문화체육관광부
국립국어원

국립국어원 개체명 분석 말뭉치 개체 연결 2021

(버전 1.2)

· **자료명:** 국립국어원 개체명 분석 말뭉치 개체 연결 2021

· **공개일**

· (버전 1.0) 2022. 3. 31.

· (버전 1.1) 2022. 9. 16.

- 2020년 구축 개체명 분석 말뭉치의 개체 연결 자료에서 부적절한 내용 포함 문서 삭제

· (버전 1.2) 2022. 12. 30.

- OGG_SPORTS 국가대표팀 관련 지침 보완 및 정밀화
- OGG_POLITICS 관련 지침 보완 및 상세화

· **자료 유형:** 텍스트

· **관련 사업:** 2021년 개체명 분석 및 개체 연결 말뭉치 연구 분석(2021)

· **자료 설명**

※ 자세한 내용은 국립국어원 누리집 > 자료 > 연구·조사 자료 > '2021년 개체명 분석 및 개체 연결 말뭉치 연구 분석' 사업 보고서 참고

· 내용

- 대상 문서 내 인식된 개체 표현(entity mention)에 대해 그 개체 표현이 문서 내에서 의미적으로 지칭하는 개체를 지식베이스에서 찾아 연결 정보를 부착함.
- 개체 유형은 '2022년 개체명 분석 말뭉치 구축 지침 ver. 2.4.'에 따라 150개 세분류 체계로 구축된 분석 말뭉치가 기준이며, PERSON(PS)과 LOCATION(LC), ORGANIZATION(OG), ARTIFACTS(AF), DATE(DT)의 세분류를 연결 대상으로 함.
- 연결할 지식베이스는 ① '한국어 위키피디아'를 기본으로 함. ①에 없는 경우, ② '영어 위키피디아'로 연결함. ②에 없는 경우, '없음'에 해당하는 정보(NA)를 표시함.
- '2021년 개체 연결 말뭉치 구축 지침 ver. 1.6.'('2021년 개체명 분석 및 개체 연결 말뭉치 연구 분석' 보고서, 2021)을 기준으로 연결하나, OGG_SPORTS와 OGG_POLITICS에 관한 연결은 세부 내용이 보완된 수정사항을 기준으로 함.

· 개체명 분석 표지: 150개(※ 연결 대상 분석 표지에 바탕색 표시)

대분류	세분류
1. PERSON(PS)	PS_NAME, PS_CHARACTER, PS_PET
2. STUDY_FIELD(FD)	FD_SCIENCE, FD_SOCIAL_SCIENCE, FD_MEDICINE,

	FD_ART, FD_HUMANITIES, FD_OTHERS
3. THEORY(TR)	TR_SCIENCE, TR_SOCIAL_SCIENCE, TR_MEDICINE, TR_ART, TR_HUMANITIES, TR_OTHERS
4. ARTIFACTS(AF)	AF_BUILDING, AF_CULTURAL_ASSET, AF_ROAD, AF_TRANSPORT, AF_MUSICAL_INSTRUMENT, AF_WEAPON, AFA_DOCUMENT, AFA_PERFORMANCE, AFA_VIDEO, AFA_ART_CRAFT, AFA_MUSIC, AFW_SERVICE_PRODUCTS, AFW_OTHER_PRODUCTS
5. ORGANIZATION(OG)	OGG_ECONOMY, OGG_EDUCATION, OGG_MILITARY, OGG_MEDIA, OGG_SPORTS, OGG_ART, OGG_MEDICINE, OGG_RELIGION, OGG_SCIENCE, OGG_LIBRARY, OGG_LAW, OGG_POLITICS, OGG_FOOD, OGG_HOTEL, OGG_OTHERS
6. LOCATION(LC)	LCP_COUNTRY, LCP_PROVINCE, LCP_COUNTY, LCP_CITY, LCP_CAPITALCITY, LCG_RIVER, LCG_OCEAN, LCG_BAY, LCG_MOUNTAIN, LCG_ISLAND, LCG_CONTINENT, LC_SPACE, LC_OTHERS
7. CIVILIZATION(CV)	CV_CULTURE, CV_TRIBE, CV_LANGUAGE, CV_POLICY, CV_LAW, CV_CURRENCY, CV_TAX, CV_FUNDS, CV_ART, CV_SPORTS, CV_SPORTS_POSITION, CV_SPORTS_INST, CV_PRIZE, CV_RELATION, CV_OCCUPATION, CV_POSITION, CV_FOOD, CV_DRINK, CV_FOOD_STYLE, CV_CLOTHING, CV_BUILDING_TYPE
8. DATE(DT)	DT_DURATION, DT_DAY, DT_WEEK, DT_MONTH, DT_YEAR, DT_SEASON, DT_GEOAGE, DT_DYNASTY, DT_OTHERS
9. TIME(TI)	TI_DURATION, TI_HOUR, TI_MINUTE, TI_SECOND, TI_OTHERS
10. QUANTITY(QT)	QT_AGE, QT_SIZE, QT_LENGTH, QT_COUNT, QT_MAN_COUNT, QT_WEIGHT, QT_PERCENTAGE, QT_SPEED, QT_TEMPERATURE, QT_VOLUME, QT_ORDER, QT_PRICE, QT_PHONE, QT_SPORTS, QT_CHANNEL, QT_ALBUM, QT_ADDRESS, QT_OTHERS
11. EVENT(EV)	EV_ACTIVITY, EV_WAR_REVOLUTION, EV_SPORTS, EV_FESTIVAL, EV_OTHERS
12. ANIMAL(AM)	AM_INSECT, AM_BIRD, AM_FISH, AM_MAMMALIA, AM_AMPHIBIA, AM_REPTILIA, AM_TYPE, AM_PART, AM_OTHERS
13. PLANT(PT)	PT_FRUIT, PT_FLOWER, PT_TREE, PT_GRASS, PT_TYPE, PT_PART, PT_OTHERS
14. MATERIAL(MT)	MT_ELEMENT, MT_METAL, MT_ROCK, MT_CHEMICAL
15. TERM(TM)	TM_COLOR, TM_DIRECTION, TM_CLIMATE, TM_SHAPE, TM_CELL_TISSUE_ORGAN, TMM_DISEASE, TMM_DRUG,

