A free open source patent corpus generator

David Reymond, Roberto Faga, Andreas Molt, Celso Arruda Vanderlei, Luc Quoniam

A free open source patent corpus generator

David REYMOND[1] [0000-0003-2071-6594] and Roberto FAGA[2] and Andreas MOTL[3] and Celso VANDERLEI[2][0000-0003-4208-3353] and Luc QUONIAM[1] [0000-0002-6333-6594]

[1] Université de Toulon, Institut Méditerranéen des Sciences de l'Information et de la Communication (IMSIC - EA 4262, Toulon, France)
[2] Universidade Nove de Julio, vlab4u, Sau Paulo, Brazil
[3] IP Tools -Open source patent information software, Germany

Abstract. Patents are an unused informational source in research and education. We stated that patent documents are an information source opening a wide variety of usage (monitoring, strategic positioning, technical document for innovation and state of the art) for a wide variety of users (researchers, Small and medium Enterprises (SME) and start-ups). Hence, patents are a key resource in education but difficult to read. The European Patent Database (EspaceNet) offers 100 million of demands in free access mode to their world-wide range database. We consider it as a technological encyclopedia: most of its content is free of rights. Patent2Net (P2N) is a free open-source patent analysis tool that offers the potential to extract patents from the worldwide database and to develop new data collections in very wide variety of domains (for instance: education, banana peel, rare earth, drones, or the list of last year demands (more than 1 Million documents) and so on). As an example, we propose to explore a huge set of patents using several data mining tools and visualization techniques (dynamic networks, textometry application, classification). We use several key information representations to explore the set: cartographies of a technology using mind maps, the network of actors or applicants, k-means clusters and many other features directly in concern with innovation, knowledge, and education.

Keywords: patent document, technical encyclopedia, corpus generator.

1 Introduction

Beside their capacity in memorizing real states in technical domains, it is well known that patents are, in general, a misused information source [1]. Besides this, there is also a need to use patents as a research resource [2, 3]. The authors, among others [4], underlined that patent documents can be useful to determine the technological state-of-the-art in a domain and as an information source for innovation[5]. Valencia-Zuluaga et al. shows further that patents documents can also be used as library services resources. Compared to the publicly free online databases, we state that the value added by professional databases is marginal, for educational purposes. But, all the authors

have denoted the need for affordable tools to fill the gap in academic research and education as usual software and online services in this domain may be too expensive. See [5] and [6] for a comparison of several patent analysis tools.

Hereafter we present the Patent2Net (P2N) solution which offers an up-to-date and state-of-the-art tool suite for patent analysis; it is open source and we claim here that it can be used for specific treatment processes as corpuses creation. Furthermore, it is ready to use in educational or research programs. In the following, we will detail several aspects of patents documents to state that patent databases, traditionally used for industrial property, could be seen a technical encyclopedia. The second part consists in presenting several features of the tools showing its capability in corpuses generator.

2 A multidimensional information source

2.1 A high quality document.

A patent is described in a highly structured document, covering several states (see Table 1) recorded by database: filing stage, published, granted and expired are the most common steps. Each step before the granting one, impose the document to be checked by experts which role is also to complete the content with rich metadata set. Hence, each phase of the document may disclose different kind of information. Law requires that each patent must disclose sufficient technical information to allow skilled practitioners of the art to recreate the invention [7]. Hence, the in-depth expertise (from 3 to 8 years long) may transform the document to a patent application, and infers many corrections to the content (adding citations for instance) or expending metadata (such as classification properties). After this step, the invention may be protected by the patent document, covering some claims, describing it for anyone: the patent comes to be granted, usually for 20 years. During this period, competitors can discuss validity of the patent. If the validity period or an opposition is stated then, the invention falls into the public domain. Most of the European Patent Office documents are in this case.

2.2 Specific life cycle.

Figure 1 shows the patent idea life cycle [8, p. 15]. At the first stage of its construction: the patent document must describe the invention clearly within an abstract, a title and a primary classification. Then, a background section must describe the problem to be solved and the current state of the art. Finally, claims are written in legal terminology and this style is referred to *patentese* [8, p. 9].

