# Data anonymization patent landscape

**Mirjana Pejić Bach**[1], **Jasmina Pivar**[1,†] **and Ksenija Dumičić**[1]

[1]*Faculty of Economics and Business, University of Zagreb, J. F. Kennedy 6, 10 000 Zagreb, Croatia*
*E-mail:* ⟨*{mpejic, jpivar, kdumicic} @efzg.hr*⟩

**Abstract.** The omnipresent, unstoppable increase in digital data has led to a greater understanding of the importance of data privacy. Different approaches are used to implement data privacy. The goal of this paper is to develop a data anonymization patent landscape, by determining the following: (i) the trend in data anonymization patenting, (ii) the type of technical content protected in data anonymization, (iii) the organizations and countries most active in patenting data anonymization know-how; and (iv) the topics emerging most often in patent titles. Patents from the PatSeer database relating to data anonymization from 2001 to 2015 were analyzed. We used the longitudinal approach in combination with text mining techniques to develop a data anonymization patent landscape.
The results indicated the following. The number of single patent families is growing with a high increase after 2010, thus indicating a positive trend in the area of patenting data anonymization solutions. The majority of patenting activities relate to the G Physics section. Organizations from the USA and Japan assigned the majority of patents related to data anonymization. The results of text mining indicate that the most often used word in titles of data anonymization patents are "anonym*", "method", "data" and "system". Several additional words that indicated the most frequent topics related to data anonymization were: "equipment", "software", "protection", "identification", or "encryption", and specific topics such as "community", "medical", or "service".

**Keywords:** data anonymization, patent landscape, PatSeer, data mining, association rules, text mining

## 1. Introduction

Part of the difficulty of working with data that come from sensitive sources, such as health or financial data, is protecting the privacy of individuals or organizations based on such data. Such types of data need to be anonymized with some of the

---

† Corresponding author

data anonymization techniques and methods, which is a prerequisite of data utilization, while at the same time retaining the data privacy. Various means are used for data anonymization, which includes algorithms and physical equipment. Cormode and Srivastava [5] state that "the result of data anonymization is anonymized data which are, essentially, a set of possible worlds, one of which corresponds to the original data".

Data anonymization has been the subject of research and patenting activities in recent years. Development of data anonymization is being reinforced across many industrial applications. Hence, it is important that both researchers and entrepreneurs become aware of data anonymization patents cover the requirements of different companies, industries, and countries, which is the purpose of this paper. According to the World International Property Organization, the patent is an exclusive right granted for an invention - a product or a process which provides a new way of doing something or a new technical solution to a problem [15]. It offers the exclusive right to stop or prevent others from commercially making, using, distributing, importing or selling the patented invention without the patent owner's permission. Those rights are only valid in the country or region where a patent has been filed or granted [11]. Also, the protection is granted for a limited period, generally 20 years from the filing date of the application [15].

Many countries use national patent systems based on "world patent application" that are made under the Patent Cooperation Treaty [15]. The World Intellectual Property Organization maintains a database of published international patent applications, using the International Patent Classification system (IPC), that was established in 1971 by the Strasbourg Agreement, and is nowadays used in more than 100 countries worldwide [15, 9]. Additionally, there are two important classification systems employed by the largest patent offices, both based on the IPC: 1) The Cooperative Patent Classification (CPC) system - developed by the European Patent Office (EPO) and the United States of America, and 2) File Index (FI) – Japanese patent classification system [11].

The patent documents are highly structured, providing a rich source of information [9]. They contain fields such as patent title, description, simple family ID, publication/issue year, filing/application year, priority country code, assignee country, assignee original/inventors, IPC codes, CPC codes, FI codes and others. The patent landscape, also known as a Competitive Technical Intelligence Report, White Space Analysis or Technical Gap Analysis, is a study which uses a large set of patents data to extract useful information for understanding a particular field [14]. It aims to give an overview of a particular area and provide insights to decision makers. The insights can, for example, be the publication trend (time) of patents or filing trend (technology) of patents; the top assignees or the companies filing patents and the number of them; and the way patents are spread across countries [14]. Other approaches are also often used [7]. For example, Noh, Jo and Lee [10] focused on keywords strategies for patent analysis and offered guidelines

on the selection and processing of keywords for patent analysis, and Brügmann et al. [2] presented an operational prototype of a workbench for patent document analysis and summarization. Text mining and visualization based approaches had also been used for analyzing the patent content in the vast body of literature [1]. Numerous researchers have developed the patent landscape for different technology fields. Some examples will be provided with a brief presentation of the methodology. Han and Sohn [8] identified technological convergence in standards related to information and communication technology. They applied social network analysis and association rules analysis. Choi and Hwang [4] analyzed the patents related to light emitting diodes and wireless broadband fields by using trend analysis, and a method that combines network-based and keyword-based research. The patent analysis was used to explore virtualization technology development in the USA [13], analyzing technology life cycle, assignee organization and country, patent classification and patents citations. In [3], authors investigated technological pervasiveness and a variety of innovators in Green ICT, using network analysis.

