KoPDP Overview

Ver.1.0

5/31/2019

- 본 Dataset은 "한국 특허 데이터 프로젝트(Korea Patent Data Project, KoPDP)"의 결과이다. KoPDP는 한국특허청(KIPO) 및 미국특허청(USPTO) 특허데이터를 STATA dta 파일로 구축하고 각 특허데이터의 출원인을 DataGuide의 기업식별코드와 매칭하였다.[1]

  This dataset is the result of "Korea Patent Data Project (KoPDP)", providing a comprehensive and accessible dataset of South Korea & US patents from KIPO and USPTO bulk data together with their assignees matched to DataGuide, a Korean firm-level database. All data files are in STATA dta format.

- Dataset은 "CODES" 폴더와 "DATA" 폴더로 구성되어 있으며, "CODES" 폴더에는 "DATA" 폴더에 있는 '.dta' 파일을 만드는 데에 쓰인 '.py' 및 '.do' 파일이 담겨있다.

  The Dataset consists of two folders: "CODES" and "DATA". The "CODES" folder contains all '.py' and '.do' files used for constructing the '.dta' files in "DATA".

- DataGuide 기업식별코드와 KIPO 및 USPTO 출원인 간의 매칭 결과는 "DATA" 폴더 'matching_table.dta' 파일에 해당한다.

  The matching result of DataGuide firm identifiers and the assignees from KIPO and USPTO is 'matching_table.dta' which can be found in the "DATA" folder.

- KIPO 데이터는 저작권 문제로 업로드하지 않았으나 KoPDP의 KIPO 데이터 구축을 위한 모든 코드는 "CODES" 폴더에 업로드 되어있다. KIPO 데이터 구축을 위해서는 KIPRIS Plus Open API 구매가 필요하다.

  KIPO '.dta' files are not uploaded due to copyright issues but all the construction codes can be found in the "CODES" folder. For constructing KIPO '.dta' files, one may purchase "KIPRIS Plus Open API".

- 이 문서는 본 Dataset의 전반적인 사용법에 관한 설명이며 "CODES" 폴더에는 각 폴더 별로 'readme.txt' 파일에 '.py' 및 '.do' 파일의 상세한 사용법이 적혀있다.

  This document provides an overview of how to use or replicate the KoPDP dataset. Each sub-folder of "CODES" folder has 'readme.txt' file for the specific '.py' and '.do' files in the corresponding folder.

---

[1] KoPDP에 관한 자세한 내용은 이지홍 외. (2019) 참조.

1. 한국특허청 특허 자료

한국특허청 특허 자료는 본 Dataset에 업로드 되어 있지 않지만, "CODES" 폴더를 통해 구축할 수 있다. 모든 데이터 파일은 특허의 출원번호(appnum)를 기준으로 정리된다.

(1) basicinfo.dta, assignee.dta, invt_loc.dta, ipc.dta

- 관련 코드: CODES > 1_KIPRIS > Biblio
- 1948-2016년 동안 KIPO에 출원된 특허의 출원번호 별로 출원 및 등록 시점과 등록여부(baiscinfo.dta), 출원인 정보(assignee.dta), 발명자 정보(invt_loc.dta), IPC 정보('ipc.dta')에 관한 데이터가 정리되어 있다.[2]
- Biblio 폴더의 '.py' 코드로 KIPO XML 파일을 파싱하여 '.csv' 파일을 만든 후 STATA에서 바로 '.dta' 파일로 변환하거나 Biblio 폴더의 '.do' 파일을 이용하여 '.dta'로 변환하여 사용한다.
- 각 특허의 출원번호를 기준으로 STATA의 'merge' 명령어를 사용하여 모든 데이터를 하나의 '.dta' 파일에 모을 수 있다. 이 때 'basicinfo.dta' 파일에는 출원번호 하나 당 한 개의 observation만이 존재하지만 'assignee.dta', 'invt_loc.dta', 'ipc.dta'에는 출원번호 하나 당 한 개 이상의 observation이 존재한다.[3]
- 특히 'ipc.dta'는 'industry.dta'와 결합하여 각 특허의 산업 분류를 식별할 수 있다. KoPDP에서는 특허의 IPC를 기준으로 특허의 산업을 NBER PDP를 따라 분류한다.[4] 'IPC-산업분류' 결과는 'industry.dta'에 있으며 'ipc.dta'에서 각 특허의 출원번호 별로 subclass 수준에 해당하는 IPC 변수를 만들어(STATA 명령어: gen subclass = substr(ipc,1,4)) 'industry.dta'와 결합하여 특허의 산업을 식별할 수 있다.[5]

---

[2] IPC에 관한 자세한 내용은 이지홍 외. (2019) 2절 6항 참조.