Table 1: Life-cycle of patents documents adapted from [9]

Phase	**When**	**Information disclosed**
Patent application	Filing date	Some metadata (min: title, IPC, abstract, applicant)

Patent application – published	18 months after the filing date	Full text and metadata
Granted patent	From 2 up to 5 years or more after the filing date	Amended full text and revised metadata
Expired or ceased patent (public domain)	20 years after the filing date (or before in case of unpaid fees)	Amended full text and revised metadata

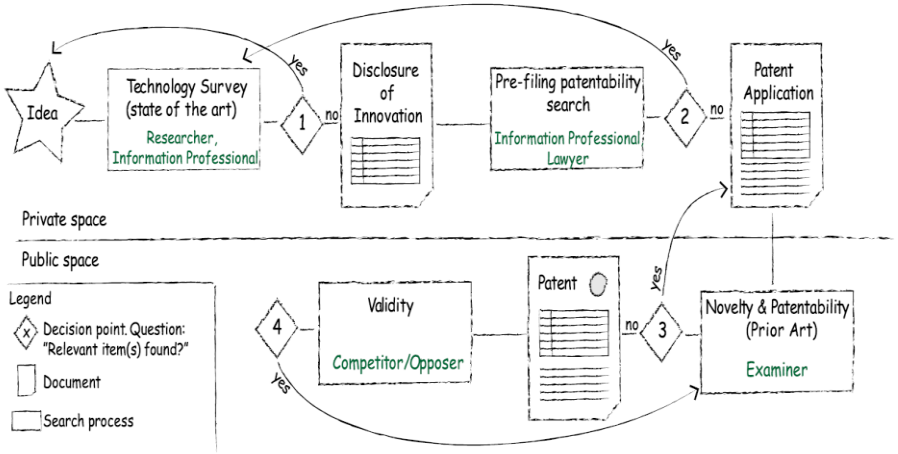

Figure 1: life-cycle Patent idea, from [10, p. 9]

Patent information is also typically more detailed and, exhaustive than scientific papers [11]. The language of patents is unique and contains highly specialized or technical words not found in everyday language [12]. Patent documents are often written in "patent jargon," they are more difficult to utilize and understand [13]. Most patents document are also completed by images and diagrams to explicit their contents.

2.3 A unique classification schemes.

The World Intellectual Property Organization maintains a standard hierarchical taxonomy for patent classification, named *International Patent Classification* (IPC). Updates are provided every three months for the deepest levels and every three years for the core classification. The IPC is adopted by more than 100 nations worldwide [9], [14]. The IPC classification constitutes a rich controlled taxonomy to describe inventions, it is used by experts as an international and independent language. Searching patents using the classification scheme allows the user to make abstraction of primary language of the patent. Hence, classification system is considered as a good alternative to full text research, not proposed for patent information retrieval [15] .

To refine more again patent search, EPO and USPTO started recently to fusion their own classification in a common classification system compatible with IPC structure, named the Cooperative Patent Classification.

As everyone can write a patent, all this checking process from expert evaluation to expert classification, makes for each step, if granted, the convergence of the patent document to a new performed document version. That process control, similar in some sense to traditional evaluating process for academic stuff, brings the patent document to have a non-negligible quality and bibliographic consistent metadata description.

2.4 A patent database freely accessible.

Among the many patent databases freely accessible on the web (see World Intellectual Property Organization, 2009, p 5-6), Esp@cenet is a free patent search service offered by the *European Patent Office* (EPO). The "worldwide" database is currently centralizing over 90 databases and over 50 patent authorities. Esp@ceNet lacks tools for analyzing data but, in 2006, the European Patent Office, released an *Application Programming Interface* (API) to their databases called the Open Patent Service [16]. The API allows specific crawler to get access to all the content provided by the database, freely under conditions of fair-robot and fair loading of server's compliance. Using the API, one can download 2.5 GB per week.

Esp@ceNet "worldwide" database is the world's largest free collection of technical information. "Espacenet offers free access to more than 100 million patent documents worldwide, containing information about inventions and technical developments from 1836 to today". From there, one can consider this database as a huge technical encyclopedia, centralizing a worldwide feed proceed, controlled by experts and classified by almost one classification taxonomy. These points make us view a patent database as technical encyclopedia.

