Technology Analysis & Strategic Management

ISSN: 0953-7325 (Print) 1465-3990 (Online) Journal homepage: https://www.tandfonline.com/loi/ctas20

Comparing data sources for identifying technology trends

N. Mikova & A. Sokolova

To cite this article: N. Mikova & A. Sokolova (2019) Comparing data sources for identifying technology trends, Technology Analysis & Strategic Management, 31:11, 1353-1367, DOI: 10.1080/09537325.2019.1614157

To link to this article: https://doi.org/10.1080/09537325.2019.1614157

Published online: 10 May 2019.

TECHNOLOGY ANALYSIS & STRATEGIC MANAGEMENT
2019, VOL. 31, NO. 11, 1353–1367
https://doi.org/10.1080/09537325.2019.1614157

Comparing data sources for identifying technology trends

N. Mikova and A. Sokolova

Institute for Statistical Studies and Economics of Knowledge, National Research University Higher School of Economics, Moscow, Russia

ABSTRACT

This paper considers the strategies for working with different data sources for identifying technology trends. For this purpose, a comparative analysis of technology monitoring results using various data collections (scientific publications, patents, media, foresight projects, conferences, international projects, dissertations and presentations) is conducted. Guidance on how to use them to ensure the greatest output is presented. Green energy is taken as an example for comparative analysis and provides improvements in reducing inputs (time) and increasing output (coverage). The factors that affect data processing results are considered and discussed to more efficiently use quantitative and qualitative procedures for identifying, correcting and updating technology trends. The results of the study can be interesting for government bodies financing foresight studies and setting priorities in science and technology, for companies scanning disruptive innovations in the markets to support their corporate strategies and academic community developing the methodology for technology trends monitoring.

ARTICLE HISTORY
Received 18 April 2018
Revised 3 March 2019
Accepted 23 April 2019

KEYWORDS
Technology foresight; technology trends monitoring; data sources; green energy

Introduction

Earlier identification and regular monitoring of technology trends, which have a key impact on social and economic development in the long term, provide stakeholders of different levels (global, national, corporate) with a distinct advantage for reacting quickly to technological changes and making strategic decisions in a timely manner. The timeliness and accuracy of technology monitoring depends on the right choice of a research methodology and data sources. Theoretical studies on technology monitoring are mainly based on the combination of qualitative and quantitative methods that complement each other. Along with extensive use of expert procedures, automated and semi-automated methods for extracting technology trends are increasingly being developed and becoming more important as evidence-based practice. A broad range of information sources can be used for processing such quantitative data: scientific publications, patents, media, business resources and others.

In academic works that aim to discover technology trends, the most widely used data sources are scientific publications (e.g. Chen 2006; Cobo et al. 2011; Daim et al. 2006; Guo, Weingart, and Borner 2011; Kajikawa et al. 2008; Upham and Small 2010) and patents (e.g. Daim et al. 2006; Lee, Lee, and Yoon 2011; Trappey et al. 2006; Tseng, Lin, and Lin 2007; Wang, Chang, and Kao 2010; Yoon and Park 2004). Some of these studies propose the general methodology for monitoring emerging technologies, research fronts, 'white spots' and potential research areas. The other discuss the advantages and disadvantages of using various methods and techniques for technology monitoring. In addition,

CONTACT N. Mikova nmikova@hse.ru Institute for Statistical Studies and Economics of Knowledge, National Research University Higher School of Economics, Myasnitskaya st., 20, Moscow 101000, Russia

a number of works apply the data visualisation methods, such as technology maps and technology networks, and develop and utilise the program software for data processing (e.g. Vantage Point, CiteSpace, Science of Science). Although many authors tend to choose one key information source – either scientific publications or patents – while processing data on technology trends, there have also been attempts to use a combination of them. For instance, Shibata, Kajikawa, and Sakata (2010) compare the structure of citation networks of publications and patents in order to discover the differences between the two and identify non-commercialised gaps between science (publications) and technology (patents).

The possibilities of using other information sources for identifying technology trends have also been discussed in the literature. For example, much useful data can be found in additional sources: *specialised conferences* (Porter and Cunningham 2005); *newspapers and social media* (Farber 2016; Krzywicki et al. 2016); *business-related resources*, such as Lexis-Nexis (Porter and Cunningham 2005); *statistics on venture capital funds and start-ups* (Cozzens et al. 2010); *technical reports and 'grey' literature* (Porter and Cunningham 2005) and others. Chen, Chiang, and Storey (2012) study the evolution of data types used in business intelligence and analytics (BI&A). They distinguish three generations of BI&A: BI&A 1.0 – uses structured *business-based commercial data*, BI&A 2.0 – is based on unstructured web-based content, such as *forums, online groups, web blogs, social networking sites, social multimedia sites (for photos and videos)*, and even *virtual worlds and social games* (O'Reilly 2005); and BI&A 3.0 – employs mobile and sensor-based content: *large- scale and fluid mobile and sensor data*. In addition, specific data sources can be used for BI&A in science and technology (S&T): e.g. *S&T instruments and system-generated data*, *sensor and network content*. Jordan and Mitchel (2015) discuss the machine learning methods that have been developed for capturing and mining large quantities of data from the following sources: *medical records*, *traffic data*, *crime data*, *large experimental data sets*, etc. Segev, Jung, and Choi (2015) suggest the method for analysing technology trends through investigation their development over time, based on different information sources. They retrieve the trends not only from patents and academic articles, but also from *web searches*, *news articles*, *book publications*, and compare them with *Gartner technology predictions*. The research results show, that more research-oriented data sources such as academic articles and patents have a longer predictive time window and higher accuracy than news, books and web searches.

