결같은 성질을 말한다[8]. 공공성이란 한 개인이나 단체가 아닌 일반 사회 구성원 전체에 두루 관련되는 성질을 말한다[9]. 활용성 이란 잘 이용할 수 있는 특성을 말한다[10]. 신뢰성이란 굳게 믿고 의지할 수 있는 성질을 말한다[11]. 이와 같이 데이터 품질 속성들을 정의하여 데이터 오류를 탐지하였고 품질 속성을 기준 삼아 데이터를 개선할 수 있었다.

## 2.2 데이터 품질 관리 수준

공공데이터법 제22조(공공데이터의 품질관리)는 공공데이터의 안정적 품질관리 및 적정한 품질수준의 확보를 위하여 품질 진단·평가, 개선 지원 등 필요한 시책을 수립하고, 사회적·경제적 파급효과가 큰 공공데이터에 대한 품질 평가를 실시하고 결과를 공표할 수 있다 고 규정하고 있다[6]. 데이터 품질에 대한 인식을 기반으로 행정안전부에서는 공공기관의 데이터 표준화 지침[12]을 운영하고 있다. 지침에는 공공기관이 생성 또는 관리하는 데이터베이스의 표준화 원칙, 방법, 조직, 지표 등을 정의하고 있다. 공공기관들의 업무 및 데이터베이스에서 공동 활용해야 하는 표준 용어와 행정 코드를 정의하고 있다. 공공데이터를 개방할 때 적용해야 하는 메타데이터, 제공 방식, 공통항목들도 표준으로 정의하고 있다. 또한 데이터를 수집, 개방하는 기관의 데이터 품질관리 조직과 역할, 데이터 품질을 검증하기 위한 주기와 방법도 정의하고 있다.

공공데이터 품질관리 내실화를 위하여 공공데이터 품질심사원을 양성해오고 있으며, 인증을 시행하고 있지만, 여전히 공공데이터 품질은 좀 더 개선이 필요하다고 지적 있다. 우리나라 공공데이터의 경우 매우 편향된 평가를 받고 있는데 이는 실무적 차원에서 공공데이터 개념 정립이 미흡하고 수요자 관점에서 실효성 있는 데이터의 공개가

부족하기 때문이다[13].

공공데이터 품질평가를 위해 개방 표준 데이터를 수집하고, 완전성과 정확성에 대한 평가를 수행한 연구 결과에 따르면 완전성 지표는 0.88(1점 기준), 정확성 지표는 0.78(1점 기준)을 보였다. 공공데이터에서 보이는 주요 오류로는 공백 오류, 특수문자 오류, 값 오류였다. 공공 데이터의 완전성과 정확성 지수가 낮은 수준은 아니지만, 국가가 인증한 표준 데이터라면 더 높은 수준의 품질 제공이 필요하다.

이러한 오류들 중에는 빈번하게 나타나면서 오류에 패턴이 존재하기도 하는데, 이를 처리하는 전처리기 알고리즘을 개발한다면 공공 데이터 전체에 적용하여 품질을 개선할 수 있을 것이다.

## 3. 데이터 오류 유형

### 3.1 공공데이터 품질의 문제점

공공데이터를 수집할 수 있도록 제공하는 사이트는 많이 존재한다. 국내사이트의 경우 공공데이터 포털, 서울열린데이터광장, e-나라지표, 국가통계포털, 서울특별시 빅데이터 캠퍼스, 통계청 등이 있다. 해외사이트의 경우 구글 데이터 검색, 캐글, Awesome Public Datasets Github, Data and Story Library 등이 있다. 공공데이터의 형태는 txt, csv, excel, API호출, hwp, pdf 등이 있다.