3 Patent2Net

Patent2Net (P2N) was born in 2014, in the promotion of the master *Intelligence Économique et Territoriale* of the University of Toulon. The main objective was to give to the students some capabilities in IP practice and some skills in complex data treatment in their Competitive Intelligence course. Its original denomination came from visualizing bibliographic patents' data with networks, abbreviated now into P2N.

3.1 General architecture.

Written in python language, the script suite is under Cecil-B License (GNU GPL under French laws), and is distributed on an "AS IS" basis, without warranties or con-

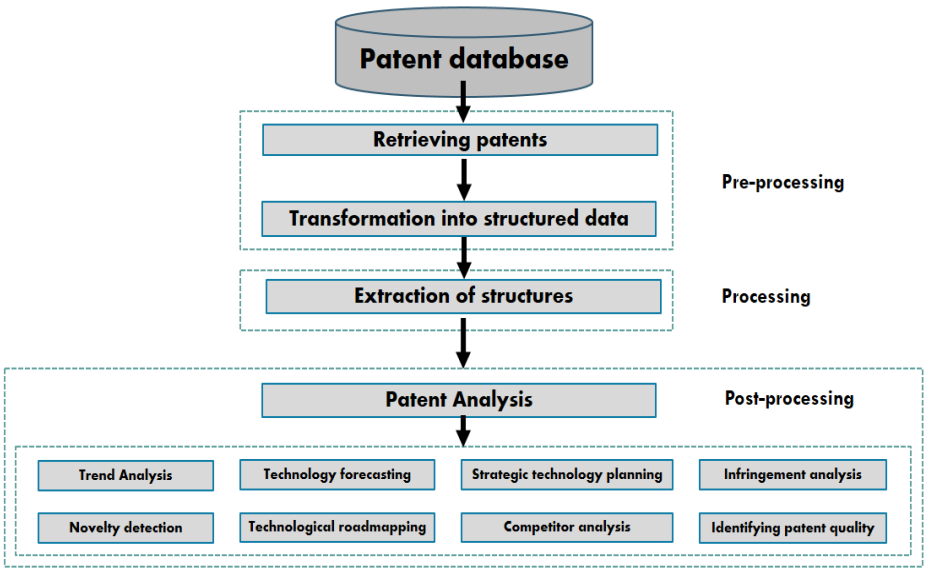

Figure 2: Generic patent analysis workflow from [14]

ditions of any kind, either expressed or implied. Anyone can get the code or join the team on git-hub [17].

P2N suite structure follows the generic patent analysis [18] delivered in three main steps from pre to post processing (see Figure 2): gather a patent universe corresponding to a request, filter and segment data according to specific bibliographical fields (pre-processing) and deliver interoperable content for patent analysis (processing). Specific software and tools are then convened to provide high-level analysis features (post-processing). We develop the three steps here after, with needed specific technical precisions.

3.2 The gatherer.

The pre-processing script is the gatherer. This script is fully compatible with the OPS API [19] as is offers the capability to retrieve all the bibliographic data attached to a request formulated in CQL (*Contextual Query Language*). Pre-process is built on a client OPS library [20]. It handles and API credential couple (log, pass) that allows the users to gather in free mode (up to 2.5 GB a week). High-level consumer data can choose the paying mode in OPS registrar (that augments the amount and rate of allowed data transfers). The main feature in this step is, once the credential configured, to re-

cursively interrogate the worldwide database in order to download the entire set of patents data according to one request. The result set contains the same or even greatest patent data set as is would be downloaded from the *Smart Search web frontal* feature. Smart search and advanced search features are presented in [21]: P2N is fully compatible with advanced search field identifiers (search in *titles, titles or abstract, author, applicant, publication date, priority number or demand, CPC or IPC)* and the smart search operators (*Boolean operators, proximity, comparison, truncation…*). In addition, optional switches, in the configuration file, allows the gatherer to retrieve complementary bibliographic data, patents contents (descriptions and claims), and patents families as described in OPS.

A data patent content set retrieved through a CQL request is called hereafter *the patent universe*.

3.3 Processing step.

Processing step will threaten the bibliographic data, the textual content and the images of Patent Universe (PU) to facilitate data presentation and content managing tools. This step is separated in two main features: the processors that provide interoperability with external analysis tools and the FormateExport* sub-suite, created to present data for general analysis compatible with the Firefox navigator in HTML format using several libraries (DataTable, the Pivot Table, Zotero, D3.js, etc.)