[3] 하나의 특허에 여러 명의 출원인 혹은 발명자가 존재할 수 있기 때문이다.

[4] 산업 분류에 관한 자세한 내용은 이지홍 외. (2019) 6절 참조.

[5] 'industry.dta'는 이지홍 외. (2019) 6절에서 설명하고있는 과정을 거쳐 수작업으로 만든 표를 STATA '.dta' 파일로 옮긴 것이다.

(2) citation.dta

- 관련 코드: CODES > 1_KIPRIS > Citation
- 2002-2016년에 등록된 KIPO 특허들의 인용 관계에 대한 정보이다.
- Citation 폴더의 '.py' 코드로 KIPO XML 파일을 파싱하여 '.csv' 파일을 만든 후 '4. citation_byappnum' 폴더에서 2002-2016년에 등록된 KIPO 특허를 추려 출원번호 기준으로 정리한 '.dta' 파일을 만들어 사용한다. XML 파일은 출원번호를 기준으로 인용 관계가 정리되어 있지 않다.
- 'citing' 변수는 인용한 특허의 출원번호이고, 각 'citing' 변수에 대응되는 'cited_num' 변수는 해당 'citing' 특허의 인용을 받은 특허의 출원번호이다. 인용 순서('order')는 중요도와 같은 특별한 의미가 있는 변수는 아니며, 표준인용식별코드('cited_code')와 피인용특허 인용구분 코드('cited_typecode')로 원하는 종류의 인용 관계를 한정할 수 있다.
- 'citing' 혹은 'cited_num' 변수의 변수명을 'appnum'으로 바꾸어 'basicinfo.dta', 'assignee.dta', 'invt_loc.dta' 등의 파일에 있는 기본 정보, 출원인 정보, 발명자 정보와 결합하여(merge) 사용할 수 있다.

(3) family.dta

- 관련 코드: CODES > 1_KIPRIS > Family
- 각 출원번호 별로 관련 패밀리특허에 대한 정보가 정리되어 있다.
- Family 폴더의 '.py' 코드로 KIPO XML 파일을 파싱하여 '.csv' 파일을 만든 후 STATA에서 '.dta' 파일로 변환하여 사용한다.
- 한국특허청 패밀리특허 정보(familykind == "FAMILY")와 EPO 패밀리특허 정보(familykind == "DOCDBFAMILY")가 모두 담겨있으며, 그 외에 패밀리 특허의 국가 및 특허의 종류('litkind') 등이 정리되어 있다. 변수에 대한 자세한 내용은 이지홍 외. (2019) 참조.
- 이 데이터는 이후 USPTO 출원인을 DataGuide 기업식별코드에 매칭하는 작업에 사용된다.

(4) RND.dta

- 관련 코드: CODES > 1_KIPRIS > RnD

- RnD 폴더의 '.py' 코드로 KIPO XML 파일을 파싱하여 '.csv' 파일을 만든 후 STATA에서 '.dta' 파일로 변환하여 사용한다.
- 각 출원번호 별로 국가에서 해당 특허와 관련된 연구비를 받은 연구 과제에 관한 내용이 정리되어 있다. 각 연구 과제의 과제 번호('tasknum') 연구부처('ministry')와 주관기관('inst') 및 연구사업명('project')과 연구과제('task') 등이 정리되어 있다.
- 마찬가지로 'appnum'을 이용하여 다른 '.dta' 파일과 결합하여 사용할 수 있으며, 복수의 연구사업에서 연구비를 받는 특허도 존재한다.