Nevertheless, investigation of the comparative efficiency of different data sources and the specificity of how they are selected is limited in the literature. Martino (2003) is one of the examples of such research, in which the author studies the differences in emphasis contained in technology trends based on the stage of their life cycle. According to this criterion, the data sources are divided into five categories (see Table 1).

Nevertheless, such factor as *a technology life-cycle stage* is not the one that exhausts all possibilities and ways of employing diverse information sources. The selection of them depends also on *the purpose of a study*, its *methodology*, *available resources* and other parameters of the project. In addition, this choice can be determined by *what is understood as 'a technology trend*.' Different authors propose various interpretations of this term and use other notions associated with it (Mikova and Sokolova 2014). These concepts can differ depending on their *expected effects* (e.g. disruptive innovations, game-changers, emerging technologies, key enabling technologies, technology

Table 1. Choosing data sources based on the technology life-cycle stage

R&D stage	Typical source
Basic research	Science Citation Index
Applied research	Engineering Index (Compendex)
Development	USPTO patents database
Application	Newspapers Abstracts Daily
Social impacts	Business and popular press

Source: Martino (2003).

applications), *the life-cycle stage* (e.g. emerging technologies, technological applications and products), *the scale* (e.g. mega trends, micro trends). In addition, technology trends can differ in *the way how they are identified*. For example, research fronts are defined as clusters of documents detected on the basis of co-citation analysis (Upham and Small 2010). The present study considers technology trends in more general terms, such as topical, cutting-edge and quickly developing technology areas that can significantly affect the development of the economy and society in the long term.

Therefore, taking into account the fundamental impact of the choice of data sources on technology monitoring outputs and significant underdevelopment of this issue in the scientific literature, the purpose of this paper is to conduct the exploratory study of the strategies for working with different information sources for technology trends monitoring, as well as of the factors that should be taken into account when choosing them. The authors compare on a systemic basis the results of technology monitoring, obtained from various data collections (scientific publications, patents, media, foresight projects, conferences, international projects, dissertations, presentations), using the area of green energy as an example.

According to the integrated approach of UNDP (2016), the green (renewable) energy lies at the intersection of several development areas, such as achievement of climate targets, reduction of disaster risks associated with rising temperatures and better recovery from the disaster events. Green energy technologies contribute to a zero-carbon, risk-informed and sustainable development (UNDP 2016). Selection of green energy as an example for comparing information sources has several advantages. First, it is an interdisciplinary area that has many links with other fields and adjacent technologies. Second, green energy has no clear boundaries, that is why the results largely depend on the correct strategy for extracting trends and this area gives more opportunities for exploring diverse search strategies. Third, the interest in green energy area is also caused by the difference in keywords used in various data sources, which enables one to further discuss the influence of the terminology on the choice of information sources and thus on technology monitoring results.

The paper is organised as follows. At first, the theoretical background of the specificity of different information sources in terms of their usage and possible data search strategies is provided. Next, the methodological approach is described. Subsequently, the case study of green energy is presented. Finally, the paper discusses the results and draws a conclusion.

Theoretical background: choosing information sources, collecting and processing data

The right choice of data sources for identifying technology trends has a crucial influence on the final outcome. One of the criteria for this selection is the goal and *the tasks* of the project. Table 2 illustrates the examples of correspondence between the information sources, the possible databases and the research tasks.

Another important issue is how to work with these information sources: how to select a database and what methods to use for collecting and processing data. While choosing a database, it may be necessary to account for its characteristics as a whole, ensuring that it meets a set of minimum standards in line with information objectives. Considerations may include suitability and comprehensiveness of coverage, biases, content quality, record structure and keywords availability (Porter and Cunningham 2005).

When the database is identified, the important task is to form a search strategy and choose the methods for collecting data. Different options may be applied for this purpose. These can be *a broad query* that describes the subject area in general (for example, 'green energy' or 'sustainable energy'); *a list of keywords* (anything from one phrase to combinations of 30–50 keywords in order to describe the subject area in a complete manner) selected through consulting with experts (Lee, Yoon, and Park 2009; Morris et al. 2002); *the keywords from the most important documents* (Kim,

Table 2. Possible databases and research tasks for using different information sources

Sources	Possible databases	Possible tasks
Scientific publications	*Interdisciplinary:* • Web of science • Scopus • EI Compendex and INSPEC • Pascal • ResearchIndex, etc. *Specific:* • PubMed (MEDLINE) • Chem abstracts • Biological abstracts, etc.	• To study the dynamics of S&T development and growth of scientific interest in certain areas • To track research fronts and emerging technologies • To define 'white spots' that are not yet developed but are potentially significant • To analyse leading countries and research groups engaged in S&T research • To identify exciting research activity and profile research and development (R&D) activity across a specific region to look for potential collaborators
Patents	• Questel-Orbit • Derwent World Patents Index • MicroPatent • Delphion • WIPS • Patbase • PatenCafe.com • IFI CLAIMS, etc.	• To search for information on technology problems and solutions in a subject area • To track the evolution of technology development in a specific area • To analyse leading countries, research groups and companies engaged in S&T research • To identify exciting research activity and profile R&D activity across a specific region to look for potential collaborators • To find information about technology applications and technology products developed in a specific area
Media	• Factiva • LexisNexis • Internet Securities, etc.	• To understand technology supply and demand dynamics in a wide variety of socio-economic areas • To monitor highly discussed topics in S&T areas (news from business sites, news channels, etc.) • To track public interests in general through newspaper article compilations • To find information about social acceptance of technologies and possible barriers for their development
Foresight projects	• European Foresight Monitoring Network (EFMN)	• To detect technology trends and priorities in different areas • To identify and analyse key emerging issues relevant for the future of S&T development • To analyse leading countries and research groups engaged in S&T research • To detect the main institutions sponsoring the future research around the world
Conferences	• Conference websites	• To identify key technology areas, which are of a great interest from representatives in specific areas of knowledge (scientists, business companies and others) • To assess dynamics and prospects for the implementation of novel technologies • To track research fronts and emerging technologies • To identify leading conferences on particular topics and highly active researchers to seek out • To analyse leading countries and research groups engaged in S&T research
International projects	• CORDIS Europe • European Framework Programs (FP) for Research and Technological Development, etc.	• To monitor political incentives and priorities in specific technology areas • To evaluate demand side and markets, S&T performance of different countries • To analyse possibilities for future cooperation in S&T areas • To detect the main institutions sponsoring the future research around the world • To track research funding emphases and research participation patterns