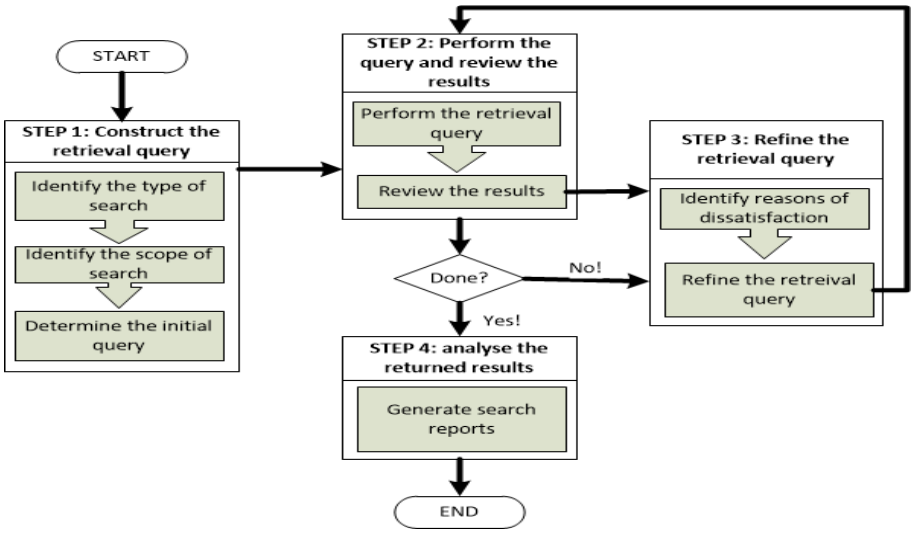

Figure 3: Patent search procedure from (Zhang, Li, & Li, 2015)

Aside these generic tools to study the patent universe; P2N offers interoperability to many in-depth analysis tools through post-processing data using bibliographical fields to tag contents.

3.4 External interoperability for in-depth analysis.

P2N provides four main connectors for in-depth analysis of the patent universe. The first one is a complete sub-suite for generating networks of interest from bibliographic fields. The second allow a micro lexical analysis, using IRAMuTeQ textometry suite. The third one enters a macro level representation formatting abstract data to the Carrot2 classification tool. In addition, the last one, at a Meta level, proposes a mind-map representation of IPC's in a human readable description.

Networks.

GEXF (Graph Exchange XML Format) is a language for describing complex networks [26] enabling lot of characteristics (partitioning, coloring, dynamic graphs). We choose it to stand for the interoperability chain from bibliographical data to networks.

P2N offers the distinction of three main kids of directed networks:

- Simple and mixed networks that handles cooperation's networks (inventors, applicants or IPC's networks) or inter category networks. Within these nets, all nodes are tagged by their category (different color in visualization), an url to address further information. Nets are also qualified in time for dynamically follow evolution on the patent universe.
- References networks addresses networks of citations. These nets expose two sub kinds of network citations: *UniverseReferences.gexf* (which presents the cited documents by the patent, making the distinction of nodes belonging to Patents database or external references such as academic publications) and *UniverseCitations.gexf* that represents the citations of the document itself by other patents, but includes previous net also.
- Finally, Families networks handles a **hierarchical representation** of family in OPS sense. At the time of writing this last category is not yet transposed in the last version of P2N.

Each network processed (three citations networks, three mono fields nets and four cross-fields nets) offers through Gephi [27] software, the capability the user to analyses metadata interconnections. Among classical description on any node by category, url, several characteristic network properties (degree, centrality and time) permits works on relations between nodes and better understanding on the PU.

Figure 4: Gexf interface showing part of PU applicant-inventor networks through Firefox. Left side bar provides node information and link to Wipo or OPS website).