2. 미국특허청 특허 자료

미국특허청 특허 자료는 크게 USPTO 전체 특허데이터와 USPTO 한국 특허데이터 2 가지가 존재한다. USPTO 전체 특허데이터는 미국특허청에 1976-2017년 기간 동안 등록된 모든 실용 특허(utility patent)에 대한 데이터이고, USPTO 한국 특허데이터는 출원인 혹은 발명자의 국가 정보가 한국에 해당하는 특허를 따로 추려 놓은 것이다.[6] USPTO 전체 특허데이터 '.dta' 파일에는 'all'이라는 접미사가, 한국 특허데이터 '.dta' 파일에는 'kr'이라는 접미사가 붙어있다. 미국특허청 전체 특허 자료는 '.dta' 파일로 가공하였으나 용량의 문제로 '.csv' 파일로 업로드 하였으며, 나머지는 모두 '.dta' 파일로 업로드 하였다. 각 '.dta' 파일 구축을 위한 관련 코드도 업로드 되어있다.

'assignee_uspto_all.dta'의 'engname2'는 1976-2015년 기간의 출원인 영문명으로, USPTO에서 제공하는 'ASG NAMES UPRD 69 15NUMSORT.txt' 파일에서 가져왔으며 해당 파일의 첫 번째 열에 있는 변수는 USPTO에서 부여한 출원인 코드('assignee harmonization code')에 해당한다. 이를 USPTO에서 제공하는 'PN_ASG_UPRD_69_15.txt' 파일의 '출원번호-출원인 코드' 정보를 이용하여 출원번호 별로 정리하였다. 또한 1976-2015년 기간의 출원인 국적 및 주소 정보는 USPTO에서 제공하는 'ASSIGNEE_FULLFILE_14.txt'에서 가져왔다. 특히 출원인 영문명에는 띄어쓰기 혹은 철자 등에 약간의 차이가 있을 수 있으나 모든 정보는 'CODES > 2_USPTO'에 있는 코드를 이용한 XML 파싱 결과를 사용해도 무방하다.

한편, "CODES > 2_USPTO" 폴더의 XML 파싱(parsing) '.py' 코드는 2016-2017년 기간의 XML 파일에 최적화 되어있다. USPTO XML bulk data는 tagging scheme이 기간별로 달라지는 경우가 존재한다. 예컨대 2015년까지는 특허 분류체계를 USPC (US Patent Classification)를 사용하다가 이후 CPC (Cooperative Patent Classification)로 변경하면서 XML 파일에 해당 정보의 위치가 변경되었다. 따라서 2016-2017년 기간 외에 다른 기간의 XML 파싱을 진행할 때에는 XML 파일을 확인하여 파싱 위치를 수정해주어야 한다. "CODES > 2_USPTO > 5.ipc"에 IPC 정보를 파싱하는 코드를 기간 별로 다르게 업로드하였다. 이를 참조하여 다른 정보를 파싱할 때에도 코드를 기간 별로 다르게 수정하여야 한다.

---

[6] 미국의 실용 특허(utility patent)가 한국의 특허에 해당한다. 본문에서는 구분 없이 '특허'라는 용어를 사용하였다.

(1) baicinfo_uspto_all.dta, basicinfo_uspto_kr.dta

- 경로: DATA > USPTO
- 관련 코드: CODES > 2_USPTO > 1. basic.py
- 1976-2017년 기간 동안 미국특허청에 등록된 특허의 출원번호 별로 출원 및 등록일자가 정리되어 있다. 'basicinfo_uspto_kr.dta'의 한국 특허는 각 특허의 첫 번째 출원인의 국가 정보가 "KR"(한국)인 경우를 기준으로 추려 내었다.
- '1. basic.py' 코드로 USPTO XML 파일을 파싱하여 '.csv' 파일을 만든 후 STATA '.dta' 파일로 읽어서 이 중 출원번호('wku'), 등록일자('isd'), 출원일자('apd')를 남긴다. 한국 특허를 추리기 위한 국가 정보는 'assignee_uspto_all.dta'를 만든 후 출원번호('wku')와 출원인 국가('cnt')를 이용하여 첫 번째 출원인 국가 정보가 "KR"인 경우의 출원번호('wku')를 추려서 이를 이용하여 'basicinfo_uspto_kr.dta'를 만들었다.
- KIPO 데이터와 마찬가지로 출원번호를 기준으로 다른 데이터와 결합하여 이용할 수 있다.