Dissertations	• ProQuest • EBSCO's Open Dissertations database • PQDT Open • EThOS, etc.	• To study the dynamics of S&T development and growth of scientific interest in certain areas • To track research fronts and emerging technologies • To analyse the state of art in a specific area • To count the number of scientists engaged in R&D • To identify leading research groups and relationships between them
Presentations	• SlideShare • Scribd, etc.	• To study the dynamics of S&T development and growth of scientific interest in certain areas • To identify key technology areas which are of a great interest from representatives in specific areas of knowledge (scientists, business companies, etc.) • To collect the concentrated information about new innovation ideas, technologies and their applications • To detect individuals and research organizations who share their ideas, conduct research, connect with each other and generate decisions for their businesses

Sources: Cozzens et al. 2010; Ena et al. 2016; Kostoff et al. 2008; Porter and Cunningham 2005; consultations with the technology mining experts.

Table 3. Advantages and disadvantages of possible search strategies

Possible search strategies	Advantages	Disadvantages
A broad query	(+) may save time at the stage of collecting data from the database (+) may be useful for the clearly defined areas (+) allows to capture many articles across multiple disciplines or fields (transfer of ideas)	(–) may retrieve too many documents (–) may provide less precision and risk to extract more results that are not pertinent to the subject of the analysis (more time needed to clean the data afterwards)
A list of keywords	(+) is more narrow query capturing fewer articles of a more specific nature (+) requires the use of the expert knowledge to create a list of keywords, which is an additional validation of them	(–) may not take into account all the synonyms (–) may require more deep involvement of experts, which increase time and the costs of getting information
The keywords from the most important documents	(+) represents the scope of research confirmed by the authors or the database administrators (+) may save time and reduce dependence on the experts during creating the data collection	(–) may not take into account the adjacent documents (from the other disciplines) (–) is limited to the analysis of only main documents, while important seeds of innovation from other documents may be neglected
A nominal query (e.g. articles from specialised journals, patents from specific classes, etc.)	(+) can be particularly important for competitive technological intelligence (+) saves time for creating 'the right query' in order to determine the scope of the area more precisely (specific topics of investigation, specific countries, specific authors, etc.)	(–) requires compiling an initial list of specialised journals or patent classes for collecting data (–) narrows the search to the concrete topics, patent classes, etc., and therefore may not include the important documents from the other related areas (–) may not account for the problem of ambiguities of attribution of the documents (publications, patents, etc.) to different classes and possible classification errors
The most highly cited documents	(+) represents the 'hot research topics' in a specific area (+) may be valuable in competitive technological intelligence, where it is important to determine the experts in a field through citation analysis	(–) requires citation information, which is presented in only few databases (–) may cause 'the Matthew effect', when the papers already cited become easier to find and more attractive to scientists looking for key references, and as a result 'the rich get richer' (–) requires suitable normalisation and expert review during citation comparisons among separated research communities

Sources: based on Porter and Cunningham (2005).

Suh, and Park 2008); or *a combination of these methods* (Kim, Suh, and Park 2008; Porter and Cunningham 2005). An alternative search strategy is to compile a set of documents based on *a nominal query*, such as articles from specialised journals (Cobo et al. 2011; Guo, Weingart, and Borner 2011; Kajikawa et al. 2008; Kostoff et al. 2008), the most highly cited publications (Upham and Small 2010), patents from specific classes of the International Patent Classification (Corrocher, Malerba, and Montobbio 2003; Lee, Lee, and Yoon 2011) or patents sponsored by a particular government agency (Tseng, Lin, and Lin 2007). The advantages and disadvantages of these search strategies are briefly summarised in Table 3.

In the first two search strategies, a query can be applied to different data fields: title, abstract, keywords or full text. Each possibility has pros and cons. *Title* usually represents the compact information about the documents' content and often includes the main keywords, but some names of the documents serve more to attract the reader's attention and not precisely reflect the content. In *abstract* the authors describe essential ideas of work briefly and clearly, trying to incorporate the key terms. But at the same time, for example, patent applicants may not want to communicate their intents fully and directly (Porter and Cunningham 2005). *Keywords* represent the concentrated and specific information about the documents' content, but usually are limited to the specific number of phrases, which may be not enough to fully describe a document. *Full text* gives more detailed

and complete information about the content, but such 'unstructured' materials may include a lot of noise and require more time and technical capabilities to process (Porter and Cunningham 2005). Therefore, the final choice of the data fields depends on the project task, the methodology, the subject area, the information sources and other parameters.