To the best of our knowledge, there has not been an analysis of the patents related to data anonymization. As an attempt to develop a patent landscape of data anonymization draws closer, this study is expected to help understand this area, and shed some more light on the means of protecting data privacy.

## 2. Methodology

The development of the patent landscape consists of four stages related to (i) the patent selection and trend analysis, (ii) the areas of technology analysis, (iii) assignee country and organization analysis, and (iv) text mining analysis.

## 2.1. Stage 1: patent selection and trend analysis

Our source for the patent search and selection was the PatSeer database, which is an online global patent database covering patent activity in 121 countries and stored in the form of simple patents and patent families. The patent family consists of a set of patent applications assigned to different countries, in order to protect the innovation in a wider geographical area.

To detect the patents related to data anonymization, we have searched patents that have the word "data" and one of the following words in their title: "anonymizing", "anonymization", "anonymized", "anonymizy" and "anonymize". Therefore, the PatSeer database was searched on 13 May 2016, using the search string (TA:(data AND anonym*)), with an option for searching simple patent families. The following keywords associated with data anonymization were used: "anonymizing", "data", "anonymization", "anonymized", "anonymizy" and "anonymize". The British English spelling of the words was also used, e.g.

anonymisation. The possible status of a patent is active, inactive-rejected, refused, suspended or inactive – withdrawn/surrendered. In our analysis, we have focused only on active simple patent families.

## 2.2. Stage 2: patent analysis based on technology areas

With our goal of determining technical content on data anonymization and protected by the patenting process, we conducted an analysis using the International Patent Classification (IPC) system [15]. The IPC separates the entire body of technical knowledge using hierarchical levels in descending order [15]. Figure 1 shows the hierarchical levels of the IPC: section, class, subclass and group. The contents of lower hierarchical levels are subdivisions of the contents of the higher hierarchical levels, and the lower levels are subordinated to the higher hierarchical levels [15].
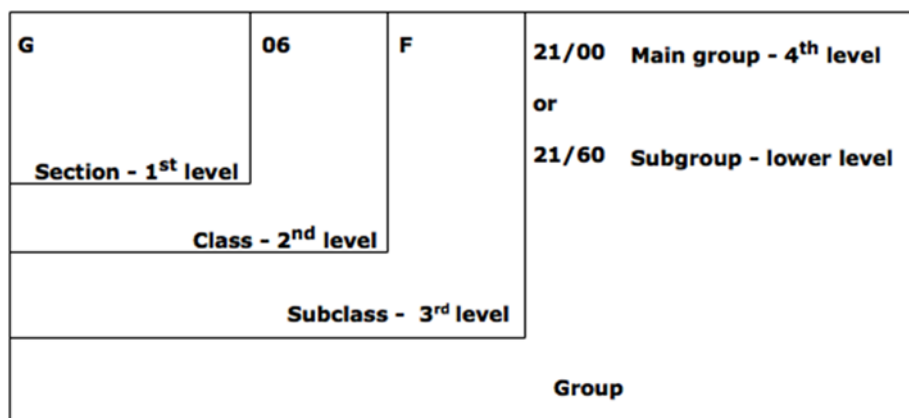


**Figure 1**: *IPC hierarchical levels [15]*

The section is the highest level of the IPC hierarchy. It is considered a very broad indication of technological contents [15]. IPC contains eight sections, is divided into classes, with each class referring to one or more subclasses. Finally, each subclass is broken down into groups [15]. Patents related to data anonymization are most often patented under sections G Physics and H Electricity. Some examples of the sections for data anonymization patents are G06F - Electric digital data processing, and G06Q - Data processing systems or methods. Examples of groups are H04L9/00 - Arrangements for secret or secure communication and G06F21/62 - Security arrangements for protecting computers, components thereof, programs or data against unauthorized activity - Protecting access to data.

In this research, we will analyze the active simple patent families related to data anonymization according to sections, subclasses, and groups. Furthermore, we will

use association rules to conduct an analysis at the IPC Group level in order to determine the heterogeneity of the technical content protected by the patenting process.