(2) assignee_uspto_all.dta, assignee_uspto_kr.dta

- 경로: DATA > USPTO
- 관련 코드: CODES > 2_USPTO > 1. basic.py & 2. assg_location.py
- 1976-2017년 기간 동안 미국특허청에 등록된 특허의 출원번호 별로 첫 번째 출원인의 출원인 ID, 영문명 그리고 주소 정보가 정리되어 있다 (국가, 주, 도시). 'assignee_uspto_kr.dta'의 한국 특허는 각 특허의 첫 번째 출원인의 국가 정보가 "KR"(한국)인 경우를 기준으로 추려 내었다.
- 'assignee_uspto_all.dta': '1. basic.py'을 이용한 파싱 결과로부터 출원번호('wku')와 출원인 영문명1('engname1')을 가져왔으며 '2. assg_location.py'을 이용한 파싱 결과로부터 출원인 종류('asscode'), 출원인 국가('cnt'), 출원인 주('sta'), 그리고 도시 단위의 출원인 주소('city')를 가져왔다. 출원인 영문명2('engname2')는 앞서 언급한 바와 같이 USPTO가 제공하는 'ASG NAMES UPRD 69 15NUMSORT.txt'로부터 가져왔으며 2015년까지만 데이터가 존재한다. 출원인 ID('assg_id')는 Lee and Lim (2019)에서 1976-2015년 기간 동안의 출원인 영문명2와 2016-2017년

기간 동안의 출원인 영문명1을 이용하여 USPTO 출원인을 식별하는 고유의 ID를 부여한 결과를 가져왔다.[7] 이 ID는 'assignee_uspto_kr.dta'의 출원인 ID('assgid')와는 전혀 다른 식별 번호이다.[8]

- 'assignee_uspto_kr.dta': 'assignee_uspto_all.dta'에서 출원인 국가('cnt')가 "KR"에 해당하는 경우를 추린 것이다. 'assignee_uspto_kr.dta'의 출원인 영문명('engname')은 'assignee_uspto_all.dta'의 출원인 영문명 1('engname1')에 해당한다. 'assignee_uspto_kr.dta'의 출원인 ID('assgid')는 KoPDP에서 출원인 영문명('engname')을 기반으로 부여한 출원인 고유의 식별 코드이다.[9]
- 'assignee_uspto_kr.dta' 파일은 이후 한국 출원인을 이름 및 법인등록번호, 패밀리정보 등을 이용하여 DataGuide 기업식별코드와 매치하는 데에 사용하였다.

(3) invt_loc_uspto_all.dta, invt_loc_uspto_kr.dta

- 경로: DATA > USPTO
- 관련 코드: CODES > 2_USPTO > 4. invent.py
- 1976-2017년 기간 동안 미국특허청에 등록된 특허의 출원번호 별로 첫 번째 발명자의 영문명 그리고 주소 정보가 정리되어 있다 (국가, 주, 도시). 'invto_loc_uspto_kr.dta'의 한국 특허는 각 특허의 첫 번째 발명자의 국가 정보가 "KR"(한국)인 경우를 기준으로 추려 내었다. 이 때 발명자 주소 정보에 띄어쓰기 혹은 철자 등에 오류가 많아 2002- 2017년 기간 동안 등록된 특허를 추려서 수작업으로 주소 정보를 정리한 후 추가로 광역자치단체 코드('dist') 및 광역자치단체 이름('name') 변수를 구축하였다.
- '4. invent.py' 코드로 USPTO XML 파일을 파싱하여 '.csv' 파일을 만든 후 STATA '.dta' 파일로 변환한다. 이후 'invt_loc_uspto_kr.dta'의 경우

---

[7] 출원인 ID 부여와 관련된 자세한 내용은 Lee and Lim (2019) 부록 참조.

[8] 특히 Lee and Lim(2019)에서는 한국 기업들에 대한 출원인 이름 정리 혹은 식별 코드 정리에 주목하지 않기 때문에 한국 기업의 출원인 ID가 KoPDP에 비해 부정확하다.

[9] 자세한 내용은 이지홍 외. (2019) 5절 참조. 매칭 과정에서 출원인 ID 변수를 구축한 후 출원번호를 기준으로 데이터 파일에 해당 변수를 추가하였다.

'cnt'가 'KR'인 경우를 남기고 'wku'로 'basicinfo_uspto_all.dta'에서 해당 'wku'의 날짜 정보를 가져와서 등록일자('isd')가 2002-2017년에 해당하는 특허를 남겼다. 그리고 해당 특허의 'cty' 변수에 있는 주소 정보를 토대로 수작업으로 광역자치단체 코드('dist') 및 광역자치단체 이름('name') 변수를 구축하였다.