공공데이터의 일반적인 품질 문제는 다양하게 존재한다. 그 사례로 공중화장실, 전기차 충전정보 데이터에서 결측 데이터, 데이터 불일치, 데이터 값 오류 총 3가지 오류가 나타났다. 공공데이터는 행정안전부에서 제공·운영하는 지방행정 인허가 데이터 개방: LOCALDATA의 데이터와 공공데이터 포털에서 데이터를 수집할 수 있다. 데이터의 수집은 공공데이터 포털에서 제공하는 API를 활용할 수 있지만, 이렇게 받은 데이터에는 결측치가 너무 많았다. 동일한 공공데이터 포털에서 엑셀 파일로 다운 받은 결과, API를 통해 받은 경우보다 결측치가 적었기에 지방행정 인허가 데이터 개방 사이트와 공공데이터 포털에서 엑셀로 다운 받는 방식으로 데이터를 수집하였다.

다른 공공데이터는 공공데이터포털에서 제공하는 데이터 중 전국 데이터를 포함하고 있는 데이터를 선정하였고, 필요한 컬럼을 모두 포함하고 있는 한국환경공단에서 제공하는 API를 사용하기로 결정하였다.

공공데이터포털에서 현재까지 업데이트 되고 있는 API중에서는 한국전력공사에서 제공하는 API 또한 있었으나, 보유 데이터의 수가 한국환경공단 API에 비해 월등히 적어 제외하였다.

### 3.2 오류유형 1: 결측 데이터 (Missing data)

데이터가 존재하지 않는 필드들이 다수 존재함을 발견할 수 있다. 예를 들어서 주소로 소재지도로명주소 또는 소재지지번주소를 저장하도록 정의하고 있지만, 소재지 도로명만 추가 되어있거나 소재지번주소만 추가되어 있는 경우가 존재하였다. 또한 위도/경도 값이 존재하지 않아 지도에 표시할 수 없는 경우들도 다수 존재한다. 설치연월과 시설 담당자 연락처 등은 공백으로 자료가 없는 경우들이 대부분으로 고장 신고에 필요한 서비스 앱 개발에 필요한 정보가 없는 경우에 해당한다.

| 전화번호 | 개방시간 | 설치연월 | WGS84위도 | WGS84경도 |
|---|---|---|---|---|
| | 0 | | | |
| | 0 | | | |
| | 0 | | | |
| | 0 | | | |
| | 0 | | | |
| | 0 | | | |

[그림 1] 데이터 결측 사례 1

[Fig. 1] Cases of Missing Data 1

| 소재지도로명주소 | 소재지지번주소 |
|---|---|
| | 강원도 화천군 화천읍 |
| | 강원도 화천군 화천읍 |
| | 강원도 화천군 화천읍 |
| | 강원도 화천군 화천읍 |
| | 강원도 화천군 화천읍 |
| | 강원도 화천군 화천읍 |
| | 강원도 화천군 화천읍 |
| | 강원도 화천군 화천읍 |
| | 강원도 화천군 화천읍 |
| | 강원도 화천군 화천읍 |
| | 강원도 화천군 화천읍 |
| | 강원도 화천군 화천읍 |

[그림 2] 데이터 결측 사례 2

[Fig. 2] Cases of Missing Data 2

### 3.3 오류유형 2: 데이터 불일치 (Data Inconsistency)

데이터 항목에 허용하는 값의 범위가 명확하지 않아서 동일한 값을 서로 다르게 표현하는 경우가 많았다. 예를 들어서 개방시간의 경우 연중무휴를 표현한 다양한 사례를 살펴보면, 24시간, 24, 00:00~24:00, 24시간(7~8월만개방), 연중개방, 연중개방(동절기 폐쇄) 등 다양함을 발견할 수 있었다. 이는 동일한 의미이지만, 입력기관별로 입력하는 데이터 값이 서로 다르기 때문으로 분석할 수 있었다. 그림 3은 개방시간 속에 저장된 다양한 값의 형태를 보여주는 예이다.

