Computers in Industry 100 (2018) 244–257

Contents lists available at ScienceDirect

Computers in Industry

journal homepage: www.elsevier.com/locate/compind

Extracting and mapping industry 4.0 technologies using wikipedia

Filippo Chiarello[a,*], Leonello Trivelli[b], Andrea Bonaccorsi[a], Gualtiero Fantoni[c]

[a] Department of Energy, Systems, Territory and Construction Engineering, University of Pisa, Largo Lucio Lazzarino, 2, 56126 Pisa, Italy
[b] Department of Economics and Management, University of Pisa, Via Cosimo Ridolfi, 10, 56124 Pisa, Italy
[c] Department of Mechanical, Nuclear and Production Engineering, University of Pisa, Largo Lucio Lazzarino, 2, 56126 Pisa, Italy

ARTICLE INFO

Keywords:
Industry 4.0
Digital industry
Industrial IoT
Big data
Digital currency
Programming languages
Computing
Embedded systems
IoT
Internet of things

ABSTRACT

The explosion of the interest in the industry 4.0 generated a hype on both academia and business: the former is attracted for the opportunities given by the emergence of such a new field, the latter is pulled by incentives and national investment plans. The Industry 4.0 technological field is not new but it is highly heterogeneous (actually it is the aggregation point of more than 30 different fields of the technology). For this reason, many stakeholders feel uncomfortable since they do not master the whole set of technologies, they manifested a lack of knowledge and problems of communication with other domains.

Actually such problem is twofold, on one side a common vocabulary that helps domain experts to have a mutual understanding is missing Riel et al. [1], on the other side, an overall standardization effort would be beneficial to integrate existing terminologies in a reference architecture for the Industry 4.0 paradigm Smit et al. [2].

One of the basics for solving this issue is the creation of shared semantic for industry 4.0. The paper has an intermediate goal and focuses on the development of an enriched dictionary of Industry 4.0 enabling technologies, with definitions and links between them in order to help the user in actively surfing the new domains by starting from known elements to reach the most far away from his/her background and knowledge.

1. Introduction

Industry 4.0 is getting the center of the scene with respect to the future of production systems in advanced countries and to its economic and social implications. It is considered as the new fundamental paradigm shift in industrial production. The new paradigm is based on the advanced digitalization of factories, the Internet, and future-oriented technologies bringing intelligence in devices, machines, and systems [3]. Despite its growing popularity and the great expectations in terms of innovation impact, the concept of Industry 4.0 remains strongly linked to technologies and frameworks that have been heavily researched and analyzed in the last decades. In particular, Industry 4.0 can be seen as a smart recombination of existing technologies and some new technologies and their application to the manufacturing environment [4]. This recombinant nature has led some authors to claim that it is nothing more than a re-labeling of old technologies, such as Computer Integrated Manufacturing [5].

Yet other authors claim that this new wave of technology is fundamentally different from previous technologies and not just an amalgamation. In order to address the question whether Industry 4.0 is a new paradigm, or rather a re-labeling of existing technologies, a preliminary activity is needed, namely the delineation of the field and the clustering of technologies covered in the perimeter. It turns out that this activity is extremely challenging in the case of Industry 4.0, for a number of reasons we discuss in great detail. Faced with the complexity of Industry 4.0 existing delineation and clustering methodologies can be considered inadequate.

In this paper we develop a novel approach, test it, and show its superior performance with respect to other approaches.

The key features of the approach are as follows:

i) the description of Industry 4.0 is offered in the form of an “enriched dictionary”, or an ordered and comprehensive collection of lemmas, each of which are associated to full scale definitions and descriptions and to explicit linkages to other lemmas;
ii) the description of constituent technologies offered in the enriched dictionary is not obtained from individual experts, but is generated by accessing appropriate pages of the online encyclopedia Wikipedia;
iii) the total number of technologies covered is more than 1200, linked with more than 39,000 semantic relations;
iv) the perimeter of Industry 4.0 is not defined externally to the

* Corresponding author.
E-mail address: filippo.chiarello@destec.unipi.it (F. Chiarello).

https://doi.org/10.1016/j.compind.2018.04.006
Received 25 September 2017; Received in revised form 1 February 2018; Accepted 10 April 2018

technology (by experts, government policies or other external sources) but is generated endogenously by examining the linkages between technologies described in the Wikipedia pages;

v) the update of the descriptions of technologies in the dictionary takes place in real time due to the distributed, parallel and self-controlled activities of authors in the worldwide community of contributors to Wikipedia;

vi) new technologies are automatically included in the dictionary if they exhibit a given threshold of connectivity with those already included in the perimeter.

The paper is structured as follows. We first characterize the field of Industry 4.0 and discuss why it creates a challenge for field delineation and clustering. Second, we review the recent literature on delineation and clustering methods and we show their limitations with respect to the features of Industry 4.0. We then develop a novel methodology to identify and describe in great detail all the technologies involved in the field. A case study shows how the methodology works in operational terms. Finally, the results are discussed and future developments of the research are outlined in the last section of the paper.

2. Industry 4.0 as a multi-technology multi-stakeholder field

Industry 4.0 is the main keyword used by researchers, policy makers and entrepreneurs when describing how worldwide industrial systems will evolve in the near future by leveraging Internet connected technologies to generate new added value for organizations and society [6]. The growing interest is confirmed by the increasing number of academic papers focusing on topics that are related to the so-called “Fourth Industrial Revolution”. As shown in Fig. 1 the query “Industry 4.0” generates 967 papers. Even if the query is very sharp and does not include all the research efforts on the single “enabling technology” it demonstrates an exponential growth of the topic. In Fig. 1 a projection represented by the dotted line is included. The projection has been drawn by considering a constant increase of the derivative calculated as the average of the last 4 years. Our forecast is that in 2017 there will be 575 new papers in Scopus; this estimate is supported by the fact that about 200 papers have been already published before June 2017 (represented by the point). Since previous analyses in Scopus demonstrate a delay between publication and loading of 5–6 months our forecast seems to confirm a growing interest on the topic.

Table 1 shows how the scientific production on Industry 4.0 is divided among the main research fields (multiple attributions are possible in Scopus).

In particular, it is possible to identify field specific-technologies that refers just to one or few sectors/business areas, and general purpose technologies that can be exploited in several sectors/business areas.