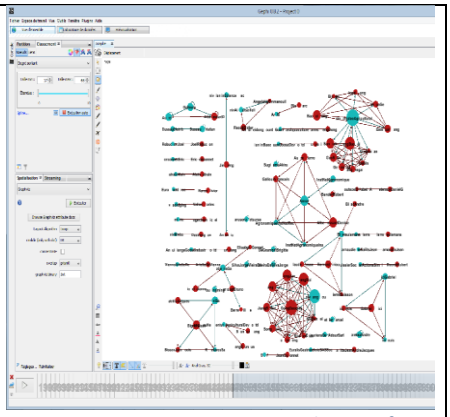

Figure 5: Same net using Gephi interface. Nodes are proportional to their out degree, colored by type (Applicant in blue, Inventors in red). Down side toolbar provide time clustering feature.

Asides the Gephi export, P2N uses Graphiz [28] and the Gexf-js [29] Firefox JavaScript interface to network to provide a downgraded but useful version to display networks on a web server. Figure 4 and Figure 5 shows an example of these tools applied to a patent universe.

3.5 Textometry.

Applied to abstracts, claims or descriptions, textometry tools provides useful information on contents of the PU, resuming it and showing terminology relations and correlations. Among many tools, IRaMuTeQ [30] stands for "*Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires*", see [31]), is selected to supply many multidimensional analysis features.

3.6 Images

An ongoing feature for Patent2Net is allowance of images extraction from patents for a given search. By downloading hundreds or thousands of images, P2N displays these images in a gallery and allows easy navigation over all patents in a single panel of thumbnails, making it easier to find patterns by simple visual analysis. In future development we expect to provide clustering techniques or even images recognition to extract relevant data.

Mind maps.

Dealing with patent content analysis as library function, Chung-Huei & Chan-Yi, (2015) denoted on classification system that an analyst would require a better tool to

take the hierarchical relationship among classification symbols into consideration. “If this kind of tool is available, we speculate that some specific technical areas may reveal a high consistency rate or similarity measure even for PCA”. P2N offers this kind of feature using mind map representation of PU. For each patent bibliographic IPC classification field, P2N extract all IPC’s number in hierarchical view described in the format compatible with the free mind mapping and knowledge management software: Freeplane [39]. Classification abbreviations of the scheme are extended in natural language to facilitate interpretation. Further manual treatment is allowed to expose easily the PU covering in order to track opportunities or missing coverage in the PU.

Mind maps can be defined as visual and non-linear representations of ideas and their relationships, composing a network of concepts on a given theme [40]. The technique was developed by Tony Buzan and presented initially in the book "Use Your Head" of 1974 as a tool of association of ideas and suggested to students as a more efficient way of taking notes [41]. Buzan's main argument is that mind maps resemble how the brain learns, that is, by non-linear associations, and thus facilitate the assimilation of knowledge [42]. Although it was created with the main objective of facilitating the association of ideas, the technique was also popularized as an aid to memorization and learning.

Patent research is considered an important tool in the development of creative thinking and the development of innovations [43]. In this process, the interactivity provided by mental maps becomes an important aid, since it allows the user to take notes, complement or comment on the information observed in the universe of selected patent records.

The use of IPC to organize the selected patent records directs the user by innovative thinking, once he or she can visualize the several applications alternatives and think about each one, going deep on the categories that suit better and despise the one that seems useless [43].

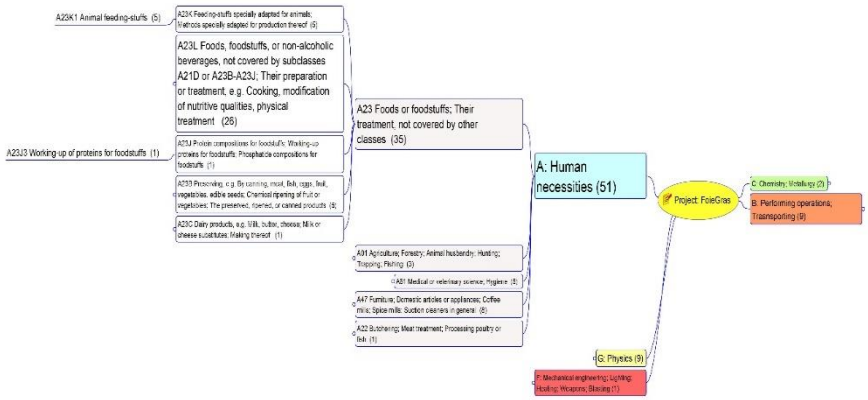

Figure 6: Freeplane native mind map processed of the previous universe. IPC's codes are distinguished by colors and explained in human readable language.