| 개방시간 | 개방시간 |
|---|---|
| 8 | 연중개방 |
| 12 | 영업시간 |
| 24시간 | (평일)09:00~18:00 |
| 9시간 | 매주 월요일 휴무+국경일 휴무 |
| 09:00 ~ 18:00 | 행사시 개방 |
| 24시간(7~8월만개방) | 연중무휴 |
| 상시 | (여름7~8월)00:00~24:00 |
| 24시간 (동절기 폐쇄) | (연중무휴)09:00~19:00 |
| 연중개방(동절기 폐쇄) | 연중무휴 |

[그림 3] 데이터 불일치 1

[Fig. 3] Data Inconsistency 1

외래어 표기법이 기관별로 다르거나, 주식회사 표기법에 입력 담당자 별로 다르게 표시하는 경우도 데이터 유형 불일치의 사례이다.

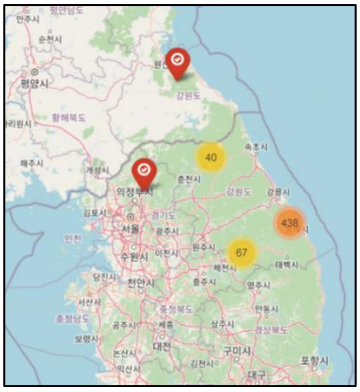
| 운영기관 명 |
| --- |
| (주)차지인 |
| 광주시 |
| 대영채비㈜ |
| 제주전기자동차서비스 |
| 포스코ICT |
| 삼성이브이씨 |
| (주)클린일렉스 |
| 퍼워큐브 |
| 파워큐브 |
| 한국전기차충전서비스 |

[그림 4] 데이터 불일치 2

[Fig. 4] Data Inconsistency 2

### 3.4 오류유형 3: 데이터 값 오류 (Data value error)

저장 값에 논리적 오류가 존재하는 유형으로 대표적인 사례가 위·경도 데이터에서 발견되었다. 위·경도를 실제 지도에 찍어 보았을 때의 그림 자료이다. [그림1]의 경우 강원도 데이터를 기준으로 지도에 화장실의 위치를 표현한 결과인데 좌표가 찍힌 위치를 보면 한 곳은 의정부시로 되어있으며, 다른 곳은 북한으로 좌표가 찍혀있다.




[그림 5] 데이터 값 오류 1

[Fig. 5] Data Value Error 1


[그림 6] 데이터 값 오류 2

[Fig. 6] Data Value Error 2

# 4. 데이터 품질 개선

## 4.1 데이터 개선

본 논문에서는 데이터 공공 데이터 오류유형별로 개선에 필요한 기법을 [표 2]와 같이 정의하였다. 결측데이타와 데이터 값 오류 처리를 위해서는 도메인 별로 특정 함수의 추가 개발이 필요할 수도 있지만, 대부분의 경우는 표준 알고리즘 또는 컴포넌트를 통해 범용적으로 적용할 수 있다.

[표 2] 데이터 품질 개선 방안

[Table 2] Data Quality Improvement Method

| 오류유형 | 개선 방법 (기술, 전략) |
|---|---|
| 결측 데이터 | 해당 데이터 행을 삭제<br>특정 속성별 결측치를 처리하기 위한 별도의 함수호출<br>평균값으로 대체 |
| 데이터 불일치 | 용어 사전을 구축하고 대표어로 통일<br>표준 용어를 관리하는 컴포넌트를 구현 및 활용 |
| 데이터 값 오류 | 시제와 관련된 경우 시제논리 처리 함수를 개발 및 적용<br>도메인에 적합한 처리 함수 개발 및 적용 |

### 4.1.1 결측 데이터 (Missing Data) 개선 방안

[그림 1]과 [그림 2]를 보면 '소재지도로명주소'와 '소재지지번주소' 중 하나만 존재하는 경우가 있다. 이럴 경우 주소변환 API를 활용하여 'Input: 도로명주소'일 경우 'Output: 지번주소'가 나오도록 하며, 'Input: 지번주소'일 경우 'Output: 도로명주소'가 나오도록 해야 한다. 본 논문에서 제안하고 개발한 결측치 데이터 개선 모듈의 알고리즘은 다음 [표 3] 과 같다.