## 2.3. Stage 3: patent analysis according to the assignee country and organization

According to the [14], the assignee is the entity that has the property right to the patent. The assignee is not necessary the inventor of new knowledge, as it is more likely that the organization will assign a patent, in which the inventor is employed. In this research, we will extensively analyze organizations and countries, focusing on the longitudinal trend when possible. The aim is to determine the top countries and organizations that assigned patents related to data anonymization.

## 2.4. Stage 4: text mining patent analysis

The text mining approach was utilized to detect the main topics that emerge in patents related to data anonymization. Text mining of simple patent family titles was used to determine what topics emerged most often as the subject of a patenting process related to data anonymization. Reducing the size of the variability of words utilizes different approaches like filtering, lemmatization or stemming [9]. We used the Statistica Text Mining software to utilize the stemming method. Examples of stemming techniques are remove the 'ing' from words, and 's' from the plural of nouns. By using the stemming algorithm, we built the stems, which are a natural group of words with similar or even equal meaning. For example, the stemming algorithm develops a stem "analy" which represents words "analysis" and "analytics".

## 3. Results

## 3.1. Patent search and trend analysis

The PatSeer database was searched on 13 May 2016, using the search string (TA:(data AND anonymiz*)), with an option for searching simple patent families. There were 346 of records for simple family IDs in total. Among these records, 313 simple patent families were active at the time of the search. Therefore, an analysis of the 313 simple patent families related to the data anonymization was conducted to achieve the goal of this research.

Figure 2 represents the patent dynamics for the period between 2001 and 2015. The increasing trend is present from 2001 to 2010 in which less than 15 simple patent families were registered per year. After that period, the number of simple patent families increases, with 82 simple patent families registered in 2015.
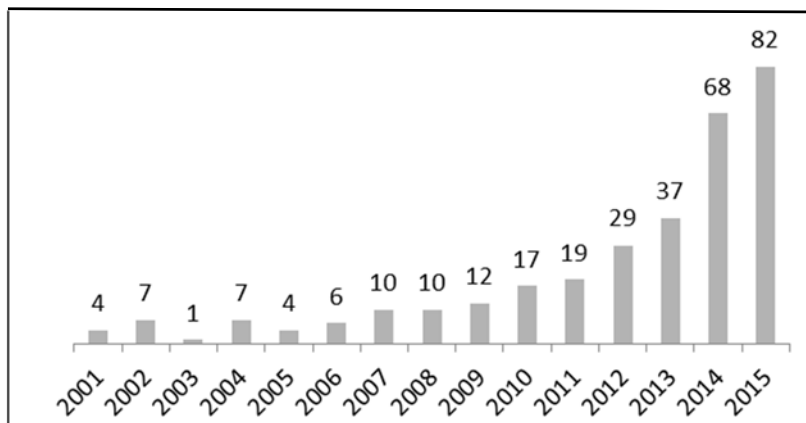
Figure 2: *Number of data anonymization simple patent families (2001-2015) [Authors; 12]*

## 3.2. Patent classification according to the areas of technology

Our search revealed that 313 simple patent families were registered under following five IPC sections: A Human necessities, B Performing Operations; Transporting, C Chemistry, Metallurgy, G Physics and H Electricity. The majority of patents were assigned to the section G Physics with the following sub-classes having the largest number of patents: G06F - Electric digital data processing (294 simple patent families) and G06Q - Data processing systems or methods (113 simple patent families). Following is the section H that covers a significant number of patents assigned to sub-classes H04L - Transmission of digital information (77 simple patent families) and H04W - Wireless communication networks (23 simple patent families). A patent can have multiple ICR codes, which is why the total number of IRC codes (583 codes) is larger than the number of patents examined (313 simple patent families). Table 1 provides detailed information.