During technology trends monitoring scientists may deal with both ***structured*** and ***unstructured data***. The most well-known methods for processing structured data are *bibliometric* and *patent analysis*, while *text mining* is a popular technique for handling huge amounts of unstructured textual documents. A number of theoretical works have been devoted to development and usage of automated software for processing unstructured data, including linguistic and statistical methods and visualisation tools (Chen 2006; Dereli and Durmusoglu 2009; Guo, Weingart, and Borner 2011; Morris et al. 2002; Palomino, Vincenti, and Owen 2013; Porter and Cunningham 2005). These programs may work online (e.g. Carrot, PAS) or offline (e.g. Vantage Point (Porter and Cunningham 2005), CiteSpace (Chen 2006), DIVA (Morris et al. 2002), Science of Science (Guo, Weingart, and Borner 2011), TextAnalyst (Wang, Chang, and Kao 2010), Arrowsmith (Smalheiser 2001), PackMOLE (Fattori, Pedrazzi, and Turra 2003)). As a rule, they use data from electronic databases (publications, patents, news, etc.) and have a special user interface for query creation, data filtration and visualisation. *Network analysis* and *cluster analysis* are frequently used in text mining. Network analysis is based on the graph theory and allows to detect, analyse and visualise relationships between the objects (documents, authors, thematic areas, countries, keywords, etc.). Clustering is applied to divide the prepared data into groups with similar characteristics. The most popular methods include: k-means (Kim, Suh, and Park 2008; Trappey et al. 2006), hierarchical (Kostoff et al. 2008; Lee, Lee, and Yoon 2011) and topological clustering (Kajikawa et al. 2008; Shibata, Kajikawa, and Sakata 2010), k-nearest neighbours algorithm (Tseng, Lin, and Lin 2007) and others.

Methodology

In this study comparative analysis of eight data sources (scientific publications, patents, media, foresight projects, conferences, international projects, dissertations and presentations) was started with extracting potential trends from each of them independently, using the same methodological approach (see Figure 1). After that the results were validated by the experts in order to compile the final list of trends.

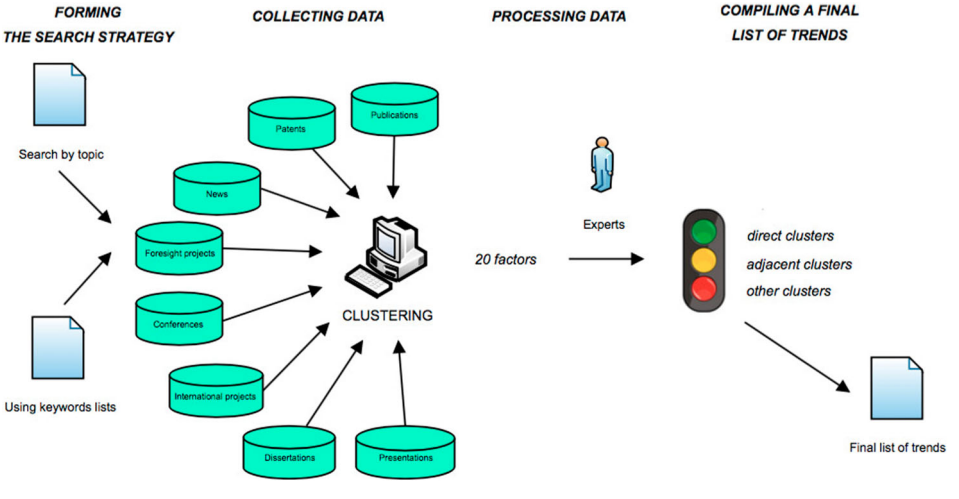

Figure 1. Research methodology

The methodology includes four stages:

- *Stage 1.* Forming the search strategy
- *Stage 2.* Collecting data for identifying technology trends
- *Stage 3.* Processing data
- *Stage 4.* Compiling a final list of trends

At the first stage, two main methods were chosen for collecting data about technology trends: *search by topic* and *using the keywords provided by experts*. These methods are combined in order to increase the consistency of collections. First, the main topics for data searching were selected (they differ in the various databases: 'sustainable energy,' 'renewable energy,' 'alternative energy sources', etc.). Second, specific keywords were used for creating the sets of documents found in these chosen areas. A list of 30 keywords was created for the green energy area in consultation with experts. The experts were asked to compile the lists of keywords, which outline the main scope of green energy area and include all important subareas. For example, this list includes such keywords as 'wind energy,' 'solar energy,' 'biofuels,' 'geothermal energy' and others. The patent search was performed using *the specific patent classes*, which were preliminarily discussed with the experts, since due to more technical character of patent data, using the same list of keywords as for the other resources (publications, dissertations, etc.) did not help collect the relevant documents. In order to increase the technology monitoring efficiency, different ways of data collecting can be explored in detail at the next stages of the research.

At the second stage, eight data collections were compiled based on a combined search strategy described above. Table 4 shows the aggregated information: data sources, the number of documents and data fields being processed. Eight collections of English-language documents covering the period of the last 10 years were generated. The data field 'keywords,' which represents the concentrated information about the documents' content, was used as a priority field for processing data (scientific publications, dissertations). If it was not available in a database, the data fields 'abstract' (for patents) and 'title' (for media) were applied. To provide the same level of analysis, the field 'claim' was not used for patents, since it contains the specification of the document in too technical terms. In the case of dealing with unstructured data (foresight projects, conferences and international projects) processing the full texts was the only option.