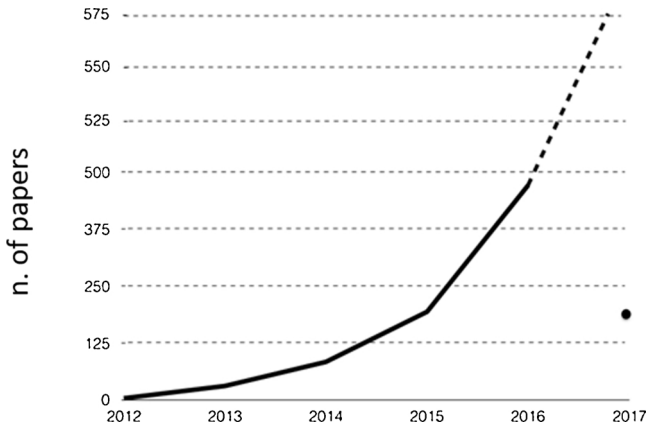

Fig. 1. Trend of publications on Industry 4.0 (Title, Abstract, Keywords). Source: Scopus. Date: 06/06/2017

Table 1
Breakdown of industry 4.0 papers per research field.
Source: SCOPUS date: 06/06/2017

Subject Area	Number of Publications
Engineering	645
Computer Science	410
Business, Management and Accounting	185
Decision Science	134
Material Science	90
Mathematics	87
Chemistry	52
Physics And Astronomy	45
Social Sciences	34
Energy	30

Formulated initially in Germany in 2011, the Industry 4.0 paradigm has been quickly translated, adapted and reinterpreted in developed and developing countries. Table 2 offers a compilation of official documents of governments, agencies and international organizations that in a few years after the initial formulation have embraced the concept.

Despite this rapid and impressive convergence of interest (in itself a clear demonstration of the interdependence of policies across the world), there is no common ground in the definition and delineation of the field even if a first definition of the goal of industry 4.0 have been presented since 1998 [7].

More precisely, while there is a reasonable convergence on the architectural definition of Industry 4.0, as defined in a relatively loose way, there is still considerable disagreement and misalignment with respect to constituent technologies [1,2,8].

Furthermore, many constituent technologies are included in the definition of Industry 4.0, and hence described in these documents, from a variety of perspective that reflect mainly the huge variety of application domains. In other words, technologies are often described not only with respect to their fundamental engineering principles and related dimensions of performance, but with respect to specific applications to various manufacturing or service operations. In these applications the specific working of technologies and the associated dimensions of performance are indeed quite diverse.

Grangel and González (2016) develop a deductive rule-based system able to identify conflicts among AutomationML documents, named ALLIGATOR. It is interesting for the present work to notice how ALLIGATOR has the function to interoperate and align information models between a vast variety of areas (manufacturing, security, logistics) at a micro/plant level. In other words, this paper highlights the fact that one of the main problems of Industry 4.0 is the integration of models and concepts typically developed in their respective domains.

To offer an example of this state of affairs, let us consider the case of RFID (Radio Frequency Identification and Detection) technology. One of the main uses of RFID technology, that is, the detection of the location of a tag moving along a known path with known speed, as illustrated in one of the basic patents [9], can indeed be applied for largely different purposes (safety, tracking, localisation) and in various company areas (production, logistics, maintenance). In practice, each of these applications will develop the basic technology in different directions. Depending on the applications we will find largely different descriptions of the technology involved. Fig. 2 offers a Value Chain-like representation of Industry 4.0, showing the wide range of applications of constituent technologies.

An interesting consequence of this state of affairs is that there is disagreement also at the higher level of government documents describing Industry 4.0 as the main object for innovation and industrial policies: when describing the main components of Industry 4.0 the French government uses 47 technologies, against 39 technologies for the Italian government.

Summing up, the recombinant nature of Industry 4.0 creates several

Table 2
Technical document and Scientific articles used to create the Technology Seed List according to the workflow in Fig. 3.

Author(s)	Name/title of the source	Pages	Country	Year	Publisher	Typology
France Government	New Industrial France – Building France's industrial future	112	France	2016	French Government	Technical Document
Geissbauer et al.	Industry 4.0: Building the digital enterprise – Global Industry Survey	36	United Kingdom	2016	Price Watehouse & Coopers	Technical Document
German Trade & Invest	Industrie 4.0 – Smart manufacturing for the future	40	Germany	2014	German Trade & Invest	Technical Document
Heng, S.	Industry 4.0 – Upgrading of Germany's industrial capabilities on the horizon	16	Germany	2014	Deutsche Bank	Technical Document
National Intelligent Factories Cluster	Research andInnovation Roadmap	88	Italy	2015	National Intelligent Factories Cluster	Technical Document
Rüßmann et al.	Industry 4.0 – The future of productivity and growth in manufacturing industries	20	United Sates	2015	Boston Consulting Group	Technical Document
Siemens	On the Way to Industrie 4.0–The Digital Enterprise	27	Germany	2015	Siemens	Technical Document
Smit et al.	Industry 4.0 – Study for the ITRE Committee.	94	Europe	2016	European Parliament	Technical Document
The Government Office for Science	The future of manufacturing: a new era of opportunity and challenge for the UK	250	United Kingdom	2016	The Government Office for Science	Technical Document
Wee et al.	Industry 4.0 – How to navigate digitization of the manufacturing sector	62	United States	2015	McKinsey Company	Technical Document
Gorecky et al.	Human-machine-interaction in the industry 4.0 era	6	Germany	2014	Industrial Informatics (INDIN) − 2014 12th IEEE International Conference	Scientific Article
Hermann et al.	Design principles for industrie 4.0 scenarios	10	Germany	2016	System Sciences (HICSS) − 2016 49th Hawaii International Conference	Scientific Article
Jadzi N.	Cyber physical systems in the context of Industry 4.0	3	Germany	2014	Automation, Quality and Testing, Robotics − 2014 IEEE International Conference	Scientific Article
Lasi et al.	Industry 4.0	4	Germany	2014	Business & Information Systems Engineering	Scientific Article
Lee et al.	A cyber-physical systems architecture for industry 4.0-based manufacturing systems	7	United States	2015	Manufacturing Letters	Scientific Article
Lee et al.	Service innovation and smart analytics for industry 4.0 and big data environment	6	United States	2014	Procedia CIRP	Scientific Article
Monostori L.	Cyber-physical production systems: Roots, expectations and R&D challenges	5	Hungary	2014	Procedia CIRP	Scientific Article
Posada et al.	Visual computing as a key enabling technology for industry 4.0 and industrial internet	15	Spain	2015	IEEE Computer Graphics and Applications	Scientific Article
Shrouf et al.	Smart factories in Industry 4.0: A review of the concept and of energy management approached in production based on the Internet of Things paradigm	5	Italy	2014	Industrial Engineering and Engineering Management (IEEM) − 2014 IEEE International Conference	Scientific Article
Zhan et al.	Cloud computing resource scheduling and a survey of its evolutionary approaches	33	China	2015	ACM Computing Surveys (CSUR)	Scientific Article

Fig. 2. A Porter-like Value chain framework for Industry 4.0 *(courtesy of Towel Publishing)*.

interrelated problems for profiling and mapping:

(a) the number of constituent technologies is very large
(b) the description and performance of constituent technologies depend critically on the specific application, hence on the business function/company area affected
(c) the stakeholders are located in several organizational positions
(d) the technical progress is very fast, with many (even if not all) constituent technologies facing rapid changes in their nature and performance.