| Code | Code description | Simple patent families |
|------|------------------|------------------------|
| | A Human necessities | |
| A61B | Diagnosis; Surgery; Identification | 2 |
| | B Performing operations; transporting | |
| B60Q | Arrangement of signaling or lighting devices | 1 |
| B65G | Transport or storage devices | 1 |
| | C Chemistry; Metallurgy | |
| C12N | Micro-organisms or enzymes; compositions | 1 |
| | G Physics | |
| G01C | Measuring distances, levels or bearings; surveying; navigation; gyroscopic instruments; photogrammetry; videogrammetry | 2 |
| G01D | Measuring not specially adapted for a specific variable and variables not covered by a single another subclass | 2 |
| G01R | Measuring electric and magnetic variables | 3 |
| G06F | Electric digital data processing | 294 |
| G06K | Recognition and presentation of data; record carriers | 11 |
| G06N | Computer systems based on specific computational models | 4 |
| G06Q | Data processing systems or methods | 113 |
| G06T | Image data processing or generation | 9 |
| G07B | Ticket-issuing apparatus; taximeters; apparatus for collecting fares, tolls or entrance fees; franking apparatus | 1 |
| G07C | Time or attendance registers; registering or indicating the working of machines; generating random numbers; voting or lottery apparatus; arrangements, systems or apparatus for checking | 4 |
| G07F | Coin-freed or like apparatus | 4 |
| G08B | Signaling or calling systems; order telegraphs; alarm systems | 2 |
| G08C | Transmission systems for measured values and control signals | 1 |
| G08G | Traffic control systems | 2 |
| G09B | Educational or demonstration appliances | 1 |
| G09C | Ciphering or deciphering apparatus | 11 |
| G10L | Speech analysis or synthesis, recognition, processing, coding or decoding | 3 |

| H Electricity | | |
|---|---|---|
| H02J | Circuit arrangements or systems for supplying or distributing electric power; systems for storing electric energy | 1 |
| H04H | Broadcast communication | 1 |
| H04L | Transmission of digital information | 77 |
| H04M | Telephonic communication | 4 |
| H04N | Pictorial communication | 4 |
| H04Q | Selecting | 3 |
| H04W | Wireless communication networks | 23 |

Table 1: *Number of patents related to data anonymization according to the IPC system - Sub-class level [Authors; 12]*

Analyzing the IPC group revealed the interesting results. The majority of data anonymization simple patent families were assigned to the group G06F17/30 - Digital computing or data processing equipment or methods adapted for specific functions - Information retrieval; Database structures (71 simple patent families). The second most frequent IPC group is G06F21/60 - Security arrangements for protecting computers, components, programs or data against unauthorized activity - Protecting data (59 simple patent families). More detailed information on the number of patents related to data anonymization at the IPC group level is presented in Table 2.

| Code | Code Description | Count |
|---|---|---|
| G06F15/16 | Digital computers in general -Combinations of two or more digital computers each having at least an arithmetic unit, a program unit and a register | 14 |
| G06F17/00 | Digital computing or data processing equipment or methods, specially adapted for specific functions | 11 |
| G06F17/30 | Digital computing or data processing equipment or methods adapted for specific functions - Information retrieval; Database structures | 71 |
| G06F19/00 | Digital computing or data processing equipment or methods, specially adapted for specific applications | 15 |
| G06F21/00 | Security arrangements for protecting computers, components thereof, programs or data against unauthorized activity | 16 |
| G06F21/24 | Information storage device, information storage program, verification device and information storage method | 7 |

| | | |
|---|---|---|
| G06F21/60 | Security arrangements for protecting computers, components, programs or data against unauthorized activity - Protecting data | 59 |
| G06F21/62 | Security arrangements for protecting computers, components thereof, programs or data against unauthorized activity - Protecting access to data | 44 |
| G06Q10/00 & 06 | Administration; Management & Resources, workflows, human or project management; Enterprise planning; Organizational models | 17 |
| G06Q30/00 | Commerce | 14 |
| G06Q30/02 | Commerce -  Marketing | 13 |
| G06Q40/00 | Finance; Insurance; Tax strategies; Processing of corporate or income taxes | 10 |
| G06Q50/00 | Systems or methods adapted for a specific business sector | 11 |
| G06Q50/22 | Systems or methods adapted for a specific business sector - Healthcare | 11 |
| G06Q50/24 | Systems or methods adapted - Patient record management | 8 |
| G09C | Ciphering or deciphering apparatus for cryptographic or other purposes | |
| G09C1/00 | Apparatus or methods whereby a given sequence of signs is transformed into an unintelligible sequence of signs by transposing the signs or groups of signs or by replacing them by others according to a predetermined system | 11 |
| H04L29/06 | Arrangements, apparatus, circuits or systems, not covered by a single one of groups H04L 1/00-H04L 27/00 - Characterized by a protocol | 25 |
| H04L29/08 | Arrangements, apparatus, circuits or systems, not covered by a single one of groups H04L 1/00-H04L 27/00  - Transmission control procedure | 9 |
| H04L9/00 | Arrangements for secret or secure communication | 7 |
| H04L9/32 | Arrangements for secret or secure communication - Including means for verifying the identity or authority of a user of the system | 10 |
| H04W12/02 | Security arrangements; Authentication; Protecting privacy or anonymity | 7 |