[표 3] 결측 데이터 개선 알고리즘

[Table 3] Algorithm for Missing Data Improvement

```
if 결측 데이터 유형 = 주소 then
        주소변환 API 를 호출하여 도로명주소 또는 지번 주소를 얻어 채운다
if 결측 데이터 유형 = 정량데이터 then
        해당 데이터의 평균값으로 대체한다
otherwise
                        해당하는 행을 삭제한다.
```

### 4.1.2 데이터 불일치 (Data Inconsistency) 개선 방안

[그림 4]를 볼 경우 동일한 회사명 데이터도 작성방법에 따라 다르게 구분되어 등록이 되어있는 것을 알 수 있다. 이 경우 데이터의 용어 표준화가 필요하다. 예를 들어 ['파워큐브', '퍼워큐브', '파워큐브코리아']가 있다고 생각해보자. 이때 우리는 위 3가지 데이터가 하나의 회사를 가르킨다는 것을 알 수 있다. 따라서 사용자에게 '대표어'와 '유사어'를 입력받아 유사어를 모두 대표어로 바꿔줄 수 있다. 위와 같은 사례에서는 대표어를 '파워큐브'로 설정하고, 유사어를 '퍼워큐브', '파워큐브코리아'로 설정할 경우 '퍼워큐브'와

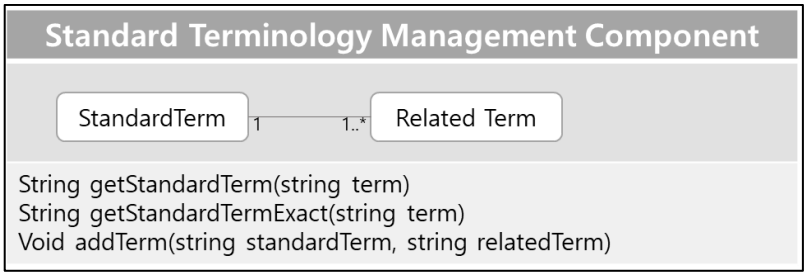
‘파워큐브코리아’는 ‘파워큐브’라고 변환되어 저장된다. 이와같은 방식을 알고리즘으로 [표 4]와 같이 표현하였다.

[표 4] 데이터 불일치 개선 알고리즘

[Table 4] Algorithm for Data Inconsistency Improvement

```
if 원천 데이터 = 대표어 then
        pass
else if 원천 데이터 = 유사어 유형 then
        미리 정의한 대표어로 교체
otherwise
                        해당하는 행을 삭제한다.
```

데이터 불일치 개선을 위한 컴포넌트 설계도는 다음 [그림 7]과 같다.




[그림 7] 표준용어관리컴포넌트 다이어그램

[Fig. 7] Diagram for Standard Terminology Management Component

#### 4.1.3 데이터 값 오류 (Data value error) 개선 방안

[그림 5] 와 [그림 6]을 보면 위 · 경도값을 실제 지도에 나타냈을 때 논리적인 오류가 있다는 것을 알 수 있었다. [그림 5]의 경우 경기도와 북한에 화장실이 위치하고 있다고 나왔으며, [그림 6]의 경우 아프리카에 화장실이 위치한다고 지도에 나타났다. 이처럼 데이터에 오류가 발생할 경우 다른 데이터로 유추가 가능한 경우 API등을 활용하여 오류 데이터를 논리적으로 맞는 데이터 값으로 수정하거나, 제거하여 이후 처리에서 문제가 없도록 해야 한다. 이와 같은 처리를 통해 고품질의 데이터를 확보할 수 있고 2차 가공을 통한 다른 특징을 추출하기에 적합하도록 개선될 수 있다.