Faced with this situation, a traditional approach to profiling and mapping would require a massive effort of keyword definition. A number of experts would be recruited in order to offer a representation of the field from their disciplinary or industry perspective. Extensive domain knowledge would be mobilized in such a way to build up detailed yet comprehensive maps of technologies. Based on these maps, governments, statistical offices and international organizations would work for a few years in line in the effort to identify the trends of technologies, the most important actors, the shares of individual countries or regions in the global landscape. For sure, this is the approach underlying most of the documents illustrated in Table 2 and this the approach that will be pursued in the years to come.

We strongly suggest this approach does not deliver the expected result. Keyword-based approaches to emerging technologies are too dependent on subjective judgments of experts. Even when the experts involved are top class and disinterested (often the best researchers or industrialists), their vision is inevitably partial. Even more importantly, keyword-based representations cannot be updated with the same speed of technology. The set of keywords identified by experts becomes inevitably obsolete in a few months.

This paper leverages on publications and open source repositories to design, develop and test a methodology that provides delineation and clustering of technologies of the Industry 4.0 paradigm. The methodology is based on a dictionary concerning the enabling technologies for industry 4.0 with full definitions and connections between them. Given the fast growth and the uncertainty that characterizes industry 4.0 technologies, our methodology is designed to be a bottom-up and continuous evolving tool. The structure and measurements made on the tool refer to July 2017.

3. Mapping a complex emerging technology

3.1. Delineating emerging technologies

The first task for mapping a new technology is field delineation, or the definition of the perimeter of the field. Industry 4.0 is not a new technology, but a novel combination of partly existing, partly new technologies driven by the convergence of their trajectories. As a matter of fact, it is clearly an example of emerging technology [10], as it shares the features of rapid growth, technological uncertainty, and market uncertainty.

The delineation of new fields of science and technology is an issue addressed since the late ‘70s, after the pioneering period of bibliometrics. Field delineation is a necessary step when existing classifications do not offer timely, reliable or comprehensive coverage of a topics, for example of a new technology or a new technological field. Moving beyond existing classifications require undertaking a search which, in general, may follow a lexical approach, a citationist approach, or a mix between the two [11]; Kreuchauff and Korzinov, 2017. In all cases there is a need to initialize the process, i.e. to identify a set of elements that constitute the starting point for searching.

The main approach has been based on keywords, to be identified in various regions of documents (title, abstract, keywords, full text of an article; title, abstract, claims, full text of a patent) and to be used as queries. A query is a structured sequence of words, connected by logical elements, such as “or”, “and” and the like, to be launched on a database in order to build up profiling, indexing, or clustering a given field. In general the initial set of keywords are provided by experts in the field, usually organized in an expert panel.

There are several limitations of the keyword/expert approach. First, expert based keyword definition, or patent classification is a very

expensive activity [12].

Second, the keyword selection is based on subjective judgment, and when experts are asked to decide on relatedness measures (e.g. synonims, hypernims or hyponims), they do not apply systematic rules [12,13]. Experts may be subject to a number of biases, such as for example the desirability bias (attributing higher probability of occurrence to preferred events) and many others [5,14,15]. Panels of experts are not immune by biases, such as group thinking. There is little research on the impact of expert subjective judgments on the delineation of emerging fields, but there is reason to believe it may be significant.

Third, the delineation of perimeter of emerging technologies is not robust to slight differences in the queries. As it has been shown by Zitt and Bassecoulard, little differences in the wording of queries, or on the time window, may end up in completely different sets of documents (Bassecoulard and Zitt, 2006; [16]. Therefore there is no proof that the method is reliable.

Finally, and more problematic for the case of Industry 4.0, the methodology is static, as it is based on a fixed sets of words. This set can (and in practice often is) updated, but this introduces a delay in the process and does not deliver reliable results. Keeping updated a collection of keywords in a dynamic technological landscape is extremely difficult.

These limitations have become evident in the last two decades, after the efforts of many authors to produce reliable perimeters of the emerging field of Nanotechnology. The initial efforts have been based on a classical expert-based approach. Panel of experts provided lists of keywords that were transformed into database queries. Among them, a consulting company called Lux, the Fraunhofer ISI in Germany, and CWTS in the Netherlands were the most active. Most studies delivered largely different delineations (Bassecoulard and Zitt, 2006; Mogoutov and Kahane, 2007; [17,18,19].

In turn, these limitations opened the way to massive efforts to reduce the dependence on experts and exploit systematically the new opportunities opened by text mining, following what has been called "full text based scientometrics" [20]. Starting from the late '90 s several attempts have been made to apply text mining techniques to the patent corpus and the field is currently burgeoning [12,21]; Kreuchauff and Korzinov, 2017; [19].

Within the text mining field applied to patents a well developed branch of studies makes use of lexical and/or synctactic structures that are derived from substantive engineering knowledge. Structures derived from theories of engineering invention, in particular TRIZ and Functional Analysis, are used in order to classify documents or to train algorithms for text mining. Based on an extensive processing of the text of patents, the TRIZ methodology has identified a number of evolutionary trends in technology [22], whose description in synctactic terms can be used to identify technology trends [23,24,25] or predict the future use of innovative products [26].

A more recent approach makes use of synctactic structures that reflect the functional aspects of invention, that is, the actions performed by a subject upon an object. This approach is rooted in a venerated tradition in Engineering design theory, called Functional analysis, which has become more operational through its combination with text mining techniques [27,28].

Large part of the recent literature deals with SAO structures, or subject-action-object linguistic patterns that are derived automatically from the full text of patents. These triplets of words describe the abstract functioning of technologies, in which a subject (device, mechanism, artifact) carries out an activity on a given object, producing a physical effect. These expressions are meant to represent formally the content of inventions in functional terms. Similarity matrices based on SAO sequences are then computed using various measures of similarity.

This approach has the advantage of being replicable and scalable and has been applied repeatedly in recent years [29,30,31,32,33,34]. A limitation of the SAO approach is that the identification of synctactic structures is not rooted in a substantive engineering-based knowledge that is made transparent to the reader. The synctactic notion of "action" may in fact cover either actions with a functional meaning, or actions that have no relation with the functional content of the invention. In addition, this literature has not yet provided a formal proof of completeness of the collection of SAO structures extracted from patents.