Table 2: *Number of patents related to data anonymization according to the IPC system - Group level [Authors; 12]*

In order to detect the degree of heterogeneity of technical content related to data anonymization and protected by the patenting process, we used association rules analysis [9]. Most of the patents were assigned to more than one IPC group, and 583 groups were identified for the 313 simple patent families. This indicates that one simple patent family is registered to approximately 2 IPC groups in average. Therefore, 28 rules are generated, under a minimal support and confidence at 1% level, which indicates that finding the dependencies between different IPC group level codes is difficult. The results are presented Figure 3.
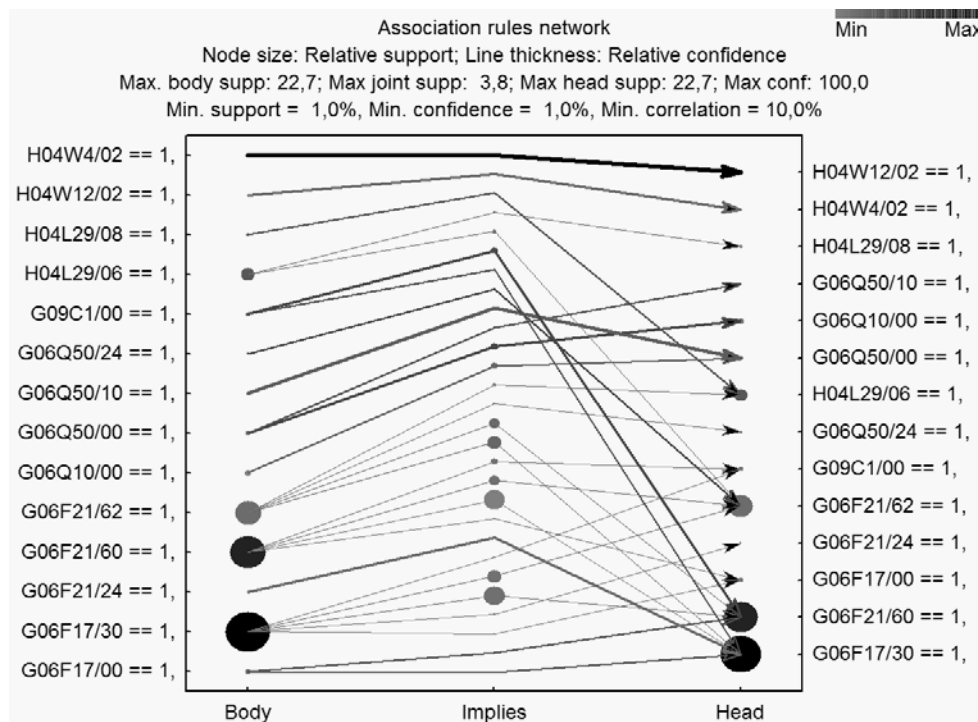


**Figure 3**: *Association rules network at IPC Group level patents [Authors; 12; Statistica Text Miner]*

According to the set limitations and generated rules, we can conclude that heterogeneity is not characterize of protected technical content related to data anonymization. The results reveal the following IPC groups were most often registered together: G06F21/60 - Security arrangements for protecting computers, components, programs or data against unauthorized activity - Protecting data, G06F21/62 - Security arrangements for protecting computers, components thereof, programs or data against unauthorized activity - Protecting access to data via a platform, G06F17/30 - Digital computing or data processing equipment or methods adapted for specific functions - Information retrieval; Database stru-

ctures, and G06F17/00 - Digital computing or data processing equipment or methods, specially adapted for specific functions.