	TMI_HW, TMI_SW, TMI_SITE, TMI_EMAIL, TMI_MODEL, TMI_SERVICE, TMI_PROJECT, TMIG_GENRE, TM_SPORTS

· 분량

총 약 1,100만 어절(웹 500만, 문어 300만, 구어 300만 어절)

- 국립국어원 개체명 분석 말뭉치 2020(버전 2.1): 500만 어절(웹)
- 국립국어원 개체명 분석 말뭉치 2021(버전 1.0): 600만 어절(문어 300만, 구어 300만 어절)

· 파일 형식: JSON(UTF-8 인코딩)

· 파일 수 및 크기: 파일 323개, 총 255MB(ZIP 파일 기준)

· **인용:**

- **(국문)** 국립국어원(2022). 국립국어원 개체명 분석 말뭉치 개체 연결 2021(버전 1.2). URL: https://kli.korean.go.kr/corpus
- **(영문)** National Institute of Korean Language (2022). NIKL Named Entity Linking 2021 (v.1.2). URL: https://kli.korean.go.kr/corpus

· **파일 명명 규칙**

<table>
<tr><th>자리</th><th>1</th><th colspan="2">2</th><th>3</th><th>4</th><th>5</th><th>6</th><th>7</th><th>8</th><th>9</th><th>10</th><th>11</th><th>12</th><th>13</th><th>14</th></tr>
<tr><td>속성</td><td>유형</td><td colspan="2">구성 방법</td><td colspan="2">주석 단계</td><td colspan="2">구축 연도</td><td colspan="8">일련번호(8자리)</td></tr>
<tr><td rowspan="2">정의 값</td><td>N: 신문
E: 웹</td><td rowspan="2">X: 추출</td><td></td><td rowspan="2" colspan="2">EL: 개체 연결</td><td rowspan="2" colspan="2">21: 2021년</td><td rowspan="2" colspan="8">00000001 ~ 99999999
(여덟 자리 일련번호)</td></tr>
<tr><td>S: 구어</td><td>A: 공적 독백
B: 공적 대화</td></tr>
<tr><td colspan="15">※ 예시: NXEL2102203310.json 2021년에 구축한 신문 자료 가공 개체 연결 파일
EXEL2105170001.json 2021년에 구축한 웹 자료 가공 개체 연결 파일
SAEL2100000895.json 2021년에 구축한 구어-공적 독백 자료 가공 개체 연결 파일</td></tr>
</table>