A complementary approach has been proposed based on an explicit knowledge base called Functional dictionary, which allows the exploration of the engineering content of the description of functions, while still being manageable for computer processing [35,5]. Robust engineering knowledge is needed to test, validate and cross-validate functional dictionaries

Summing up, the existing approaches to the delineation of emerging technologies, taken together, suffer from the following limitations:

i) dependence on expert judgment
ii) lack of robustness to alternative definitions of the perimeter
iii) delay in update of technologies (static approach)
iv) lack of transparency in the modeling of technology.

We now turn to the second main task in text mining of technology documents, that is, clustering.

3.2. Clustering complex emerging technologies

Once the perimeter of the field is delineated, a task usually included in the mapping is clustering, or the creation of groups of entities in such a way to reduce the complexity of the representation. Within text mining the clustering of documents is based on various kinds of linkages that are considered a signal of similarity in topics [36]. In general, linkages among documents can be generated by citations (citation-based clustering) or by the extraction of features in texts (text-based clustering) [37,38,39].

The most used approaches to clustering assume that members of a cluster:

- cite each other (*citation analysis*: [39,40,41,42];
- share certain words (*co-word analysis*: [43,44,45,46,47,48];
- share a reference in their bibliography (*bibliographic coupling*: [49,50];
- share the same sub-fields in a classification (*co-subfield analysis*: [51];
- are cited by the same documents (*co-citation analysis*: [52,11,53];
- are cited by the same authors (*author co-citation analysis*: [54].

In this paper we exploit the hyperlink feature of Wikipedia in order to introduce a new approach to clustering. Hyperlinks are introduced by authors in order to establish a semantic linkage between the page and other pages. We exploit this feature as follows: members of the same cluster are those pages that share the hyperlinks to other pages, according to thresholds defined by an appropriate algorithm.

Note that hyperlinks are only superficially similar to citations. The origin page "cites" another page by hyperlinking it, but in effect this linkage is not a citation (that is, a reference to a previous work) but a signal of semantic similarity, intended to guide the reader in the network of meanings.

We suggest that the hyperlinks are similar to citations under some respect, but very different under some other respects. Like citations, hyperlinks are introduced in the text by authors and reflect intentionality. Unlike citations, they reflect semantic relations, not relations of credit assignment or tribute to scientific authority. Perhaps more importantly, citations are introduced only by the author(s) of a paper and remain unchanged after publication. Hyperlinks, on the contrary, are introduced also by subsequent readers of the Wikipedia page. If the introduction of hyperlinks is not considered appropriate by the community of contributors, as it may happen due to vandalism, they are immediately removed (see the discussion below). This means that

they reflect all possible semantic connections among the pages, as collectively stated by a large community of authors, in a reliable and robust way. We believe this methodology offers a remarkable improvement with respect to existing approaches.

3.3. On the use of wikipedia in technology delineation and clustering

With respect to the literature on field delineation and clustering applied to science and technology we innovate by introducing a new source of knowledge- Wikipedia. We start with a small number of documents following a procedure which is expert-independent, in order to minimize the distortions from subjective judgment. We then exploit the properties of Wikipedia in order to delineate the field and identifying the linkages between technologies. Before entering into the details of the methodology let us first comment on the role of Wikipedia as a source of knowledge for automatic text extraction, field delineation and clustering.

The use of Wikipedia as source of knowledge started more than a decade ago and has been validated repeatedly in a variety of text mining applications (text annotation, categorization, indexing, clustering, searching: [55]. In addition to the large and growing size in terms of number of articles, the structure of Wikipedia has a number of useful features that make it a good candidate for text mining applications.

First, Wikipedia pages are considered reliable in many knowledge fields, including the ones we are mostly interested here, i.e. engineering and computer science Xu and Wu, 2014. The pages are regularly and systematically updated by a large global community of contributors, which includes many scientific and industrial authorities in the field. The use of Wikipedia as knowledge source for computerized text mining tools is established in the literature [56]. In addition, it is powerful in disambiguation of terms, particularly through the use of *redirect* pages and *disambiguation* pages. This means that it can be used for detection and disambiguation of named entities [57].

Second, the pages include links to other pages motivated by clear reasons on content. There are many links between Wikipedia pages, which are clues for semantic relations. This makes Wikipedia a densely connected structure, creating a classical *small world* effect: according to an often cited estimate, it takes on average 4.5 clicks to reach an article from any other article [58]. Unfortunately it is not possible to disentangle the kind of semantic relation, introducing a distinction between equivalent relations (synonymy), hierarchical relations (hyponymy/hyperonimy) and associative relations Milne, 2006, but this limitation is not relevant for our application.

Third, it makes use of categories which do not have a hierarchical structure, but a tree-like structure: categories “form a directed acyclic graph, allowing multiple categorization schemes to co-exist simultaneously” p. 11).

Fourth, it has the ability to evolve quickly [59], particularly after the development of systems such as Wikify (wikification: [60,61]. Wikipedia has by design a dynamic structure, since it is constantly growing in the number of entries and changing in their content, when this is needed due to the advancements of knowledge [62]. Furthermore the new terms that appear on Wikipedia thanks to comprehensive contributions by volunteers around the world, cannot be found in other linguistic corpora, such as WordNet Miller, 1995. Indeed, Wikipedia is the expression of a large international community, that is, of a “real community agreement” [63] or “community consensus” [64], guaranteed by permanent collective monitoring of the quality and rigour of the entries [65].

Finally, Wikipedia is free-content and multilingual. This make it possible to freely collect the information contained in the web pages and allows the possibility for future developments of the dictionary in other languages. In our opinion multilanguage is an interesting feature for the dictionary, due to the fact that Industry 4.0 is a worldwide phenomena.

These properties make Wikipedia the ideal candidate for our goal of building up a profile of Industry 4.0 that is comprehensive, dynamically updated, and, as far as possible, expert-independent.

In particular, Wikipedia entries allow an endogenous measurement of semantic relatedness. This is an exceedingly important property for our goal: we define a technology as included in the perimeter of Industry 4.0 if and only if it exhibits relatedness with other technologies already included in the perimeter. The inclusion of new technologies is therefore not dependent on experts’ subjective views, but is endogenously generated by the technological community that writes the articles for the encyclopedia and includes hyperlinks in the text of newly added pages.