(4) citation_uspto_all.dta

- 경로: DATA > USPTO
- 관련 코드: CODES > 2_USPTO > 3. cite.py
- 1976-2017년 기간 동안 등록된 USPTO 특허들이 인용한 USPTO 특허에 대한 정보가 정리되어 있다. 인용한 특허의 출원번호는 'citing'으로 인용을 받은 특허의 출원번호는 'cited'로 정리되어 있다.
- '3. cite.py' 코드로 USPTO XML 파일을 파싱하여 '.csv' 파일을 만든 후 '6. csv2dta_sample.do'에서와 같이 STATA에서 '.dta' 파일로 읽어 실용 특허를 남겨 정리한다.
- KIPO의 경우와 같이 'citing' 혹은 'cited'의 변수명을 'wku'로 변경하여 출원번호를 기준으로 다른 데이터 파일과 결합하여 사용할 수 있다.

(5) ipc_uspto_all.dta

- 경로: DATA > USPTO
- 관련 코드: CODES > 2_USPTO > 5. ipc
- '5. ipc' 폴더에는 기간 별로 XML 파일의 구조가 달라지는 것을 고려하여 1976-2017년 기간 동안 등록된 USPTO 특허들의 IPC 정보를 파싱하는 '.py' 파일이 담겨있다. 이를 이용하여 출원번호 별로 IPC 정보가 담긴 '.csv'를 만든 후 이를 하나로 합쳐 '.dta' 파일로 변환하여 사용한다.
- KIPO에서와 마찬가지로 'ipc_uspto_all.dta'의 ipc 변수를 토대로 subclass 수준에 해당하는 IPC 변수를 만들어(STATA 명령어: gen subclass = substr(ipc,1,4)) 'industry.dta'와 결합하여 특허의 산업 분류를 식별할 수 있다.

3. ‘DataGuide – 특허 출원인’ 매칭

KoPDP에서는 KIPO 출원인과 USPTO 한국 출원인을 한국 기업데이터인 DataGuide의 기업식별코드(Symbol)에 매치하였다. 본 Dataset은 KoPDP의 매칭 결과 및 매칭 작업과 관련된 모든 코드를 제공한다. 자세한 매칭 방법 및 알고리즘은 이지홍 외. (2019) 5절에 상세히 설명되어 있다.

KoPDP의 매칭 작업은 크게 ‘DataGuide 기업 – KIPO 출원인’ 매칭과 ‘DataGuide 기업 – USPTO 출원인’ 매칭으로 나누어진다. ‘DataGuide 기업 – KIPO 출원인’ 매칭은 크게 법인등록번호 이용 단계와 문자열 알고리즘 이용 단계로 나뉘며, ‘DataGuide 기업 – USPTO 출원인’ 매칭은 크게 패밀리특허정보 이용 단계와 문자열 알고리즘 이용 단계로 나뉜다. 모든 매칭 관련 코드는 “CODES > 3_MATCHING”에 담겨있으며 각 세부 폴더에 상세한 ‘readme.txt’ 파일이 있다.

(1) CODES > 3_MATCHING > 1. corpnum 폴더

- ‘DataGuide 기업 – KIPO 출원인’ 매칭에서 법인등록번호를 이용한 매칭 단계에 관련된 코드

(2) CODES > 3_MATCHING > 2. family 폴더

- ‘DataGuide 기업 – USPTO 출원인’ 매칭에서 패밀리특허정보를 이용한 매칭 단계에 관련된 코드

(3) CODES > 3_MATCHING > 3. stringmatching 폴더

- kipris 폴더: ‘DataGuide 기업 – KIPO 출원인’ 매칭에서 문자열 알고리즘을 이용한 매칭 단계에 관련된 코드
- uspto 폴더: ‘DataGuide 기업 – USPTO 출원인’ 매칭에서 문자열 알고리즘을 이용한 매칭 단계에 관련된 코드

(4) CODES > 3_MATCHING > 4. cleaning 폴더

- (1)-(3)의 결과를 토대로 중복 처리 및 출원인 식별 ID 정리 등을 통해 ‘matching_table.dta’를 만드는 코드