· **예시**

1. 기본 예시

{
"id": "SARW1900000895.1.1.24",
"form": "그래서 지금 제가 동영상을 하나 갖고 왔는데 그 고대 그리스의 천동설이

```
예측하는 행성의 움직임이 어떤 것인지를 보여주는 동영상이에요.",
                    "word": [
                        {
                            "id": 1,
                            "form": "그래서",
                            "begin": 0,
                            "end": 3
                        },
                        {
                            "id": 2,
                            "form": "지금",
                            "begin": 4,
                            "end": 6
                        },
                        {
                            "id": 3,
                            "form": "제가",
                            "begin": 7,
                            "end": 9
                        },
                        {
                            "id": 4,
                            "form": "동영상을",
                            "begin": 10,
                            "end": 14
                        },
                        {
                            "id": 5,
                            "form": "하나",
                            "begin": 15,
                            "end": 17
                        },
                        {
                            "id": 6,
                            "form": "갖고",
                            "begin": 18,
                            "end": 20
                        },
                        {
                            "id": 7,
                            "form": "왔는데",
                            "begin": 21,
                            "end": 24
                        },
                        {
                            "id": 8,
                            "form": "그",
                            "begin": 25,
                            "end": 26
                        },
                        {
                            "id": 9,
                            "form": "고대",
                            "begin": 27,
                            "end": 29
                        },
```

```
                {
                    "id": 10,
                    "form": "그리스의",
                    "begin": 30,
                    "end": 34
                },
                {
                    "id": 11,
                    "form": "천동설이",
                    "begin": 35,
                    "end": 39
                },
                {
                    "id": 12,
                    "form": "예측하는",
                    "begin": 40,
                    "end": 44
                },
                {
                    "id": 13,
                    "form": "행성의",
                    "begin": 45,
                    "end": 48
                },
                {
                    "id": 14,
                    "form": "움직임이",
                    "begin": 49,
                    "end": 53
                },
                {
                    "id": 15,
                    "form": "어떤",
                    "begin": 54,
                    "end": 56
                },
                {
                    "id": 16,
                    "form": "것인지를",
                    "begin": 57,
                    "end": 61
                },
                {
                    "id": 17,
                    "form": "보여주는",
                    "begin": 62,
                    "end": 66
                },
                {
                    "id": 18,
                    "form": "동영상이에요.",
                    "begin": 67,
                    "end": 74
                }
            ],
            "NE": [
```

```
                    {
                        "id": 1,
                        "form": "하나",
                        "label": "QT_COUNT",
                        "begin": 15,
                        "end": 17,
                        "kid": "09030000013095",
                        "wikiid": "NA",
                        "URL": "NA"
                    },
                    {
                        "id": 2,
                        "form": "고대",
                        "label": "DT_DYNASTY",
                        "begin": 27,
                        "end": 29,
                        "kid": "07070000000019",
                        "wikiid": "378315",
                        "URL":
                            "https://ko.wikipedia.org/wiki/%EA%B3%A0%EC%A0%84_%EA%B3%A
                            0%EB%8C%80"
                    },
                    {
                        "id": 3,
                        "form": "그리스",
                        "label": "LCP_COUNTRY",
                        "begin": 30,
                        "end": 33,
                        "kid": "05000000000197",
                        "wikiid": "1458",
                        "URL":
                             "https://ko.wikipedia.org/wiki/%EA%B7%B8%EB%A6%AC%EC%8A%
                             A4"
                    },
                    {
                        "id": 4,
                        "form": "천동설",
                        "label": "TR_SCIENCE",
                        "begin": 35,
                        "end": 38,
                        "kid": "02000000001314",
                        "wikiid": "NA",
                        "URL": "NA"
                    }
                ]
            },
```


2. 수정 지침: OGG_SPORTS 국가대표팀 관련 지침 보완 및 정밀화 예시


```
                "id": "SBRW1900012257.1.1.145",
                "form": "동생은 덴마크의 국가 대표 선수였어요.",
                "word": [
                    {
```

```
                "id": 1,
                "form": "동생은",
                "begin": 0,
                "end": 3
            },
            {
                "id": 2,
                "form": "덴마크의",
                "begin": 4,
                "end": 8
            },
            {
                "id": 3,
                "form": "국가",
                "begin": 9,
                "end": 11
            },
            {
                "id": 4,
                "form": "대표",
                "begin": 12,
                "end": 14
            },
            {
                "id": 5,
                "form": "선수였어요.",
                "begin": 15,
                "end": 21
            }
        ],
        "NE": [
            {
                "id": 1,
                "form": "동생",
                "label": "CV_RELATION",
                "begin": 0,
                "end": 2,
                "kid": "06130000000130",
                "wikiid": "NA",
                "URL": "NA"
            },
            {
                "id": 2,
                "form": "덴마크",
```

```
                "label": "LCP_COUNTRY",
                "begin": 4,
                "end": 7,
                "kid": "05000000000271",
                "wikiid": "9167",
                "URL":    "https://ko.wikipedia.org/wiki/%EB%8D%B4%EB%A7%88%ED%81%AC"
            },
            {
                "id": 3,
                "form": "국가 대표 선수",
                "label": "CV_OCCUPATION",
                "begin": 9,
                "end": 17,
                "kid": "06140000000376",
                "wikiid": "NA",
                "URL": "NA“
```


