CAPPA – A Dataset on the Career Paths of EPO Patent Attorneys

DOI: 10.5281/zenodo.3265385

Authors: Kazimir Menzel[1], Lutz Maicher[1,2]

Affiliations:
[1] Technology Transfer Research Group, Friedrich-Schiller-University Jena, Ernst-Abbe-Platz 2, D-07743 Jena, Germany

[2] Fraunhofer-Center for International Management and Knowledge Economy, Neumarkt 9-19, 04109 Leipzig, Germany

Contact email: lutz.maicher@uni-jena.de

Abstract

The CAPPA dataset contains anonymised information on the career paths of patent attorneys that are registered with the European Patent Office (EPO) tracking their employment from 2008 to 2018. It consists of four different sub-datasets (Attorney Data Frame, Employer Data Frame, Change Data Frame, and Career Path Data Frame) that are composed for specific information needs. The dataset has been computed, cleaned and made consistent based on publicly available patent records that are available through the EPO.

Value of the data

- The CAPPA dataset contains anonymised information on the career paths of patent attorneys registered with the European Patent Office, derived from all patent applications at the European Patent Office, published since 2008.
- The CAPPA dataset is to be updated every six months, released as new versions.
- The dataset and all its versions are provided as open data under the Creative Commons Attribution 4.0 International license.

Data

Patents are an important source of open data, heavily used in innovation economics for research and, through patent information systems, by practitioners. While innovation activities are extensively investigated using patent data, its usage for research about the involved patent attorneys is limited. A major reason for this is that patent data requires significant harmonisation [1], which poses a high barrier to research. In order to lower this barrier somewhat, we created the CAPPA dataset.

Based on a stringent harmonisation of the patent attorney and patent firm entries retrieved from the EP patent applications (against identifiers of the IP Industry Base[1]), the dataset contains the anonymised affiliations of patent attorneys with patent law firms and technology companies on a monthly basis since January 2008. This information is derived from all available patent applications before the European Patent Office (EPO) over the last 12 years that contain sufficient information.

The dataset was computed based on 922,959 patent applications that contained sufficient information. From these, career paths for 8,710 patent attorneys could be recorded, which were affiliated to a total of 2930 employers. The data used covers a period of 12 years, from 01/01/2008 to 31/12/2018. In order to ensure anonymity of the individual patent attorneys, the attorneys and their employers have been anonymised and are accessible only through their unique identifier within the dataset.

Structure of the dataset

The dataset consists of four data frames. The first, the *Attorney Data Frame*, has the following structure:

attorney_id	n_emp	n_pat	first_observation	last_observation	months_active
1	1	5	2009-03	2009-11	5
2	2	166	2008-04	2018-11	67
…	…	…	…	…	…

- **attorney_ID**. A unique and anonymised identifier for the patent attorney that allows to track the individual patent attorney.
- **n_emp**. Number of employers with which the attorney was affiliated over the observation period.
- **n_pat**. Number of patents, which have been assigned to the patent attorney over the observation period.
- **first_observation**. The first month given in “yyyy-mm” format, in which the attorney has been observed on a patent.
- **last_observation**. The last month given in “yyyy-mm” format, which the attorney has been observed on a patent.
- **months_active**. The duration of an attorney’s career span according to patent applications given in months.

[1] http://s.fhg.de/IPIB

The second data frame, the *Employer Data Frame*, has the structure:

employer_id	**rank**	**type**	**n_att**	**mean_ret**	**mad_ret**	**n_pat**	**med_ppa**
1	878	small	1	125	0	22	22
2	769	small	1	100	0	15	15
…	…	…	…	…	…	…	…

- **employer_id**. A unique and anonymised identifier for the registered employers of patent attorneys that allows to track the individual employer. Note that also self-employed attorneys are tracked as employers.
- **rank.** A rank over all employers that is synthetically derived from the cumulative number of employed attorneys, a period-based measure and the activity (number of patents). It is intended to propose one way of natural scaling when working with the dataset and serves also to distinguish between types of patent attorney employers. The rank is in ascending order, i.e. the higher number implies the higher rank.
- **type**. Typology of employers of patent attorneys, distinguishing between technology companies (tech) and patent law firms and within the latter between large-, medium- and small-sized companies (*large*, *medium*, *small*).
- **n_att**. The number of patent attorneys that have been observed as affiliated with the employer over the entire observation period.
- **mean_ret**. The mean duration of retaining a patent attorney given in months.
- **mad_ret**. The median absolute deviation of retaining a patent attorney given in months.
- **n_pat**. The number of patents that have been assigned to the employer over the observation period.
- **med_ppa**. The median of patents assigned to employed attorneys.