## 3.3. Patent assignee organization and country analysis

Our search revealed that the patenting activities are spread across different countries, but the USA and Japan have assigned the majority of patents related to data anonymization. In some cases, more than one organization from two or more countries were the assignees. Figure 4 outlines the patent dynamics according to countries for the period between 2001 and 2015. The USA is the leading country since its organizations began publishing patents on data anonymization in 2000. Other countries followed later. European countries that have assigned more than five patents in given period are Germany (17 simple patent families), France (7 simple patent families) and other member countries of the European Patent Organization (15 simple patent families). Table 3 lists the number of patents related to data anonymization of the countries for the period between 2001 and 2015.
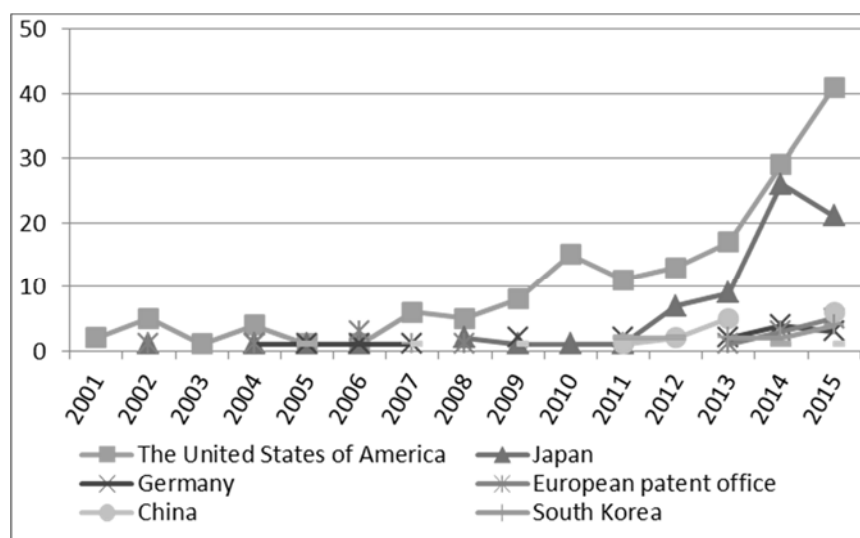


Figure 4: *Number of data anonymization simple patent families per country (2001-2015); countries with more than five simple patent families [Authors; 12]*

| Country | Country code | Number of patents | % |
|---|---|---|---|
| The United States of America | US | 159 | 50.80% |
| Japan | JP | 72 | 23.00% |
| Germany | DE | 17 | 5.43% |
| European patent office | EP | 15 | 4.79% |
| China | CN | 14 | 4.47% |
| South Korea | KR | 9 | 2.88% |
| France | FR | 7 | 2.24% |
| Great Britain | GB | 6 | 1.92% |
| India | IN | 4 | 1.28% |
| Finland | FI | 3 | 0.96% |
| Spain | ES | 2 | 0.64% |
| Canada | CA | 2 | 0.64% |
| Russia | RU | 1 | 0.32% |
| Israel | IL | 1 | 0.32% |
| Missing data | - | 1 | 0.32% |
| Total | - | 313 | 100.00% |

**Table 3**: *Number of simple patent families related to data anonymization according to country [Authors; 12]*

Table 4 shows the number of simple patent families related to data anonymization according to assignee organization, and country. The assignee of a patent is an organization that refers to a company, an academic institution and individual persons in some of the cases. The organization with the largest number of simple patent families related to the data anonymization in the observed period is NEC registered in Japan (27 simple patent families or 8.62%), followed by IBM registered in the United States of America (19 simple patent families or 6.07%). Other organizations that registered a larger number of simple patent families are also multinational organizations, such as Hitachi, Microsoft, Nifty, Siemens, Fujitsu, MasterCard, and Amazon.

| Assignee organization | Country | Count | % |
|---|---|---|---|
| NEC | Japan | 27 | 8.63% |
| IBM | USA | 19 | 6.07% |
| Hitachi | Japan | 9 | 2.88% |
| Microsoft | USA | 9 | 2.88% |
| Nifty | Japan | 9 | 2.88% |
| Siemens Ag | Germany | 7 | 2.24% |
| Fujitsu | Japan | 6 | 1.92% |

| Alcatel Lucent | France | 6 | 1.92% |
|---|---|---|---|
| Att Intellectual Property | USA | 5 | 1.60% |
| Mastercard International | USA | 5 | 1.60% |
| Google | USA | 5 | 1.60% |
| Toshiba | Japan | 4 | 1.28% |
| Accenture Global Services | Ireland | 4 | 1.28% |
| Nippon Telegraph-Telephone | Japan | 4 | 1.28% |
| Other | - | 194 | 61.98% |
| Total | - | 316 | 100% |

Table 4: *Number of simple patent families related to data anonymization according to assignee organization and country [Authors; 12]*

## 3.4. Text mining utilization for topics identification

We used text mining analysis to extract the most common words that occurred in the patent titles of simple families, using a stemming approach and available as a feature of the Statistica Text Miner program. The stem or phrases are generated as the output of the stemming algorithm. Table 5 shows the most frequently used stems or phrases in the titles of patents related to data anonymization. "Anonym*" is the most often phrase as expected. Also, the words "method", "data" and "system" appeared in more than 100 cases. Other terms often present are "inform", "devic", "apparatus", and "process".