3. 수정 지침: OGG_POLITICS 연결 정보 관련 지침 보완 및 상세화 예시


```
        "id": "NIRW2000000001.4976.3.1",
        "form": "문화재청은 서대문형무소역사관 제10·12옥사에서 1910년 경술국치부터 대한민국임시정부 환국까지 40년 동안의 역사적 상황을 재조명하는 특별전 '문화재에 깃든 100년 전 그날'을 개최한다.",
        "word": [
            {
                "id": 1,
                "form": "문화재청은",
                "begin": 0,
                "end": 5
            },
            {
                "id": 2,
                "form": "서대문형무소역사관",
                "begin": 6,
                "end": 15
            },
            {
                "id": 3,
                "form": "제10·12옥사에서",
                "begin": 16,
                "end": 26
            },
            {
```

```
            "id": 4,
            "form": "1910년",
            "begin": 27,
            "end": 32
        },
        {
            "id": 5,
            "form": "경술국치부터",
            "begin": 33,
            "end": 39
        },
        {
            "id": 6,
            "form": "대한민국임시정부",
            "begin": 40,
            "end": 48
        },
        {
            "id": 7,
            "form": "환국까지",
            "begin": 49,
            "end": 53
        },
        {
            "id": 8,
            "form": "40년",
            "begin": 54,
            "end": 57
        },
        {
            "id": 9,
            "form": "동안의",
            "begin": 58,
            "end": 61
        },
        {
            "id": 10,
            "form": "역사적",
            "begin": 62,
            "end": 65
        },
        {
            "id": 11,
            "form": "상황을",
```

```
            "begin": 66,
            "end": 69
        },
        {
            "id": 12,
            "form": "재조명하는",
            "begin": 70,
            "end": 75
        },
        {
            "id": 13,
            "form": "특별전",
            "begin": 76,
            "end": 79
        },
        {
            "id": 14,
            "form": "'문화재에",
            "begin": 80,
            "end": 85
        },
        {
            "id": 15,
            "form": "깃든",
            "begin": 86,
            "end": 88
        },
        {
            "id": 16,
            "form": "100년",
            "begin": 89,
            "end": 93
        },
        {
            "id": 17,
            "form": "전",
            "begin": 94,
            "end": 95
        },
        {
            "id": 18,
            "form": "그날'을",
            "begin": 96,
            "end": 100
```

```
            },
            {
                "id": 19,
                "form": "개최한다.",
                "begin": 101,
                "end": 106
            }
        ],
        "NE": [
            {
                "id": 1,
                "form": "문화재청",
                "label": "OGG_POLITICS",
                "begin": 0,
                "end": 4,
                "kid": "04110000001377",
                "wikiid": "72070",
                "URL":    "https://ko.wikipedia.org/wiki/%EB%8C%80%ED%95%9C%EB%A
                          F%BC%EA%B5%AD_%EB%AC%B8%ED%99%94%EC%9E%AC%EC%
                          B2%AD"
            },
            {
                "id": 2,
                "form": "서대문형무소역사관 제10·12옥사",
                "label": "LC_OTHERS",
                "begin": 6,
                "end": 24,
                "kid": "05120000004788",
                "wikiid": "NA",
                "URL": "NA"
            },
            {
                "id": 3,
                "form": "1910년",
                "label": "DT_YEAR",
                "begin": 27,
                "end": 32,
                "kid": "07040000000284",
                "wikiid": "9402",
                "URL": "https://ko.wikipedia.org/wiki/1910%EB%85%84"
            },
            {
                "id": 4,
                "form": "대한민국임시정부",
```

```
                    "label": "OGG_POLITICS",
                    "begin": 40,
                    "end": 48,
                    "kid": "04110000001207",
                    "wikiid": "9889",
                    "URL": "https://ko.wikipedia.org/wiki/%EB%8C%80%ED%95%9C%EB%AF%BC%EA%B5%AD_%EC%9E%84%EC%8B%9C%EC%A0%95%EB%B6%80"
                },
                {
                    "id": 5,
                    "form": "40년 동안",
                    "label": "DT_DURATION",
                    "begin": 54,
                    "end": 60,
                    "kid": "07000000003384",
                    "wikiid": "NA",
                    "URL": "NA"
                },
                {
                    "id": 6,
                    "form": "문화재에 깃든 100년 전 그날",
                    "label": "EV_OTHERS",
                    "begin": 81,
                    "end": 98,
                    "kid": "10040000002071",
                    "wikiid": "NA",
                    "URL": "NA"
                }
            ]
```


※ "kid": 국립국어원 개체 연결 표현 고유 번호
"wikiid": 위키피디아 문서 번호
"URL": 위키피디아 주소(분석 대상 개체 표현에 대응하는 위키피디아 문서가 없는 경우 NA)

· **자료 내용 문의:** 02-2669-9638