At this stage a total of 35986 documents in different formats (*.txt, *.html, *.doc, *.pdf, *.ppt) were collected. The collections, including semi-structured data (title, abstract, keywords), were ready for uploading into Vantage Point software[1] using specialised import filters (for Web of Science, Questel-Orbit, Factiva). The collections with unstructured data (full texts) were preliminary converted into *.smart XML format for further processing.

At the third stage, eight collections were processed in Vantage Point and the factor maps for each collection were created and analysed. To compare the results of dividing collections using different numbers of factors (clusters), the data from each collection were clustered into 5, 10, 15 and 20 factors. After the analysis of these clusters it was concluded that the data obtained by dividing

Table 4. Aggregated data on the collections

Collections	Data sources	Number of documents	Data fields for processing
Scientific publications	Web of Science	6222	Keywords
Patents	Questel-Orbit	7775	Abstract
Media	Factiva	21502	Title
Foresight projects	EFMN	40	Text
Conferences	Official websites	19	Text
International projects	CORDIS Europe	129	Text
Dissertations	ProQuest	199	Keywords
Presentations	SlideShare	100	Text

into 20 factors is a subset for the data received by using 5, 10 and 15 factors. That is why the number of 20 clusters was chosen for use in further analysis.

At the fourth stage, the final list of technology trends was created. First, the analysis of clusters was performed. Then, the experts were consulted in order to select the relevant technology trends from them. A preliminary analysis of clusters was done to understand the core idea of them, based on studying their contributing keywords. When the lists of analytical clusters were created for each collection, the experts were asked to divide them into three groups: *direct clusters* (directly related to the subject area), *adjacent clusters* (indirectly related to the subject area) and *other clusters* (related to other areas), and the final list of technology trends was compiled on this basis. The influence of expert's 'subjectivity' on the final results was relatively low, since most of the work was done analytically and the experts' role was only to validate the obtained clusters and choose the relevant ones, which represent technology mainstreams in the green energy area.

Results

The results for each data collection, obtained from division of 20 clusters into three cluster types, are presented in Table 5.

Table 5 shows that the highest number of direct clusters (related to green energy) belongs to the following sources: scientific publications, patents, media and conferences. The number of *direct clusters* represents the quality of a data collection and of a chosen search strategy (the higher the number of direct clusters, the more efficient is the data collection). The high number of *adjacent clusters* can serve as both an advantage and disadvantage of using a specific collection, which depends on the project goal, as they can be of a great interest for understanding the technology development in the border areas. The high number of *other clusters* indicates that a search strategy was too broad and not effective for a specific data collection. At the same time other clusters can provide additional relevant information, which is analysed at the end of this section.

Table 6 includes additional information about how the direct clusters are represented in the final list of trends. The names of the final (linked) trends were formulated in the following way. After automated data processing, the keywords for each cluster were retrieved from Vantage Point and then analysed and discussed with experts in order to correctly name the trends. In addition, Table 6 shows the names of the extracted trends for each collection, where the same trends from different collections are marked with the same sign (*, **, ***, etc.). For some collections the number of the linked trends is lower than the number of direct clusters, because not all direct clusters were validated and included into the final list of trends.

Through the procedures described above the lists of 20 trends obtained for each collection were studied. The analysis showed that the following topics are the best represented in the results of processing all eight collections: solar energy (especially the trend 'Solar cells'), bioenergy (the trends 'Biogas production,' 'Biodiesel production,' 'Bioethanol production'), hydrogen energy (the trend 'Hydrogen storage'), hydropower (the trend 'Large hydropower stations') and energy storage (the trend 'Electrochemical cells'). Besides there are some trends that are extracted from only one

Table 5. Aggregated data on the selected clusters

Collections	Direct clusters	Adjacent clusters	Other clusters
Scientific publications	10	5	5
Patents	8	6	6
Media	7	4	9
Foresight projects	4	1	15
Conferences	7	5	8
International projects	4	5	11
Dissertations	5	4	11
Presentations	6	2	12

Table 6. Aggregated data on the clusters and trends

Collections	Direct clusters (number)	Linked trends (number)	Linked trends (names)
Scientific publications	10	9	Solar thermal power stations* On-shore wind farms** Geothermal heat pumps*** Electrochemical cells**** Proton exchange membrane fuel cells (PEMFC)***** Solid oxide fuel cells (SOFC)****** Molten carbonate fuel cells (MCFC) Electrolysis of water******* Hydrogen storage
Patents	8	8	Solar cells# Solar thermal power stations* Off-shore wind farms Biodiesel production## Large hydropower stations### Geothermal heat pumps*** Solid oxide fuel cells (SOFC)****** Electrolysis of water*******
Media	7	5	Solar thermal collectors+ On-shore wind farms** Bioethanol production++ Small hydropower stations+++ Hydrogen storage++++
Foresight projects	4	4	Solar cells# Biogas production§ Bioethanol production++ Hydrogen storage++++
Conferences	7	6	Solar cells# On-shore wind farms** Large hydropower stations### Small hydropower stations+++ Geothermal power stations Hot dry rock geothermal power stations
International projects	4	4	Solar cells# Biogas production§ Biodiesel production## Bioethanol production++
Dissertations	5	5	Solar cells# Biogas production§ Electrochemical cells**** Proton exchange membrane fuel cells (PEMFC)***** Hydrogen storage++++
Presentations	6	5	Solar thermal collectors+ Bioethanol production++ Large hydropower stations### Electrochemical cells**** Pumped-storage power plants

particular collection (e.g. 'Off-shore wind farms,' 'Geothermal power stations,' 'Hot dry rock geothermal power stations,' 'Pumped-storage power plants,' 'Molten carbonate fuel cells (MCFC)'). From this it can be concluded that using several collections could be more efficient than employing only one of them, since in the first case, technology monitoring covers a wide range of sources and information platforms where data on technology development could be represented.