In order to provide a more intuitive insight into the themes that occur in the titles of the simple patent families related to data anonymization, a tag cloud analysis was conducted [6]. The tag cloud has become a common way of visualizing most occurring themes since it visualizes the most common words in an analyzed text, relating the size of the word and its relative frequency. Therefore, the words that occur more often are larger. The Wordle program was used to generate a tag cloud of the stems that occurred most often in the titles of the simple patent families related to the data anonymization. To increase the transparency of the cloud, we applied tag cloud algorithm to the stems that occurred more than five times in the titles of simple patent families. We excluded the stems "data" and "anonym" as they occurred in every title and given that these words were the criteria for selecting the simple patent family in the analysis. Moreover, two stems that also occurred often were omitted from the analysis: "system" and "method".

| Stem / Phrase | Number of occurrences in the title | Number of simple patent families | Examples |
|---|---|---|---|
| anonym | 192 | 165 | anonymity |
| method | 192 | 183 | method |
| data | 152 | 139 | data |
| system | 123 | 122 | system |
| inform | 78 | 59 | information |
| devic | 56 | 53 | device |
| process | 45 | 33 | process |
| apparatus | 42 | 42 | apparatus |
| program | 33 | 33 | program |
| network | 23 | 21 | network |
| use | 23 | 23 | use |
| comput | 22 | 20 | computer |
| person | 21 | 18 | person |
| privaci | 18 | 18 | privacy |
| manag | 17 | 12 | managing |
| user | 17 | 16 | user |
| medic | 16 | 14 | medical |
| protect | 16 | 14 | protected |
| provid | 15 | 14 | provider |
| behavior | 13 | 11 | behavior |
| record | 13 | 13 | record |
| servic | 12 | 11 | service |
| set | 12 | 10 | set |
| product | 10 | 10 | product |

Table 5: *Most commonly used words in titles of patents related to data anonymization; => ten patents [Authors; 12; Statistica Text Miner]*

Figure 5 indicates that the stems "method" and "system" occurred most often in the titles of simple patent families related to the data anonymization. The following groups of topics were also identified: (i) topics related to physical equipment such as "devic", "comput" or "apparatus"; (ii) topics related to software such as "program", "process", and "analy" or "manag"; (iii) topics related to the goal of the patent, such as "protect", "ident", "encrypt" or "privac"; and (iv) some specific topics related to the areas of the implementation, such as "commun", "medic", or "service".
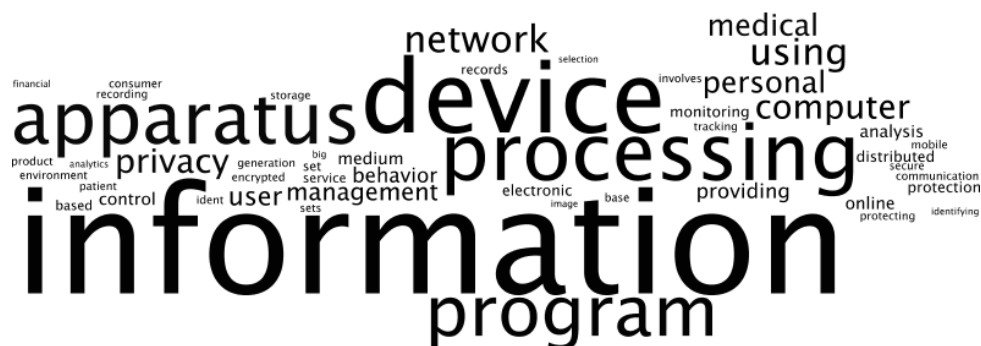
Figure 5: *Tag cloud of the most often used words in patent titles related to data anonymization; > 5 simple patent families [Authors; 12; Wordle.org]*

## 4. Conclusion

The paper presents an examination of simple patent families relating to data anonymization, and based on the data gathered from PatSeer. We analyzed 313 active simple patent families related to data anonymization assigned from 2001 to 2015. The analysis was conducted in four stages: (i) detecting the trend in data anonymization patenting, (ii) patent classification based on technology areas, (iii) assignee organization and country analysis, and (iv) the use of text mining for identifying of topics. The analysis has provided answers to the research questions, and provided insights into the data anonymization patent landscape.

The first research goal was to detect any trends in data anonymization patenting. The number of single patent families has been growing intensely since 2010, and especially as of 2014, thus indicating a positive trend in the area of patenting data anonymization solutions. This is due to an increasing awareness of the necessity of data privacy protection, and also new challenges (e.g. big data).