[표 5] 데이터 값 오류 개선 알고리즘

[Table 5] Algorithm for Data Value Error Improvement

```
if 데이터 유형 = 위경도 then
        위경도기반 주소 확인 API 호출
if 데이터 유형 = 시간흐름  then
                시제논리 확인 API 호출 //  ValudateTemporalLogic(feature 1, feature 2)
```

[표 5] 에서 보는 것 같이 값의 오류를 해결하기 위해서는 데이터 활용 목적과 항목의 논리적 의미에 대한 이해가 필수적이다. 일부 데이터 항목의 경우는 관련 데이터터 항목을 기반으로 오류 검증을 자동으로 할 수 있는데, 데이터 오류 중 가장 빈번하게 발견되는 위경도 정보가 예가 될 수 있다. 지도에 표현하기 위해 필요한 위경도 정보는 주소 데이터를 동반하는 것이 일반적이다. 그러므로 주소가 속한 국가의 위경도 정보를

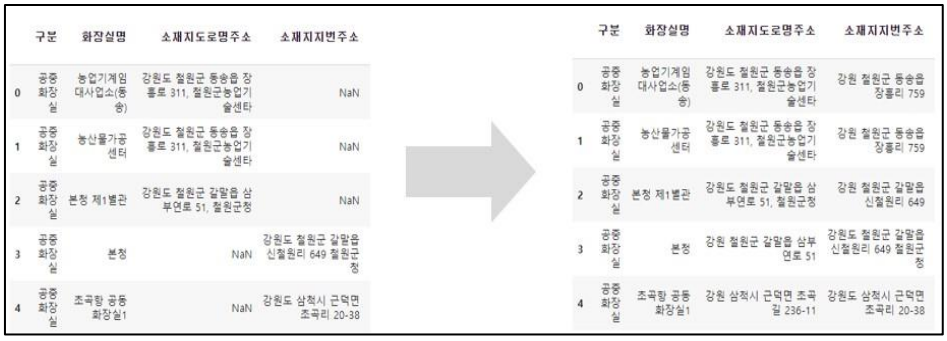
바탕으로 정의된 위경도와 비교를 통해 값의 정확성을 자동으로 검증할 수 있다. 그러나 많은 경우 데이터의 논리적 의미를 자동으로 파악하고 검증에 필요한 알고리즘을 자동으로 선정하는 것은 불가능하다. 공공데이터 중 시간과 관련한 데이터들을 흔하게 발견할 수 있다. 시작시간, 종료시간, 개방시간, 개업일, 폐업일 등이 시간 관련 데이터들의 예이다. 이런 시간 관련 데이터들은 시제논리를 기반으로 검증할 수 있다. 다만, 자동으로 어떤 시제논리를 기반으로 검증해야 하는지를 파악할 수는 없지만, 시제논리를 검증할 수 있는 알고리즘들을 미리 구축한다면 데이터 검증시 알고리즘 호출로 정제 과정을 효율화 할 수 있을 것이다. 데이터 오류의 유형을 발견하고 유형별 해결 알고리즘을 구축해 나간다면서 학습의 과정을 거친다면, 데이터 항목들의 이름들을 통해 데이터 값을 사전 검증하는 과정을 통해 오류를 개선할 수 있을 것으로 기대한다.

### 4.2 데이터 품질 개선 적용 예시

본 논문에서 정의한 데이터 품질 개선 전략의 활용 가능성을 검증하기 위하여 공중화장실 공공 데이터와 전기차 충전소 공공 데이터에 적용해보았다. 다양한 데이터 중 공중화장실을 선택한 이유는 인간이 살아가며 필수로 필요한 공간 중 하나인 화장실 데이터가 어떻게 관리되고 있는지, 자관련 법령에 맞춰 설치되고 있는지 확인하고자 선택하였다. 전기차 충전소를 선택한 이유는 국내 전기차 누적 등록 대수는 해가 지날수록 꾸준히 증가하는 추세를 보이고 있고 자연스럽게 전기차 충전소에 대한 수요 역시 증가하고 있다. 따라서 주기적으로 데이터의 값이 바뀔 수 밖에 없는 충전소 공공 데이터의 오류 유형을 확인하고 개발한 컴포넌트가 적용 가능한지를 확인하고자 선택하였다.