3.4. Entering in more technical details, how is similarity measured by using Wikipedia?

In general, semantic relatedness is a measure of the similarity between two terms. It can be computed by statistical methods without requiring a manually encoded taxonomy, for example by analyzing term co-occurrence in a large corpus [66,67]. Wikipedia has been largely exploited in the literature in order to compute semantic relatedness. Gabrilovich and Markovitch [68] developed an alternative to Latent Semantic Analysis and called this new technique Explicit Semantic Analysis (ESA). This methodology first uses a classifier that is centroid-based to map input text to a vector of weighted Wikipedia articles. Then the vectors are exploited to obtain the semantic relatedness between two terms by computing the cosine similarity. This technique could be applicable to individual words, phrases or even entire documents. Furthermore, the mapping developed in this work has been successfully utilized for documents categorization.

A new version of this kind of systems was presented by Milne [69]. While in Gabrilovich and Markovitch [68] the authors use the full text of Wikipedia articles to establish relatedness between terms, in this work only the internal hyperlinks are exploited. To compute the relatedness between two terms, they are first mapped to corresponding Wikipedia articles and then vectors are created containing the links to other Wikipedia articles that occur in these articles.

The main problem facing semantic relatedness using Wikipedia is the disambiguation of terms. Several strategies have been developed to solve this problem. A first approach, described in Strube and Ponzetto [70], exploits the order in which entries occur in the disambiguation pages of Wikipedia to find the most likely correct meaning. On the other hand, Gabrilovich and Markovitch [68] avoids disambiguation entirely by simultaneously associating a term with *several* Wikipedia articles. Milne [69] approach hinges upon correct mapping of terms to Wikipedia articles. However, when terms are manually disambiguated, it has been shown that the systems of semantic relatedness computation are more accurate than the systems of automatic disambiguation [71].

Summing up, a consistent literature in the field of computational linguistics and text mining supports the notion that the use of Wikipedia articles as a knowledge base is justified and promising.

4. Methodology

In this section we give evidence of the methodological steps undertaken to build up the enriched dictionary referred to Industry 4.0. The dictionary contains technologies related to the Industry 4.0 paradigm, each of which is associated to the full set of relations with other technologies. The dictionary is semi-automatically generated using technical documents and the Wikipedia free online encyclopedia. The linkages between technologies have all a semantic content, since they are generated within the text of the articles of Wikipedia when a related topic or entry is considered necessary for the logical flow of the definition or description.

Fig. 3 shows the methodological steps needed to generate the enriched dictionary. In the flow diagram three elements are graphically

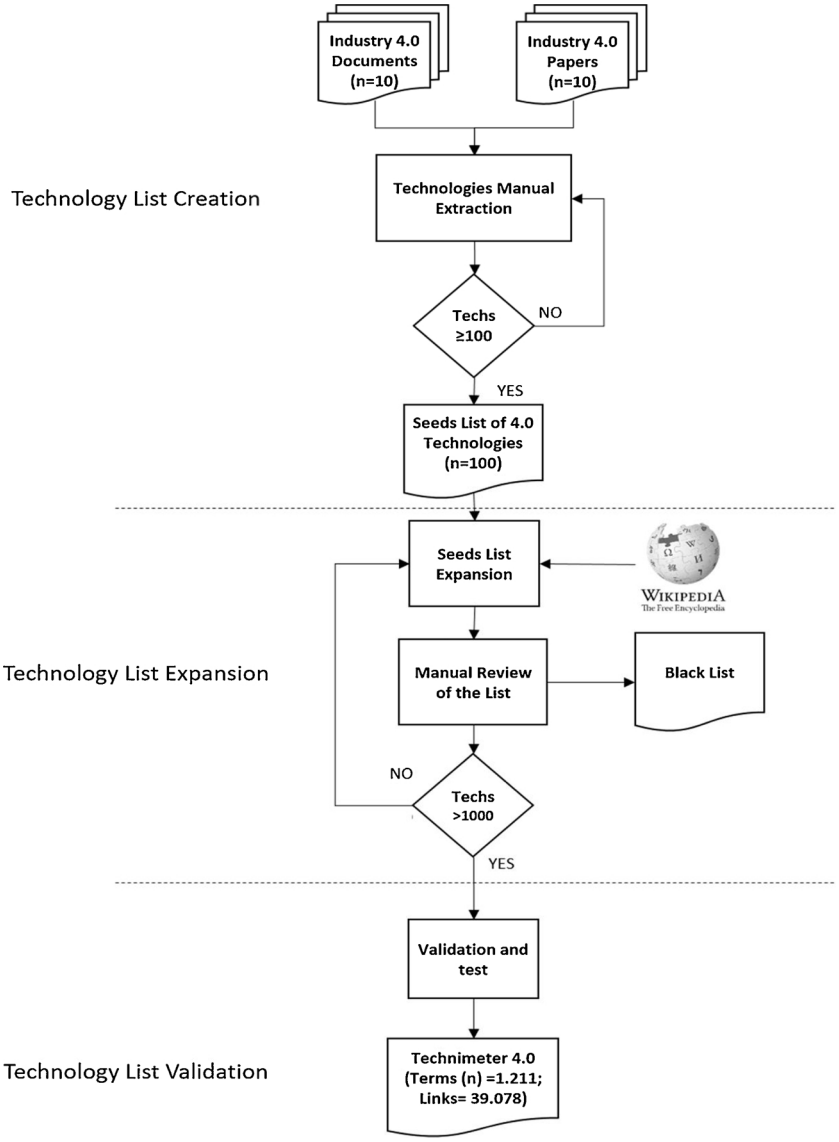

Fig. 3. Flow diagram of the adopeted methodology.

displayed: activities (rectangular shape), check points (diamond shape) and documents created from the procedure (sheet of paper shape).

4.1. Generation of the seed list

As input for our methodology we used technical documents, official government documents and the most cited academic papers in the field of industry 4.0. The selection of seed documents has been made by:

- taking the official government documents of the US government and of three large European countries (France, Germany, UK), all of which are strongly committed to the support of Industry 4.0 and are widely considered a benchmark in international documents (for example, OECD and European Union), plus a selection of technical documents cited in government papers, for a total of 10 documents;
- taking the 10 most cited papers on Industry 4.0 according to Scopus. The reference to Industry 4.0 was explicit in the title, abstract and keywords of the papers. The extraction was made on Scopus Database in June 2017.

We deliberately limit ourselves to a small number of documents. The reason is explained in conjunction with the chosen stopping rules, described in Section 4.2.1.

The documents have been manually parsed by a team of Master students in Engineering Management at the University of Pisa. The assignment was “Understand each document and extract the enabling technologies for industry 4.0”. For a definition of enabling technologies for Industry 4.0 see Section 2. The manual search in the documents continued until the team reached the goal of 100 different technologies extracted. For our purposes 100 different technologies represents a reasonable seed list of technologies, which will be used as input for the automatic expansion phase. The reason why we chose these stopping rules is explained in Section 4.2.1.

Table 3 shows a random sample of the 100 technologies included in the input seed list, with a preliminary classification (which however was not included in the assignment).

