READ ME – DOCUMENTATION STANDARDS COUNTS

This document describes the compilation of the database used to extract time series counts for standard documents as in Baron and Schmidt (2019). If you are using these time series, please cite Baron and Schmidt (2019).

We provide to all users a baseline time series of standard counts generated using the methodology described in our paper. It can be found in the folder Output_data. This baseline time series is the one used in the main analysis of Baron and Schmidt (2019). Users wanting to modify any of the parameters of the time series (such as technology class, international or national scope, weighting) to generate alternative time series using the same raw data need access to the Searle Center Database. Please follow the database access instructions and download the respective folders which contain the data to be used. In the following, we comment on the code that we provide to these users in order to assist with generating time series.

The files are written in R. Users need to install the packages openxlsx, plyr, here and readstata13.

Users who only want to access the baseline time series do not need access to the Searle Center Database.

FILE STRUCTURE:

There are two main R files that users can work with to adapt the code to their needs.

1. **Standards_database.R**
   This file compiles the database using the Searle Center Database as its main input.
2. **Standards_time_series.R**
   This file extracts the time series from the compiled database which is saved in Backup_data. Users can change the ICS classes in line 13 and the geographical dimension in line 14.

There are also two sub-files, Aggregation_time_series.R and SSO_names_harmonization.R (saved in the subfolder R_files), which are used by the main files.

There are three data subfolders:

1. Input_data
   All input data files, with the exception of the files from the Searle Center Database on technology standards and standard setting organizations, are stored in this folder.
   - SSO_manual_designation.xlsx: a spreadsheet where certain SSOs are assigned one or several ICS classes.
   - data_perinorm_eq.RData: a dataset that stores information about standard equivalences from Perinorm.
   - SSOs.dta: a dataset that stores information about the "nationality" of SSOs.
2. Backup_data
   The compiled database (Standards_database.RData) is stored in this folder.
3. Output_data
   The generated time series are saved in this folder.

In addition, the codes use the following data files from the Searle Center Database (folder SSO):

- SCDB_ics codes.dta
- SCDB_replacement.dta

- SCDB_references.dta
- SCDB_standards.dta
- SCDB_document_identifiers.dta

DATA COMPILATION PROCEDURE (STANDARDS_DATABASE.R):

Step 1: Cleaning

- Only keep unique values.
- SCDB_standards.dta (data_scdb_base):
  - There are 62 ETSI standards with the same SCDB id, but sometimes different dates. Solution: delete 33 ETSI standards. There are a total of 797749 unique standard document ids.
- SCDB_replacement.dta (data_scdb_vh):
  - Observations where replaced and replacing id are the same are deleted.
  - There are some observations that are recorded in the wrong way: the replaced doc_idn and doc_idn should be switched. This issue is taken care of by merging the publication date to the datset and switching the appropriate column entries.
- SCDB_ics codes.dta (data_scdb_ics):
  - ICS codes starting with "RW" are deleted.
  - ICS codes are aggregated to 5-digit level.
- SSO_manual_designation.xlsx (sso_manual):
  - The input from this dataset is transformed into an exploitable format.

Step 2: Merge equivalents from Perinorm

We have information on standard equivalences from Perinorm (data_perinorm_eq) that we merge via SCDB_document_identifiers.dta (data_scdb_docid) to the dataset. All equivalent standards share the same group id.

Step 3: Identify first versions

Standards are constantly updated and replaced by newer versions. Those standards that are not in the column "doc_idn" of SCDB_replacement.dta (data_scdb_vh) are considered to be the first version. In Baron and Schmidt (2019), these first versions are the ones used to identify discontinuous technological change.

Step 4: Obtain number of references within 4 years of publication

We count the number of times a standard is being referenced by other standards (these counts take into account that there are standard equivalences; we avoid double counting by grouping referencing standards via group id).

Step 5: Merge ICS info to database and fill missing info where possible

A total of 369926 standard document ids do not have ICS information (at the 5-digit level). A large number of these come from "Military and Government Specs & Standards (Naval Publications and Form Center – NPFC)".

We fill some of these missing data points by making use of the SSO that releases the standard. In particular, we use the following information:

- If the SSO is in the database and usually (i.e. at least 95% of the time) receives a certain ICS class, we also designate this ICS class to the standards issued by the respective SSO, but which does not have an ICS class in the original dataset.
- Manual designation of SSOs via SSO_manual_designation.xlsx (sso_manual), where possible.
- ETSI standards that do not have an ICS class are categorized at the 2-digit level, but receive the 5-digit class "ETSI".

Step 7: Quarterly indicator

The dataset SSOs.dta is matched to the database (via the variable issuingbody) in order to designate the geographical dimension of a certain SSO whenever it is not provided by Perinorm.

Step 7: Quarterly indicator

All standard counts are aggregated on the quarterly level: a time variable (time_q) is created which indicates the year and quarter of the publication date. Whenever monthly (and thus quarterly) information is missing, the quarter "5" is created and all standards of a given year are uniformly distributed.

Step 8: Some cleaning

- Observations without year info are deleted.
- Country distribution: Assume that missing country code implies US. (Origin code is missing for standard observations retrieved from sources other than Perinorm, such as IHS and Document Center, which focus on US standards. See Baron and Spulber, 2018.)
- Drop observations attributed to WATTC. (The World Administrative Telegraph and Telephone Conference, WATTC, in 1988 aimed at the international harmonization of telecommunication standards and led to the inclusion of a large number of already existing national standards in the ITU standard catalogue.)
- Divide number of pages by 10.
- All reference counts after 2011-Q4 receive NA values as we do not have enough data for four years.

Step 9: Create an id to identify the earliest publication within a group of equivalent standards

If a group of equivalent standards includes publications with a different nationality code ("IX" or "US" or both), the date of publication of the earliest standard in the group may differ depending on the geographic scope of the time series. Example: A standard was first published by an international SSO in 1999 and subsequently by a US-based SSO in 2001. The standard will be included in the international time series in 1999, in the US time series in 2001, and a time series including both US and international standards will use 1999 as the publication date. To this end, an identifier is created in order to identify the first publication of a group of equivalent standards.

TIME SERIES EXTRACTION (STANDARDS_TIME_SERIES.R):

Users can change the two-digit ICS class that the time series covers. More than one value is possible.

Users can change the geographical dimension by specifying if the releasing SSO is an international one ("IX"), US-based ("US") or both.

The output is saved in Output_data and the file name indicates the chosen ICS and geographical dimension.

REFERENCES

Justus Baron and Julia Schmidt (2019). Technological Standardization, Endogenous Productivity and Transitory Dynamics, mimeo.

Justus Baron and Daniel F. Spulber (2018). "Technology Standards and Standard Setting Organizations: Introduction to the Searle Center Database," *Journal of Economics & Management Strategy*, vol. 27(3), p. 462-503.