#### 4.2.1 결측치 개선 사례

공중 화장실과 충전소 정보의 공공 데이터 모두에서 결측치를 발견할 수 있었다. 전화번호, 설치연월, 위/경도, 도로명주소, 지번 주소 등이 주로 결측치를 가진 데이터이다. 위/경도와 도로명주소, 지번 주소의 경우 한가지의 데이터만 있으면 없는 데이터를 채울 수 있는 유형이라서, 본 논문이 개발한 주소 번환 컴포넌트를 통해 해결하였다. 그러나 전화번호나 설치연월의 경우 제공되는 데이터가 없을 경우 해결하기 어려운 데이터로 공중화장실의 경우는 주소지 관할 지자체 전화번호로 대체하게 해결하였으나, 충전소의 경우는 사설 충전소가 존재하므로 결측치로 남겨두었다.



| | 구분 | 화장실명 | 소재지도로명주소 | 소재지지번주소 |
|---|---|---|---|---|
| 0 | 공중화장실 | 농업기계임대사업소(통송) | 강원도 철원군 동송읍 장흥로 311, 철원군농업기술센타 | NaN |
| 1 | 공중화장실 | 농산물가공센터 | 강원도 철원군 동송읍 장흥로 311, 철원군농업기술센타 | NaN |
| 2 | 공중화장실 | 본청 제1별관 | 강원도 철원군 갈말읍 삼부연로 51, 철원군청 | NaN |
| 3 | 공중화장실 | 본청 | NaN | 강원도 철원군 갈말읍 신철원리 649 철원군청 |
| 4 | 공중화장실 | 조곡항 공동화장실1 | NaN | 강원도 삼척시 근덕면 조곡리 20-38 |

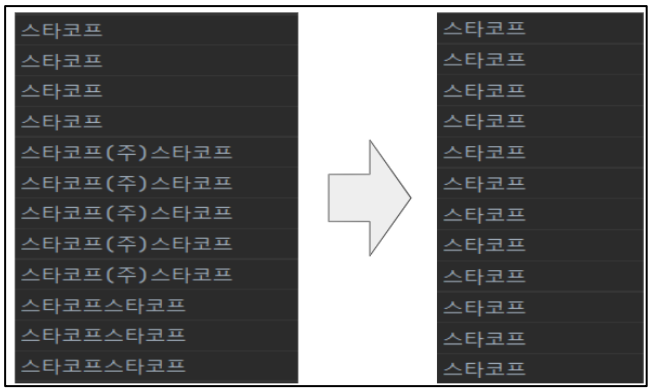
| | 구분 | 화장실명 | 소재지도로명주소 | 소재지지번주소 |
|---|---|---|---|---|
| 0 | 공중화장실 | 농업기계임대사업소(통송) | 강원도 철원군 동송읍 장흥로 311, 철원군농업기술센타 | 강원 철원군 동송읍 장흥리 759 |
| 1 | 공중화장실 | 농산물가공센터 | 강원도 철원군 동송읍 장흥로 311, 철원군농업기술센타 | 강원 철원군 동송읍 장흥리 759 |
| 2 | 공중화장실 | 본청 제1별관 | 강원도 철원군 갈말읍 삼부연로 51, 철원군청 | 강원 철원군 갈말읍 신철원리 649 |
| 3 | 공중화장실 | 본청 | 강원 철원군 갈말읍 삼부연로 51 | 강원도 철원군 갈말읍 신철원리 649 철원군청 |
| 4 | 공중화장실 | 조곡항 공동화장실1 | 강원 삼척시 근덕면 조곡길 236-11 | 강원도 삼척시 근덕면 조곡리 20-38 |