The analysis of the *other clusters* shows that they can be divided into the following groups:

(1) *Technical devices, technologies and models*

This group includes smaller-scale technology applications (at the level of sub-trends), specific technologies and models from green energy and adjacent fields, which can be used to study the process of energy transformation in different systems. Examples include the transmission electron microscope, self-propagating high-temperature synthesis and others. Clusters from this group are most widely represented in collections 'Scientific publications,' 'Patents' and 'Dissertations.'

(2) *Green energy related technology areas from the wider energy field* (energy efficiency and energy saving)*, as well as the other related areas* (such as the rational use of nature, nanotechnology and transport systems)

The fact that the analytical list includes not only the clusters from green energy and the wider energy area, but also the clusters from related disciplines, can point to the strong links between different technology fields. Examples of areas that are developing between adjacent disciplines include: interactive mapping technologies for connecting manufacturers and potential consumers of organic waste in order to produce biogas (green energy and the rational use of nature), using of environmentally-friendly heavy-duty plastics (green energy and nanotechnology) and others. The clusters from adjacent areas occur most frequently in the collections 'Media,' 'Conferences,' 'International projects', 'Dissertations' and 'Presentations.'

(3) Social, economic and political challenges

This group includes analytical clusters that involve information about the global problems (social, economic, political, etc.) that can be addressed by the technology solutions from green energy area and related fields. The examples of such trends include: the need to make life in the city more comfortable by using green technology, improving green energy education and making it more accessible in different regions of the world. Socio-economic clusters are detected in the collections 'Foresight projects,' 'International projects' and 'Presentations.'

(4) Irrelevant clusters

When analysing the analytical list of clusters across all collections, a small number of clusters (around 2.5%), which are neither directly nor indirectly related to green energy, were discovered. Examples of irrelevant clusters include: the increased attention to weight loss and low-calorie foods (such as green tea); development of the ways for diagnosing and treating Alzheimer's disease and others. Such clusters were discovered while processing the collections 'Scientific publications,' 'Patents' and 'Media.'

Therefore, the information retrieved from the analytical clusters, which were not included in the final trends list, is very useful. First, one can apply it for analysis of the specificity of different collections. Second, it may help identify the ground for selecting the best information sources. Finally, it can be used for further improvement of the methodology for monitoring and updating technology trends.

Discussion

Considering the case of green energy in detail, the possibilities and limitations of using different data collections and the lessons learned are discussed below.

The collection *'Scientific publications'* includes the greatest number of 'contributing' trends (with a value added to the final list of trends) in comparison with the other collections. Such trends are: 'Molten carbonate fuel cells (MCFC),' 'Solar thermal power stations,' 'On-shore wind farms,' 'Geothermal heat pumps,' 'Proton exchange membrane fuel cells (PEMFC),' 'Solid oxide fuel cells (SOFC)' and 'Electrolysis of water.' One should take into account the fact that in publications some emerging areas will obviously be the beginnings of the new scientific fields, not the new technologies. In order to use the possibilities of this collection more efficiently, a list of 'more technological' terms could be used.

The collection *'Patents'* is the second leader in 'contributing' trends, which are the following: 'Offshore wind farms,' 'Solar thermal power stations,' 'Geothermal heat pumps,' 'Solid oxide fuel cells (SOFC)' and 'Electrolysis of water.' In general, patents can give detailed information on specific technical devices, technologies and models (sub-trends) in the subject area. At the same time, working with this collection, one should pay attention to a specificity of patenting in different disciplines and carefully choose the patent classes for collecting information, for example, with the help of experts.

The collection '*Conferences'* can also be considered as a collection with a great added value. It includes information about the following 'contributing' trends: 'Geothermal power stations,' 'Hot dry rock geothermal power stations' and 'Small hydropower stations.' In order to improve the results, one could refer to an expert opinion on a preliminary selection of the key conferences. In addition, as data about conferences is mostly unstructured (*.html, *.doc, *.pdf, *.ppt), the databases of conferences with more structured information could be explored.

A number of 'contributing' trends are also presented in the collections '*Media,'* ('Small hydropower stations,') 'Dissertations,' ('Proton exchange membrane fuel cells (PEMFC),') and 'Presentations' ('Pumped-storage power plants.') In order to increase the efficiency of technology monitoring using these collections, one should be aware that the collection 'Media' helps examine sources written by outside observers, analysts or journalists, and may reflect the relationships between technology trends and socio-economic impacts. That is why the keyword list for this collection could include some less technical, but more 'trendy' terms like 'sustainable energy,' 'environmentally friendly energy production' and others. At the same time, the terms for creating the collection *'Dissertations'* could be much more technical.

The collection '*Presentations,'* which could act as an additional business- and technology-oriented data source, includes the concentrated information about specific technologies, technology applications and innovative ideas, since the duration of a presentation is usually limited. However, while dealing with this source, one should take into account that some important data may be represented in pictures and graphs. Therefore there is a need to convert them into a structured format for further analysis.

