

국립국어원 개체명 사전 2021

(버전 1.1)

· 자료명: 국립국어원 개체명 사전 2021

· 공개일

- (버전 1.0) 2022. 9. 16.
- (버전 1.1) 2022. 12. 30.

· 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

- 개체명 및 개체 연결 정보가 부착된 말뭉치 약 1,100만 어절에서 개체 표현, 개체 유형, 지식베이스 연결 정보를 추출하여 구축함.
- 개체 유형은 ‘2022년 개체명 분석 말뭉치 구축 지침 ver. 2.5.’(‘2021년 개체명 분석 및 개체 연결 말뭉치 연구 분석’ 보고서, 2021)에 따라 150개 세분류 체계로 구축된 분석 말뭉치가 기준임.
- 개체 연결은 PERSON(PS)과 LOCATION(LC), ORGANIZATION(OG), ARTIFACTS(AF), DATE(DT)의 세분류를 연결 대상으로 하며, 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. 지식베이스 미연결 개체 표현의 경우, 개체 연결 관련 정보는 모두 NA로 표시됨.

· 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, 신문 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(신문 300만, 구어 300만 어절)
- 파일 형식: JSON(UTF-8 인코딩)
- 파일 수 및 크기: 파일 1개, 총 61.5MB(ZIP 파일 기준)
- 인용:
 - (국문) 국립국어원(2022). 국립국어원 개체명 사전 2021(버전 1.1). URL: <https://kli.korean.go.kr/corpus>
 - (영문) National Institute of Korean Language (2022). NIKL Named Entity Dictionary 2021 (v.1.1). URL: <https://kli.korean.go.kr/corpus>

· 파일 명명 규칙

자리	1	2	3	4	5	6	7	8	9	10	11	12	13	14
속성	유형	구성 방법	주석 단계	구축 연도	일련번호(8자리)									
정의 값	X: 복합 (신문 웹 구어)	X: 추출	DX: 사전	21: 2021년	00000001									
※ 예시: XXDX2100000001.json 2021년에 구축한 복합 자료 가공 개체명 사전 파일														

· 예시

```

      "id": "0000163791",
      "form": "강남순환도시고속도로",
      "ne": [
        {
          "label": "LC_OTHERS",
          "kid": "05120000000953",
          "wikiform": "강남순환로",
          "definition": "강남순환로(江南循環路)는 서울특별시 금천구 시흥동 소하 분기점과 서초구 우면동 선암 나들목을 잇는 서울특별시의 자동차 전용도로이다".
        }
      ]
    
```

```

        "wikiLink":
"https://ko.wikipedia.org/wiki/%EA%B0%95%EB%82%A8%EC%88%9C%ED%99%98%EB%A1%9C",
        "SynonymGroup": [],
        "examples": [
            {
                "id": "NWRW1800000044.183.10.3",
                "form": "강남순환도시고속도로의 서초터널 발파로 인한 영향은
미미한 수준으로 판단했다.",
                "NE_form": "강남순환도시고속도로",
                "begin": 0,
                "end": 10
            }
        ]
    ],
},

```

- ※ "id": 국립국어원 개체명 사전 고유 번호(0000000000번부터 시작)
- "form": 개체명 형태
- "ne": 개체명 정보
- "label": 개체명 분석 표지
- "kid": 국립국어원 개체 연결 표현 고유 번호
- "wikiform": 위키피디아 문서 제목
- "definition": 위키피디아에서 정의한 정보
- "wikilink": 연결된 위키피디아 문서(분석 대상 개체 표현에 대응하는 위키피디아 문서가 없는 경우 NA)
- "SynonymGroup": 동일한 위키피디아 문서에 연결된 유의어 개체명
- "examples": 개체명 분석 말뭉치에서 해당 개체명이 나타난 문장(최대 5개 표시)

· 자료 내용 문의: 02-2669-9638