[그림 8] 결측치 개선 사례

[Fig. 8] Cases of Missing Data Improvement

### 4.2.2 데이터 유형 불일치 개선 사례

공공데이터 중 기관명, 표기법에 따른 데이터 유형의 불일치가 있는 부분이 존재하였고 위와 같은 문제의 경우 데이터의 용어 표준화로 개선하고자 하였다. [‘스타코프’, ‘스타코프(주)스타코프’, ‘스타코프스타코프’]와 같은 용어의 불일치가 보일 때, 우리는 위 3가지 데이터가 하나의 회사를 가르킨다는 것을 알 수 있었고 이에 임의로 ‘대표어’와 ‘유사어’를 설정하여 유사어를 모두 대표어로 바꿔주었다. 위와 같은 사례에서는 대표어를 ‘스타코프’로 설정하고, 유사어를 ‘스타코프(주)스타코프’, ‘스타코프스타코프’로 설정하여 ‘스타코프스타코프’와 ‘스타코프(주)스타코프’는 ‘스타코프’라고 변환되어 저장된 모습을 볼 수 있다. 이와같은 방식으로 대표어 설정을 통해 데이터를 개선하였다.

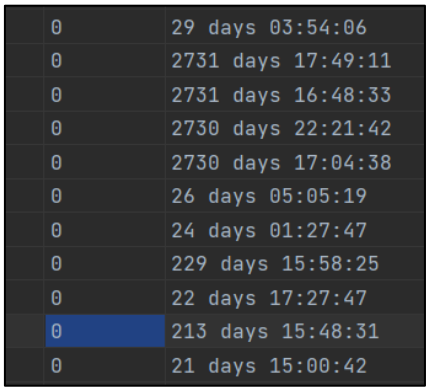



[그림 9] 데이터 유형 불일치 개선 사례

[Fig. 9] Cases of Data Inconsistency Improvement

### 4.2.3 데이터 값 오류 개선 사례

데이터 값 오류의 경우 전기차 충전소 데이터에서 다량 확인되었다. 전기차 충전소 데이터의 경우 충전시작시간과 충전종료시간 등을 제공하였는데 이와 같은 값들은 오류가 있는지 확인하기 어려운 값이었다. 이에 충전시간을 산출하여 음수값의 충전시간을 가진 경우나 비정상적으로 긴 충전시간을 가진 경우를 확인할 수 있었다.




[그림 10] 데이터 값 오류 사례

[Fig. 10] Cases of Data Value Error Improvement

이렇게 확인된 데이터 값 오류가 포함된 컬럼들을 임의로 지정한 기준에 따라 제거하거나

시제논리 함수를 통해 데이터를 개선하였다.

## 5. 결론

공공데이터포털의 경우 한 기관에서 취합하여 업로드 하는 것이 아닌 여러 기관에서 보낸 것을 그대로 등록하는 방식으로 진행된다. 따라서 데이터의 통일성이 존재하지 않았다. 또한, 공공데이터 포털에서 제공하는 데이터와 지방행정 인허가 사이트에서 제공하는 데이터 사이의 불일치도 발견할 수가 있었다.

지금까지 국내 데이터 활용에 많이 사용되는 '공공데이터포털'에서 데이터 오류의 유형을 파악하고 개선방안을 구성하였다. 이후 실제 분석에 사용하였던 공중화장실 데이터와 전기차충전소 데이터를 활용하여 데이터 품질 개선을 적용해보았고 품질속성을 만족하도록 개선됨을 보였다. 본 논문에서는 공공데이터 에서 많이 나타나는 오류 유형 3분야에 대해 해결안을 제시하였으나, 공공데이터의 품질 향상과 활용도 증대를 위해서는 전처리 과정에 활용할 수 있는 가공 전략들을 미리 제공하는 것이 바람직하다. 오류 수정과 가공에 필요한 풍부한 전략들이 라이브러리나 프레임워크 형태로 제공될 수 있다면 공공데이터의 활용 가능성이 더 높아질 것으로 기대된다.