Being the future-oriented data sources, the collections *'Foresight projects'* and *'International projects'* help reveal information on the priorities in technology fields and focus on prospective mainstreams of technology development. As in case of 'Media,' one could use some more specific keywords for creating these collections, without using difficult technical terms. Since the very projects already include the structured information about technology trends in the subject area, in some cases it could be more beneficial to process this data manually in order to increase the added value of this collection.

By comparing the results of processing different data sources for identifying technology trends in green energy, it can be concluded that the opportunities offered and the results obtained by using any of the eight represented collections may depend on the following factors:

(1) The specificity of a subject area

For example, political factors can play an important role in the area of green energy, and therefore a large portion of useful information can be contained in the projects and programs for development of energy industry in individual countries and across the globe.

(2) The technology life-cycle stage

The collections 'Scientific publications,' 'Patents' and 'Dissertations' can be used to analyse fundamental and applied research. 'Media,' 'Conferences' and 'Presentations' can be helpful in discovering the application. At the same time, 'Foresight projects,' 'International projects' and 'Presentations' can be suitable for evaluating the social and economic impact of technology development.

(3) The choice of information resources

For example, while using electronic databases it is necessary to take into account the time required to present information in a structured format. The analysis of online content makes it possible to process information in real time.

(4) The search strategy

Searching using a broad query may be applied while studying the mainstreams of technology development in a specific area, while searching by topic or using a keywords list can be more effective when analysing technology trends on a deeper level.

(5) The difference in terminology
Diverse keywords can be used for describing technology trends in the subject area, for example, more scientific terms like 'environment and energy' in the publications, and more 'trendy' terms like 'sustainable energy' in media.

(6) The choice of data fields to be processed
Information from some data fields, such as title or abstract, may be enough to search for mainstream technological trends, while processing the full text may be helpful for detecting more specific technology applications.

An important step in interpreting the results of data processing is the separation between *supply side* and *demand side* of the technology development. In order to identify these two types of trends, different data collections created by using specific keywords lists, can be applied. For instance, the collections 'Foresight projects,' 'International projects' and 'Presentations' can be more suitable for monitoring demand side (socio-economic trends), while 'Scientific publications,' 'Patents,' 'Media,' 'Conferences' and 'Dissertations' could be more helpful for discovering supply side (technology trends).

Conclusion

This study compares the results of identifying technology trends extracted from various data sources (scientific publications, patents, media, foresight projects, conferences, international projects, dissertations and presentations) in order to analyse the strategies for working with different collections, as well as the factors that may influence the technology monitoring results. The proposed approach is applied for green energy as a multidisciplinary area connected with the different fields of knowledge. The analysis shows that comprehensive coverage of data sources can be more efficient than employing only one resource, since this makes it possible to discover technology trends at the earliest stage of their development, for example, using the latest news published online or the materials from the recent conferences. Taking into consideration the specificity of the subject area and each data collection, as well as the accurate formulation of a research goal, can help select the most relevant information sources for identifying different types of technology trends. Therefore, provided that the factors, which affect the results of data processing, are taken into account as much as possible, such systemic analysis of the results can be used as an important tool for further improvement of the methodology for technology trends monitoring.

Note

1. Since the research implies working with both structured and unstructured data, there was a need for a powerful software, which combines bibliometric techniques and text mining tools and offers the visualising possibilities. Vantage Point was chosen for this purpose, since it is designed to accept virtually any structured text content, supports for over 190 different import filters, has industry leading data manipulation capability and contains mapping and visualisation tools to help analyse complex relationships (Search Technology 2016).

Acknowledgements

The article was prepared within the framework of the Basic Research Program at the National Research University Higher School of Economics (HSE) and supported within the framework of a subsidy by the Russian Academic Excellence Project '5-100'. The authors would like to thank Sergey P. Filippov for his significant expert support of this research.

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes on contributors

N. Mikova is a research fellow at Institute for Statistical Studies and Economics of Knowledge, National Research University Higher School of Economics.

A. Sokolova is a senior research fellow at Institute for Statistical Studies and Economics of Knowledge, National Research University Higher School of Economics.

ORCID

Nadezhda Mikova http://orcid.org/0000-0002-3242-1913
A. Sokolova http://orcid.org/0000-0003-0323-4374

References

Chen, Ch. 2006. "CiteSpace II: Detecting and Visualising Emerging Trends and Transient Patterns in Scientific Literature." *Journal of the American Society for Information Science & Technology* 57 (3): 359–377.

Chen, H., R. H. L. Chiang, and V. C. Storey. 2012. "Business Intelligence and Analytics: From Big Data to Big Impact." *MIS Quarterly* 36 (4): 1165–1188.

Cobo, M. J., A. G. Lopez-Herrera, E. Herrera-Viedma, and F. Herrera. 2011. "An Approach for Detecting, Quantifying, and Visualising the Evolution of a Research Field: A Practical Application to the Fuzzy Sets Theory Field." *Journal for Informetrics* 5 (1): 146–166.

Corrocher, N., F. Malerba, and F. Montobbio. 2003. "The Emergence of New Technologies in the ICT Field: Main Actors, Geographical Distribution and Knowledge Sources", TENIA project.

Cozzens, S., S. Gatchair, J. Kang, K.-S. Kim, H. J. Lee, G. Ordóñez, and A. Porter. 2010. "Emerging Technologies: Quantitative Identification and Measurement." *Technology Analysis and Strategic Management* 22 (3): 361–376.

Daim, T. U., G. Rueda, H. Martin, and P. Gerdsri. 2006. "Forecasting Emerging Technologies: Use of Bibliometrics and Patent Analysis." *Technological Forecasting and Social Change* 73 (8): 981–1012.