The third data frame, the *Career Path Data Frame*, contains the actual career paths, consisting of the stations of employment for each attorney. In total, 10,786 such stations have been recorded and are available through the following structure:

attorney_id	**employer_id**	**first_observation**	**last_observation**	**n_pat**
1	1820	2008-10	2009-11	3
2	574	2008-01	2018-07	7
2	2853	2008-01	2018-07	22
…	…	…	…	…

- **attorney_id**. A unique and anonymised identifier for the patent attorney that allows to track the individual patent attorney. The attorney_id is the same as in the *Attorney Data Frame*.
- **employer_id**. A unique and anonymised identifier for the registered employers of patent attorneys that allows to track the individual employer. Note that also self-employed attorneys are tracked as employers. The ID is the same as in the *Employer Data Frame*.
- **first_observation**. The month in which an attorney has been first observed as affiliated with an employer, given in “yyyy-mm” format.

- **last_observation**. The month in which an attorney has been observed last as affiliated with an employer, given in “yyyy-mm” format.
- **n_pat**. The number of patents that have been assigned to the attorney and the employer during his employment period.

The fourth data frame, the *Change Data Frame*, contains the changes of employment, when one attorney has actually changed from one employer to another, including work for two employers in parallel. In total, 4707 such changes have been recorded. The *Change Data Frame* has the structure:

attorney_id	**time_of_chg**	**origin**	**origin_type**	**target**	**target_type**	**change_type**
7343	2011-05	1576	tech	458	small	tech -> small
4482	2008-11	1695	medium	559	medium	medium -> medium
…	…	…	…	…	…	…

- **attorney_id**. A unique and anonymised identifier for the patent attorney that allows to track the individual patent attorney. The attorney_id is the same as in the *Attorney Data Frame*.
- **time_of_chg.** The month, in which a change of employment has been observed given in the “yyyy-mm” format. The date is set to the first month, during which an attorney has been observed with his new employer.
- **origin**. The employer_id of the employer with which the attorney was associated before the change. The employer_id is the same as in the *Employer Data Frame*. Note that a new entrance is not recorded as a change, i.e. only changes from one employer to another are recorded. New entrances can be readily seen in the *Career Path Data Frame*.
- **origin_type**. The type of the original employer as given in the *Employer Data Frame*.
- **target**. The employer_id of the employer with which the attorney is associated after the change. The employer_id is the same as in the *Employer Data Frame*. Note that it is not recorded as a change, when an attorney is no longer observed, i.e. only changes from one employer to another are recorded.
- **target_type**. The type of the target employer as given in the *Employer Data Frame*.
- **change_type**. The type of change as a combination of origin and target

Limitations

Due to the scope of the data retrieval, the dataset suffers several limitations that should be taken into account when working with it.

- **Anonymity**. Due to the entry into force of the new GDPR, it is not possible to make explicit, for which firms an attorney worked at a given time as this would allow it easily to identify the attorneys. The authors are aware of the fact that this might cause some difficulties, when seeking to combine the information with other data.
- **Focus on patent applications**. As the dataset does not take into account other aspects of a patent attorney's work, the career paths are solely derived from patent application data. This means in turn that some of the quantitative variables of patent attorneys as well as their employers reflect only a fraction of the actual activity of the attorney. This also means that the career stations of attorneys with very few patent applications are not always exact nor complete. The authors decided to keep them in the dataset in order to deliver as complete a dataset as possible.
- **Parallel employment**. A not prevalent but also not uncommon phenomenon is that patent attorneys are found in parallel employment. While the reasons for this are manifold as has been determined during interviews with selected attorneys, it makes it often difficult to decide whether an attorney is found in a transition phase between two employments or working in parallel. In these cases, there are two changes recorded, one when the attorney is first observed with her new employer and when the attorney is last observed with her old employer. In cases of doubt, it is recommended to take the first date as authoritative.
- **Insufficient or ambiguous patent records**. As most of the information has been retrieved from EPO patent applications, it was for a sizeable subset not possible to either retrieve information on an attorney or her affiliated employer. The lack of such information led to an exclusion from the dataset so that approximately 2100 registered patent attorneys are not covered in this dataset. This also led to a reduction of patents that could be assigned to the attorneys and their employers, which means that not all patents from the covered period could be included in the quantitative measures in the data set.

Acknowledgements

The creation of the CAPPA dataset has been funded by the Academic Research Programme of the European Patent Office under the grant agreement number 2017/8405.

References

[1] Thoma, G., Torrisi, S., Gambardella, A., Guellec, D., Hall, B. H., & Harhoff, D. (2010). Harmonizing and combining large datasets-An application to firm-level patent and accounting data (No. w15851). National Bureau of Economic Research.