4 CONCLUSION

The development of P2N, the free and open source tool for research and education opens wide applications in curricula. The tools allow to create wide or short corpuses and categorize the technologies developed opening a way to discuss theoretical and practical development, state strategy of research, and may help in policy decision of research. It offers the possibility to take in account the huge quantity of potential innovations to catalyze the creativity. Asides the variety of uses of the patent information across a wide variety of usages in real life, P2N stands for a corpus creation API. It offers the capability to take profit of highly structured multimedia (images and text) data provided by the database to build up wide range variety of corpuses (domain specific, cross domain or historical transversal) in several formats: texts, images or networks. Thanks to the well-structured patent documents, corpuses focusing by targeting a very large variety of contents (focusing wide range domains: for instance, high technology, huge organizations, a country, a region, etc.) to study technological dynamics. Dealing with such corpuses allow to compete various research state of the art challenges: textometry, visualization, automatic classifications and so on.

5 REFERENCES

[1] R. A. Stembridge, « Education and certification of patent information professionals in europe », in *Acs Symposium Series*, 2010, vol. 1055, p. 87- 93.

[2] D. Reymond et L. Quoniam, « A new patent processing suite for academic and research purposes », *World Pat. Inf.*, vol. 47, p. 40- 50, déc. 2016.

[3] D. Reymond et L. Quoniam, « Patent documents in STEM and PhD education. », in *Proceedings of the IEEE EDUCON2018 International Conference:*, Santa Cruz de Tenerife, Canarias islands, Spain, 2018.

[4] A. Magid, « The road to interactive patent searching at an American University in the UAE », 2016, p. 438- 442.

[5] T. Valencia-Zuluaga, S. R. Rivera-Rodriguez, et N. Sánchez-Ortíz, « Potencial en el uso de la consulta de patentes para determinar el estado de la tecnica. Analisis en microredes con energías renovables », *Ing. Investig. Desarro.*, vol. 17, nº 2, p. 16, juin 2017.

[6] J. Feng et N. Zhao, « A New Role of Chinese Academic Librarians—The Development of Embedded Patent Information Services at Nanjing Technology University Library, China », vol. 41, nº 3, p. 292 – 300, 2015.

[7] J. Bessen, « Patents and the diffusion of technical information », *Econ. Lett.*, vol. 86, nº 1, p. 121- 128, janv. 2005.

[8] M. Lupu et A. Hanbury, « Patent retrieval », *Found. Trends Inf. Retr.*, vol. 7, nº 1, p. 1- 97, 2013.