Dereli, T., and A. Durmusoglu. 2009. "A Trend-Based Patent Alert System for Technology Watch." *Journal of Scientific and Industrial Research* 68: 674–679.

Ena, O., N. Mikova, O. Saritas, and A. Sokolova. 2016. "A Methodology for Technology Trend Monitoring: The Case of Semantic Technologies." *Scientometrics* 108: 1013–1041.

Farber, M. 2016. "Using a Semantic Wiki for Technology Forecast and Technology Monitoring." *Program* 50 (2): 225–242.

Fattori, M., G. Pedrazzi, and R. Turra. 2003. "Text Mining Applied to Patent Mapping: A Practical Business Case." *World Patent Information* 25: 335–342.

Guo, H., S. Weingart, and K. Borner. 2011. "Mixed-Indicators Model for Identifying Emerging Research Areas." *Scientometrics* 89 (1): 421–435.

Jordan, M. I., and T. M. Mitchel. 2015. "Machine Learning: Trends, Perspectives and Prospects." *Science* 349 (6245): 255–260.

Kajikawa, Y., J. Yoshikawa, Y. Takeda, and K. Matsushima. 2008. "Tracking Emerging Technologies in Energy Research: Toward a Roadmap for Sustainable Energy." *Technological Forecasting and Social Change* 75 (6): 771–782.

Kim, Y. G., J. H. Suh, and S. C. Park. 2008. "Visualisation of Patent Analysis for Emerging Technology." *Expert Systems with Applications* 34 (3): 1804–1812.

Kostoff, R. N., M. B. Briggs, J. L. Solka, and R. L. Rushenberg. 2008. "Literature-Related Discovery (LRD): Methodology." *Technological Forecasting and Social Change* 75 (2): 186–202.

Krzywicki, A., W. Wobcke, M. Bain, J. C. Martinez, and P. Compton. 2016. "Data Mining for Building Knowledge Bases: Techniques, Architectures and Applications." *The Knowledge Engineering Review* 31 (2): 97–123.

Lee, H., S. Lee, and B. Yoon. 2011. "Technology Clustering Based on Evolutionary Patterns: The Case of Information and Communications Technologies." *Technological Forecasting and Social Change* 78 (6): 953–967.

Lee, S., B. Yoon, and Y. Park. 2009. "An Approach to Discovering New Technology Opportunities: Keyword-Based Patent Map Approach." *Technovation* 29 (6-7): 481–497.

Martino, J. 2003. "A Review of Selected Recent Advances in Technological Forecasting." *Technological Forecasting and Social Change* 70 (8): 719–733.

Mikova, N., and A. Sokolova. 2014. "Global Technology Trends Monitoring: Theoretical Frameworks and Best Practices." *Foresight-Russia* 8 (4): 64–83.

Morris, S., C. DeYong, Z. Wu, S. Salman, and D. Yemenu. 2002. "DIVA: A Visualisation System for Exploring Document Databases for Technology Forecasting." *Computers and Industrial Engineering* 43 (4): 841–862.

O'Reilly, T. 2005. "What Is Web 2.0? Design Patterns and Business Models for the Next Generation of Software." https://mpra.ub.uni-muenchen.de/4580/1/MPRA_paper_4580.pdf.

Palomino, M. A., A. Vincenti, and R. Owen. 2013. "Optimising Web-Based Information Retrieval Methods for Horizon Scanning." *Foresight (Los Angeles, CA)* 15 (3): 159–176.

Porter, A. L., and S. W. Cunningham. 2005. *Tech Mining: Exploiting New Technologies for Competitive Advantage*. Hoboken, NJ: John Wiley & Sons, Inc.
Search Technology. 2016. The Vantage Point. https://www.thevantagepoint.com.
Segev, A., S. Jung, and S. Choi. 2015. "Analysis of Technology Trends Based on Diverse Data Sources." *IEEE Transactions on Service Computing* 8 (6): 903–915.
Shibata, N., Y. Kajikawa, and I. Sakata. 2010. "Extracting the Commercialization Gap Between Science and Technology – Case Study of a Solar Cell." *Technological Forecasting and Social Change* 77: 1147–1155.
Smalheiser, N. R. 2001. "Predicting Emerging Technologies with the Aid of Text-Based Data Mining: The Micro Approach." *Technovation* 21: 689–693.
Trappey, A. J. C., F.-Ch. Hsu, Ch.V. Trappey, and Ch-I. Lin. 2006. "Development of a Patent Document Classification and Search Platform Using a Back-Propagation Network." *Expert Systems with Applications* 31: 755–765.
Tseng, Y.-H., C.-J. Lin, and Y.-I. Lin. 2007. "Text Mining Techniques for Patent Analysis." *Information Processing and Management* 43 (5): 1216–1247.
UNDP. 2016. Renewable Energy. http://www.undp.org/content/undp/en/home/climate-and-disaster-resilience/sustainable-energy/renewable-energy.html.
Upham, S. P., and H. Small. 2010. "Emerging Research Fronts in Science and Technology: Patterns of New Knowledge Development." *Scientometrics* 83 (1): 15–38.
Wang, M.-Y., D.-S. Chang, and C.-H. Kao. 2010. "Identifying Technology Trends for R&D Planning Using TRIZ and Text Mining." *R&D Management* 40 (5): 491–509.
Yoon, B., and Y. Park. 2004. "A Text-Mining-Based Patent Network: Analytical Tool for High-Technology Trend." *Journal of High Technology Management Research* 15: 37–50.