공공 데이터를 기반으로 서비스를 개발하거나 AI 학습 데이터로 사용하기 위해서는 다음의 연구 개발이 필요하다.

첫째 공공 데이터의 데이터 오류 유형별로 정제 (Cleansing) 하는 소프트웨어 개발을 통해 서비스 개발 목적에 맞는 정확한 데이터를 활용해야 한다.

둘째, 공공 데이터 수집 시 사전 데이터 오류를 검증하는 체제를 갖추어야 한다. 즉, 공공 데이터를 수집한다는 것은 향후 활용을 목적으로 이루어져야 하는 활동이다. 그러므로 수집할 때 데이터 항목별 허용 값의 범위, 필수와 선택 여부, 값의 무결성 등을 검증하고 수집할 필요가 있다. 서비스 개발의 목적 별로 데이터 정제 프로그램을 개발하는 것 보다는 수집 시 오류를 검증하고 정확한 데이터를 수집할 수 있는 기반을 갖추는 것이 공공 데이터 관리의 목적에 부합한다고 볼 수 있을 것이다.

## 6. 감사의 글

본 과제(결과물)는 2022년도 교육부의 재원으로 한국연구재단의 지원을 받아 수행된 지자체-대학 협력기반 지역혁신 사업의 결과입니다.(2022RIS-005)

## References

[1] https://www.data.go.kr/, Mar 12 (2023)

[2] https://aihub.or.kr/, Mar 12 (2023)

[3] https://www.data.go.kr/bbs/rcr/selectRecsroom.do?pageIndex=1&originId=PDS_0000000000000516, Mar 12 (2023)

[4] Haklae Kim, Quality Evaluation of the Open Standard Data, The Journal of the Korea Contents Association, (2020), Vol.20, No.9, pp.439-447.
DOI: http://doi.org/10.5392/JKCA.2020.20.09.439

[5] Go Eun Park, Chang Jae Kim, Quality Characteristics of Public Open Data, Journal of digital convergence, (2015),

Vol.13, No.10, pp.135-146.
DOI: http://dx.doi.org/10.14400/JDC.2015.13.10.135

[6]https://stdict.korean.go.kr/search/searchResult.do?pageSize=10&searchKeyword=%EC%99%84%EC%A0%84%EC%84%B1, Jan 10 (2023)

[7] https://ko.dict.naver.com/#/entry/koko/485c5ad0c9f84b0d8dda7a0be44cba8c, Jan 10 (2023)

[8] https://ko.dict.naver.com/#/entry/koko/ce2368dff5744d3ab4d0b6aa0302edc6, Jan 10 (2023)

[9] https://ko.dict.naver.com/#/entry/koko/3139ceacd9d7482d8ed6e44da7258943, Jan 10 (2023)

[10] https://ko.dict.naver.com/#/entry/koko/f6711e9070d44638853ec9f10b5e816a, Jan 10 (2023)

[11] Cho Jane, A Study about Library-Related Open Data through Public Data Portals, Journal of the Korean BIBLIA Society for Library and Information Science, (2018), Vol.29, No.2, pp.35-56.
DOI: https://doi.org/10.14699/kbiblia.2018.29.2.035

[12]https://www.mois.go.kr/frt/bbs/type001/commonSelectBoardArticle.do?bbsId=BBSMSTR_000000000016&nttId=69428#none, Mar 12 (2023)

[13] https://ko.dict.naver.com/#/entry/koko/998d1e60df304ce3868795f869851bad, Jan 10 (2023)