The second research goal was to detect protected technical content related to data anonymization and classified using the IPC system (at the sub-class and group level). The majority of simple patent families relating to data anonymization were assigned to the section G Physics of IPC system. G section sub-classes with most patents are G06F - Electric digital data processing and G06Q - Data processing systems or methods. Within this sub-class, the majority were assigned to the

group G06F17/30 - Digital computing or data processing equipment or methods adapted for information retrieval and database structures. Therefore, the protection of data privacy in databases and information retrieval has been the biggest focus of inventors, and is due to the omnipresent digitization of information. Association rules analysis has revealed that patents with more than one IPC group were homogenous, since all of the co-occurring IPC groups were from the class G06F- Electric digital data processing.

The third research goal was to identify which organizations from countries patented their data anonymization innovations. According to the patent analysis, data anonymization technology is spread across different countries, but the majority of simple patent families relating to data anonymization have been assigned by the USA and Japan organizations. In the observed period, the NEC company, registered in Japan, has assigned the greatest number of patents, followed by IBM registered in the USA. Numerous multinational corporations, such as Google, Microsoft, Amazon and MasterCard have also registered a substantial number of data anonymization patents.

The fourth research goal was to identify the topics emerging most often as the subject of patenting process related to data anonymization. The most often used word in the titles of patents related to data anonymization was "anonym*", followed by "method", "data" and "system". Several additional groups that indicate the most frequent data anonymization topics were also detected: physical equipment, software, protection, identification, encryption or privacy, and specific topics such as community, medical, or service.

The limitations of this work result from the fact that we have focused only on simple patent families possessing the word "data" and one of the following words: "anonymizing", "anonymization", "anonymized", "anonymizy" and "anonymize". Hence, patents with these words in the abstract but not in the title were omitted from the analysis.

## References

[1] Abbas, A., Zhang, L., and Khan, S.U. (2014). A literature review on the state-of-the-art in patent analysis. World Patent Information. 37, 3-13

[2] Brügmann, S. et al. (2015). Towards content-oriented patent document processing: intelligent patent analysis and summarization. World Patent Information. 40, 30–42.

[3] Cecere, G., Corrocher, N., Gossart, C., and Ozman, M. (2014). Technological pervasiveness and variety of innovators in Green ICT: a patent-based analysis. Research Policy, 43(10), 1827-1839.

[4] Choi, J., and Hwang, Y.-S. (2014). Patent keyword network analysis for improving technology development efficiency. Technological Forecasting and Social Change. 83, 170–182.

[5] Cormode, G., Srivastava, D. (2010). Anonymized data: generation, models, usage. In: 26th IEEE International Conference on Data Engineering, pp. 1211–1212. IEEE, Long Beach

[6] De Spindler, A., Leone, S., Nebeling, M., Geel, Matthias, and Norrie, M.C. (2011). Using synchronised tag clouds for browsing data collections. In: Mouratidis, H. and Rolland, C. (Eds.). Advanced Information Systems Engineering. LNCS, vol. 6741 (pp. 214-228). Heidelberg: Springer

[7] Grant, E., Van den Hof, M., and Gold, E. R. (2014). Patent landscape analysis: a methodology in need of harmonized standards of disclosure. World Patent Information. 39, 3-10.

[8] Han, E.J., and Sohn, S.Y. (2016). Technological convergence in standards for information and communication technologies. Technological Forecasting and Social Change. 106, 1–10.

[9] Kim, J., and Lee, S. (2015). Patent databases for innovation studies: A comparative analysis of USPTO, EPO, JPO and KIPO. Technological Forecasting and Social Change. 92, 332–345.

[10] Noh, H., Jo, Y. and Lee, S. (2015). Keyword selection and processing strategy for applying text mining to patent analysis. Expert Systems with Applications. 42(9),4348–4360.

[11] Patent Lens. Patent Laws Around the World. Available at: http://www.bios.net/daisy/patentlens/ip/around-the-world.html [Accessed 21/4/2016].

[12] PatSeer. (2016). Available at: http://patseer.com/ [Accessed 6/5/2016].

[13] Sheau-Pyng, J., Ming-Fong, L., and Chin-Yuan, F. (2012). Using patent analysis to analyze the technological developments of virtualization. Procedia-Social and Behavioral Sciences. 57, 146-154.

[14] Sinha, M., and Pandurangi, A. (2015). Guide To Practical Patent Searching And How To Use Patseer For Patent Search And Analysis. Gridlogics Technologies, Pune.

[15]World Intellectual Property Organization. (2015). Guide to the IPC. WIPO. Available at: http://www.wipo.int/export/sites/www/classifications/ipc/en/guide/guide_ipc.pdf [Accessed 21/4/2016].