[9] D. Bonino, A. Ciaramella, et F. Corno, « Review of the state-of-the-art in patent information and forthcoming evolutions in intelligent patent informatics », *World Pat. Inf.*, vol. 32, nº 1, p. 30- 38, mars 2010.
[10] M. Lupu et A. Hanbury, « Patent Retrieval. », *Found. Trends Inf. Retr.*, vol. 7, nº 1, p. 1–97, 2013.
[11] D. Bonino, A. Ciaramella, et F. Corno, « Review of the state-of-the-art in patent information and forthcoming evolutions in intelligent patent informatics », *World Pat. Inf.*, vol. 32, nº 1, p. 30-38, mars 2010.
[12] H. Joho, L. A. Azzopardi, et W. Vanderbauwhede, « A Survey of Patent Users: An Analysis of Tasks, Behavior, Search Functionality and System Requirements », in *Proceedings of the Third Symposium on Information Interaction in Context*, New York, NY, USA, 2010, p. 13–24.
[13] P. P. Paranjpe, « Patent Information and Search », *DESIDOC J. Libr. Inf. Technol.*, vol. 32, nº 3, 2012.
[14] H. Wongel, « The reform of the ipc—consequences for the users », *World Pat. Inf.*, vol. 27, nº 3, p. 227-231, 2005.
[15] S. R. Adams, *Information sources in patents*. De Gruyter Saur, 2012.
[16] P. Kallas, « Open patent services », *World Pat. Inf.*, vol. 28, nº 4, p. 296-304, 2006.
[17] David REYMOND, *Patent2Net (P2N)*. Toulon: Université de Toulon, 2015.
[18] A. Abbas, L. Zhang, et S. U. Khan, « A literature review on the state-of-the-art in patent analysis », *World Pat. Inf.*, vol. 37, p. 3-13, 2014.
[19] P. Kallas, « Open patent services », *World Pat. Inf.*, vol. 28, nº 4, p. 296-304, 2006.
[20] George Song, *python-epo-ops-client*. 2014.
[21] « Espacenet. Free access to 90 million patent documents worldwide », European Patent Office, Vienna, Austria, 2015.
[22] *datatable*. SpryMedia Ltd, 2007.
[23] L. Zhang, L. Li, et T. Li, « Patent Mining: A Survey », *SIGKDD Explor Newsl*, vol. 16, nº 2, p. 1–19, mai 2015.
[24] Nicolas Kruchten, *PivotTable.js*. Datacratic, 2013.
[25] *Zotero*. Fairfax, Virginia 22030: Roy Rosenzweig Center for History and New Media, 2015.
[26] Heymann, Sebastien *et al.*, *GEXF File Format*. Paris, France: The Gephi Consortium - Médialab Sciences Po, 2007.
[27] Bastian, Mathieu, *Gephi*. Paris, France: Médialab Sciences Po, 2009.
[28] A. Bilgin, J. Ellson, E. Gansner, Y. Hu, et S. North, *windows | Graphviz - Graph Visualization Software*. AT&T Research, Yahoo, Google, Apple, 2015.
[29] Raphaël Velt, *GEXF-JS*. Paris, France, 2011.
[30] Ratinaud, Pierre, *Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires*. LERASS, 2008.
[31] E. Marty, P. Marchand, et P. Ratinaud, « Les médias et l'opinion-eléments théoriques et méthodologiques pour une analyse du débat sur l'identité nationale », *Bull. Sociol. Methodol. Méthodologie Sociol.*, vol. 117, nº 1, p. 46-60, 2013.

[32] M. Reinert, « Alceste une méthodologie d'analyse des données textuelles et une application: Aurelia De Gerard De Nerval », *Bull. Méthodologie Sociol.*, vol. 26, nº 1, p. 24–54, 1990.
[33] J. Stefanowski et D. Weiss, « Carrot and Language Properties in Web Search Results Clustering », in *Web Intelligence, First International Atlantic Web Intelligence Conference, AWIC 2003, Madrid, Spain, May 5-6, 2003, Proceedings*, 2003, p. 240–249.
[34] Weiss, Dawid et Osinski, Stanislaw, *Carrot2 - Open Source Search Results Clustering Engine*. 2004.
[35] S. Osiński et D. Weiss, « Carrot2: Design of a flexible and efficient web information retrieval framework », in *Advances in Web Intelligence*, Springer, 2005, p. 439–444.
[36] H. Noh, Y. Jo, et S. Lee, « Keyword selection and processing strategy for applying text mining to patent analysis », *Expert Syst. Appl.*, vol. 42, nº 9, p. 4348-4360, 2015.
[37] S. Osiński, J. Stefanowski, et D. Weiss, « Lingo: Search Results Clustering Algorithm Based on Singular Value Decomposition », in *Intelligent Information Processing and Web Mining, Proceedings of the International IIS: IIPWM'04 Conference held in Zakopane, Poland, May 17-20, 2004*, 2004, p. 359–368.
[38] A. Gonzales-Aguilar et M. Ramírez-Posada, « Carrot2: búsqueda y visualización de la información », *El Prof. Inf.*, vol. 21, nº 1, p. 105–112, 2012.
[39] Polivaev, Dimitry, *Home - Freeplane - free mind mapping and knowledge management software*. 2000.
[40] M. Davies, « Concept mapping, mind mapping and argument mapping: what are the differences and do they matter? », *High. Educ.*, vol. 62, nº 3, p. 279-301, sept. 2011.
[41] T. Buzan, *Use your head*, 2e éd. London: BBC Books, 1984.
[42] H. J.-M. Dou, « Benchmarking R&D and companies through patent analysis using free databases and special software: a tool to improve innovative thinking », *World Pat. Inf.*, vol. 26, nº 4, p. 297-309, déc. 2004.