The table shows an intriguing result in the case of Additive Manufacturing and 3D Printing technologies. While, from a technological point of view, the former concept includes the latter, we found a list of documents and papers that mention 3D Printing without explicitly citing Additive Manufacturing. Moreover, 3D Printing is mentioned more times than Additive Manufacturing. This disturbing and irrational asymmetry confirms that different vocabularies are used even when referring to the same entity, or technological field.

Table 3
List of seed documents and examples of technologies found in the seed documents.

Technology	Technical Documents	Scientific articles
Additive manufacturing	[72–74,2,75–77]	[3]
3D printing	[72,73,74] German Trade & Invest (2014), [78,2,76,79],	[3,80]
Virtual reality	[72–74,77,2]	[81]
Augmented reality	[73,74], German Trade & Invest (2014), [76,77,79]	[81,3,80]
Sensor	[72–74] German Trade & Invest (2014), [78,2,75–77,79],	[81–83,3,84–86,80,87]

Furthermore, Table 3 underlines the relevance of the term Sensor within the Industry 4.0 paradigm: indeed sensors are mentioned in almost all documents and papers used to create the seed list.

4.2. Seed list expansion

For each term in the seed list the corresponding page in Wikipedia was found. All technologies identified in the seed were covered by Wikipedia. This is a preliminary confirmation of its relevance as a source of knowledge. These pages formed the initial glossary.

The expansion procedure automatically retrieved the pages and identified all hyperlinks included in the description of the technology.

The pages that are the target of hyperlinks are classified manually according to the following categorization:

a) links to pages already in the seed: these pages are labeled “anchors”, since they provide robust indicators of technologies that are, at the same time, mentioned explicitly in the seed documents and referred to in Wikipedia pages that deal with other technologies;
b) links to pages not in the seed: these are labeled “missing technologies” and are stored in memory for later treatment as potential candidates to inclusion in the dictionary;
c) links to pages with non-technological content: they are labeled “stopwords” and are eliminated from the procedure.

The overall procedure is iterated up to the point in which a number of at least 1000 different technologies is reached. At this point the procedure will stop. The reason why we chose these stopping rules is explained in Section 4.2.1. Updates and changes of the dictionary can originate from new entries (new technologies) or from updates to existing pages.

Given that the automatic extraction of Wikipedia pages can be run on a permanent basis, each version of the dictionary has a date. The current version, illustrated in the rest of the paper, is at the time point of July 15, 2017.

4.2.1. The stopping rule for the manual and the automatic expansion

The three stopping rules follow a sequential logic of order of magnitude: we start with approximately 10^1 documents, from which we extract 10^2 names of technologies, that, used as inputs to Wikipedia, deliver approximately 10^3 final technologies. More in detail we make use of:

a) 20 input documents: all technical documents taken as reference for Industry 4.0 share the same framework (i.e. DIN:SPEC 91345:2106). Many of the documents contains the same informations/terms/technologies. Furthermore these documents are technology focused and therefore from a low number of documents we obtain a large number of technologies.
b) seed list of 100 technologies: as demonstrated in the past [35] Wikipedia is a good source also for technical terms, therefore we assumed it was able to quickly expand the technologies related to Industry 4.0. For this reason the seed list is composed of no more than 10^2 entries.
c) output list of at least 1000 technologies: If the automatic expansion works correctly the technologies should increase by an order of magnitude in few iterations. Since the list has to be revised manually we decided for 10^3 entries as a target thus reducing the impact of manual review. As a matter of fact, the Wikipedia network of semantic linkages delivers a total number of technologies related to Industry 4.0 which exceeds this target, again confirming the validity of Wikipedia as a source of knowledge.

4.3. Structure of the enriched dictionary

The enriched dictionary can be defined as a set of enabling technologies for industry 4.0, associated to their definitions and to the linkages between them. The digital version of the tool is a hyperlinked text, available at: www.industria40senzaslogan.it.

For the purpose of publication in an academic paper the tool can be represented as a table in which we have:

Column 1- *Technologies*: Enabling technologies for industry 4.0, or a broad categorisation of technologies following a clustering procedure (see below)

Column 2- *Url*: Links of the Wikipedia pages of the enabling technologies for industry 4.0

Column 3- *Definitions*: Glossary, snippets from wikipedia page of the definition of the enabling technologies for industry 4.0

Column 4 *Links*: hyperlinks to other wikipedia pages from the wikipedia pages of enabling technologies for industry 4.0

Column 5- *Anchors*: hyperlinks to other wikipedia pages that are enabling technologies for industry 4.0 from the wikipedia pages of enabling technologies for industry 4.

Table 4 shows a sample of the dictionary for four enabling technologies. An evidence is that there are conflicting definitions of “Augmented reality” (AR) and “Virtual reality” (VR); the first says that AR contrasts VR, while the second states that “AR systems may also be considered a form of VR”. This is an evidence of the ambiguity that exists in the definition of 4.0 technologies. Moreover, the table underlines how words such as “3D printing” and “Additive manufacturing” that are used in different ways within papers and technical documents (see Table 3), basically refer to the same concept.

Table 4 also shows the difference between links and anchors. Links include all hyperlinks found in Wikipedia pages. By nature, a certain fraction of these links contain information that is not relevant to our task. It is likely that, in the course of discussion of a given topics, authors quote an author (e.g. Bob Sproull in Virtual Reality), an institution (e.g. British Museum or California Institute of Technology) or an event or application (e.g. Coachella Valley Music and Arts Festival). These links do not add to our knowledge of the Industry 4.0 field. It can be seen that, following the definition of anchor given above, these terms are eliminated in column (E), which includes only the anchors, or those entries that are added to the body of knowledge.

5. Main results

The dictionary is composed of 1.211 terms and 39.078 relationships between them. This generates a graph in which the node represents a technology and the edge represents a link in the Wikipedia page.

The network structure naturally gives origin to graph-theoretic metrics. We exploit this property in order to generate a number of indicators that the readers may find it useful to examine.

5.1. Graph analysis and Sub-graph selection

Fig. 4 gives an overview of the obtained. We compute for each node the in-degree (horizontal axis), the out-degree (vertical axis) and the

Table 4
Sample of enabling technologies showing the table structure of dictionary.

Enabling technology (A)	Wikipedia Page (B)	Definition (sample) (C)	Links (sample) (D)	Anchors (sample) (E)
3D printing	en.wikipedia.org/wiki/3D printing	3D printing, also known as additive manufacturing (AM), refers to processes used to create a three-dimensional object in which layers of material are formed under computer control to create an object. […]	3D bioprinting; Actuator; Artificial brain; Artificial muscle; Bikini; Biotechnology; Blue Brain Project; CAD; Delivery drone; Forbes; Gene therapy; Injection moulding; Inkjet; Laser-powered; Phosphor display;Magnetic refrigeration	Actuator; Artificial intelligence; Artificial photosynthesis; Atomtronics; Biotechnology; CNC; Computer-aided design; DMOZ; Electron beam melting; Home automation; Number; Internet of Things
Additive Manufactu Ring	en.wikipedia.org/wiki/**3D printing**	*(same as 3D printing)*	*(same as 3D printing)*	*(same as 3D printing)*
Augmented Reality	en.wikipedia.org/wiki/Augmented_reality	Augmented reality (AR) is a live direct or indirect view of a physical, real-world environment whose elements are augmented (or supplemented) by computer-generated sensory input such as sound, video, graphics or GPS data. It is related to a more general concept called computer-mediated reality, in which a view of reality is modified (possibly even diminished rather than augmented) by a computer. […]	360° video; ARQuake; ARTag; ARToolKit; Ableton Live Accelerometer Acrobatics; Adobe Flash; Alexis Ohanian AlloSphere; Alternate reality game; American football; Android (operating system); Artificial reality; Audient; Augment (app) Augmented Reality Markup Language Augmented browsing; Augmented reality-based testing; Augmented virtuality Augmented web; Australia Council for the Arts; Automotive head-up display; Automotive navigation system; BBC; Bionic contact lens Blair MacIntyre; Blippar; Blob detection; Brain in a vat; Bruce H. Thomas; […]	Android (operating system); Bionic contact lens; Computer vision; Cyborg; DMOZ; EyeTap; GPS; Gesture recognition; Global Positioning System; Google Glass; Graphical user interface; Holography; IOS; MEMS; Mobile computing; QR code; RFID; Real-time computing Smartphone; Speech recognition; Tablet computer; Ubiquitous computing; Ultrasound; Virtual Reality; Virtual retinal display XML
Virtual Reality	en.wikipedia.org/wiki/Virtual_reality	Virtual reality (VR) typically refers to computer technologies that use virtual reality headsets, sometimes in combination with physical spaces or multi-projected environments, to generate realistic images, sounds and other sensations that simulates a user's physical presence in a virtual or imaginary environment. […]	360° video; 3D audio effect; 3D computer graphics; A-ha; ACM Computing Classification System ARToolKit; Air force Algorithm; Algorithm design; AlloSphere; Alpenexpress; Alton Towers; Amnesty International Amusement arcade; Analysis of algorithms; Anshe Chung; Antonin Artaud; Anxiety disorder; Apple Inc.; Application security; Arcade game; [.]	3D computer graphics; ACM Computing Classification System Algorithm; Artificial intelligence Augmented reality; Computer; Computer-aided design Computer network; Computer security Computer vision; Computing platfor; Cryptography; Cyberspace; Data mining; […]

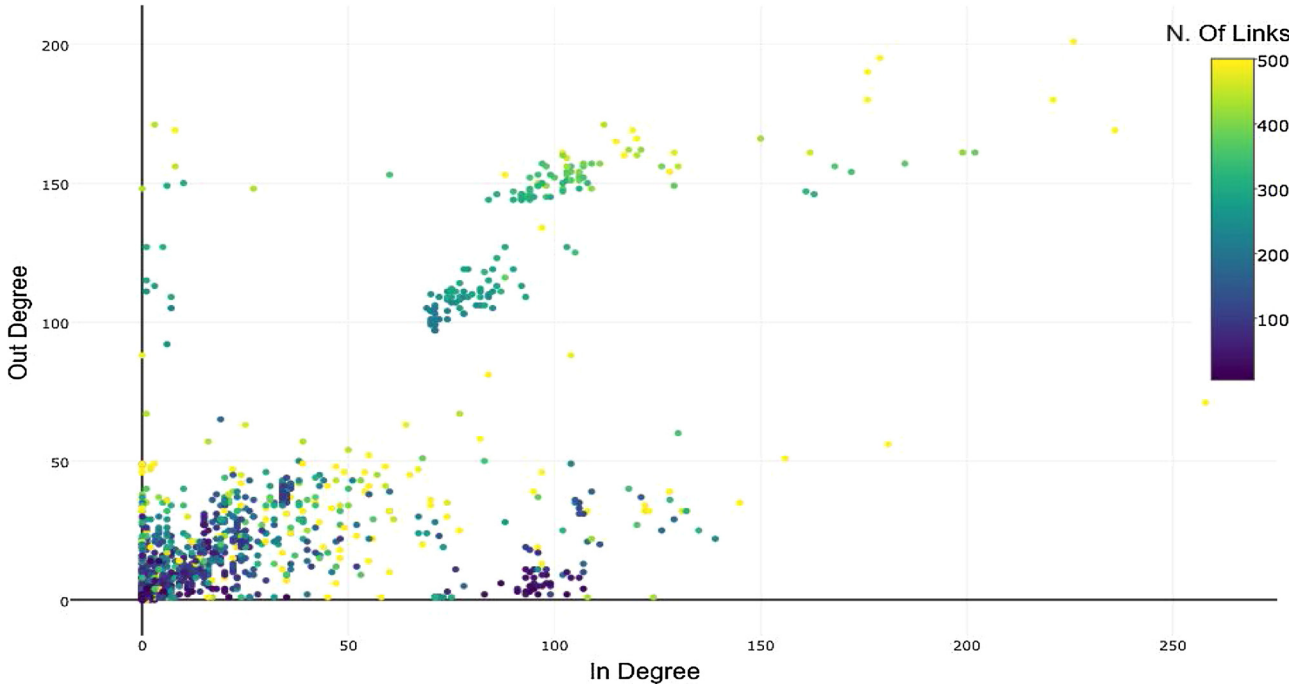

Fig. 4. Plot of the in-degree and the out-degree of the nodes of the graph. The colour of the node represents the number of Wikipedia internal links.

number of links to other Wikipedia pages that each node has (color of the node). Our results show that in terms of the analyzed variables we can identify 4 different clusters of nodes (or technologies) each one having a different behaviour.

The first group we take in consideration are the point having an in-degree greater than 70 and an out-degree greater than 70. In this group on the top right of the page we have some outliers. These are terms like *Microprocessor* and *Microcontroller*, *X86*, *8-bit*, *16-bit* and *32-bit*.

Then we observe a sub-group centered on the coordinates (150, 100). Here we have terms like *Program counter, Adressing mode, Instruction Cycle, Coprocessor, Symultaneus multitrading, SISD and MISD*. Still within the first cluster we have another sub-group of terms centered in (80,110) with terms like *Intel i860, Intel Atom, Intel 80286, Intel 80188* but also *Bloomfiled* and *Wolfdale*.

The second cluster is one in which the points have an in-degree greater than 70 and an out-degree smaller than 70. Here we have terms like *Machine learning, Artificial Neural Network, Cognitive Computing, Software, Random Access Memory, Internet, Firmware* and *C++*.

The third group collects points having an in-degree lower than 70 and an out-degree greater than 70. Here we have terms like *Processors, Micro-Operation, Micro-Assembler, Application specific integrated circuit*.

The last and most populated cluster has an in-degree smaller than 70 and an out degree smaller than 70.

This cluster is more precisely visualized in Fig. 5, for which we have as before the in-degree (horizontal axis), the out-degree (vertical axis) and the number of links to other Wikipedia pages that each node has (colour of the node). A jitter process has been implemented to the points on the graph in order to better visualize the overlapped points. In this plot only the technologies having both an in-degree and an out-degree lower than 70 are shown. This generates a subgraph composed by 931 nodes and 10673 edges.

5.2. Graph representation and cluster analysis

The structure of the graph offers it self naturally to clustering of technologies in order to obtain a readable mapping.

The clustering algorithm receives as input the collection of technology terms T of the analyzed subgraph and returns a set of terms clusters C = {C1, C2,.., Cn} that cover the whole subgraph in analysis. Each cluster Ci is a subset of terms of T, and a term may belong to only one cluster.

In Fig. 5 we show a representation of the sub-graph made using Gephi software with the Force Atlas algorithm. In this representation, two nodes in the graph are represented closely if they share an edge. In this way also nodes that belongs to the same communities of nodes (nodes that can be grouped into sets such that each set is densely connected internally) but do not share any edge are represented closely. In other words, the visualization tends to be coherent with the clustering algorithm. The size of the node is proportional to its in-degree while the colour express the cluster to which each node belong. Finally some of the labels of nodes and clusters are shown.

Each node is a technology and each edge represent a Wikipedia link between the pages. The size of the nodes is proportional to the in-degree, and the colours represent the clusters.

The algorithm we used to compute the modularity of each node and thus to assign a group to each of them is described in Blondel et al. [88]. The process resulted in 11 clusters. The content of each cluster is shown in Table 5, where for each cluster we can see the first 15 nodes in terms of in-degree (the most pointed terms).

Let us examine more in depth a cluster, such as *Identification* (cluster 11). The size and colour of the words are proportional to the in-degree of the nodes (Figs. 6 and 7).

6. Discussion and future developments

Faced with a complex, multitechnology and multistakeholder technology, there is a need for delineation and clustering techniques that are rooted in solid technical knowledge. At the same time, there is a need for comprehensive coverage and automatic knowledge update.

The enriched dictionary can be used for tagging and classification of documents in a variety of applications.

Among them we consider the following applications

a) mapping the field of Industry 4.0 by projecting the dictionary onto the corpus of publications and patents
b) comparing the field delineation obtained with the dictionary with

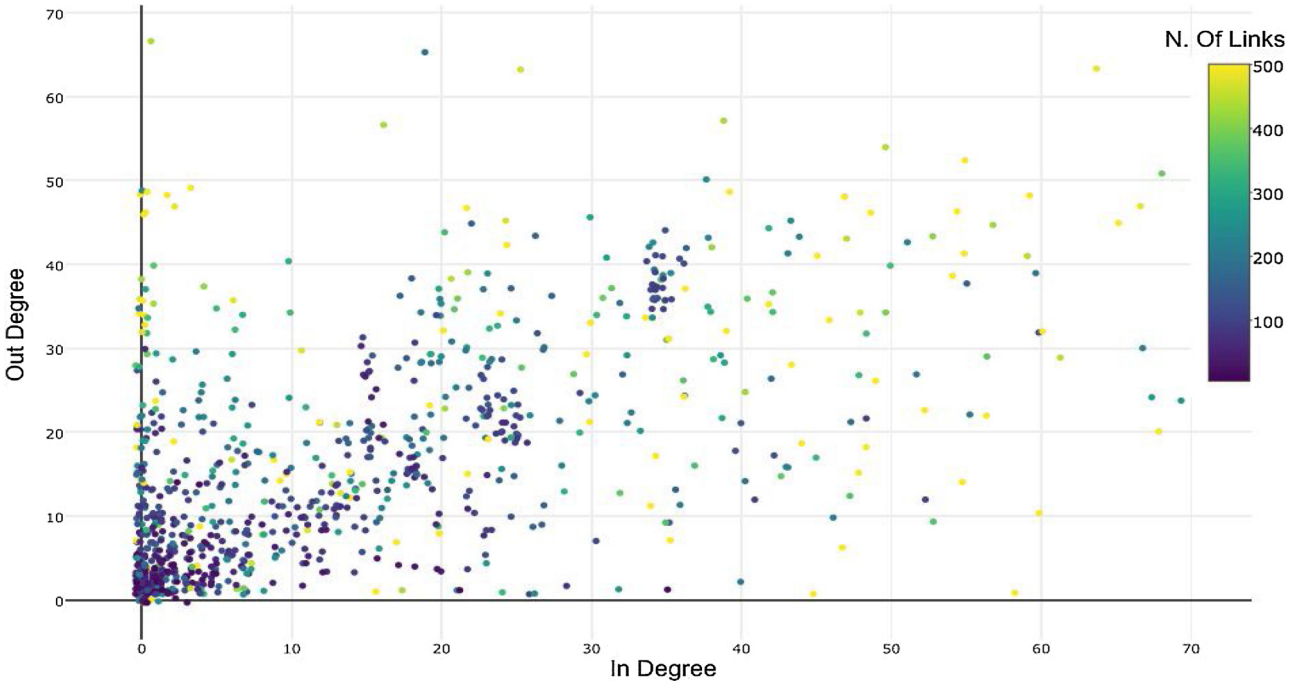

Fig. 5. Plot of the in-degree and the out-degree of the nodes of the sub-graph. Selection of the nodes having an in-degree and an out-degree lower than 70. The colour of the node represent the number of Wikipedia internal links.

those obtained with established methodologies (keyword search, evolutionary search, IPC or CPC classification)

c) mapping the field of Industry 4.0 by using the dictionary as a device for crawling the websites of companies in search of words and/or sentences that relate to the field, by applying measures of similarity

d) matching words and/or sentences in the abstract or full text of research proposals submitted to funding agencies, ministries or the European Commission, in order to evaluate the similarity to the Industry 4.0 field and detect possible cases of re-labeling

e) matching words and/or sentences to the abstract, claim or full text of patents in order to support IP services to assess the similarity to Industry 4.0 in a way that does not depend on existing (or future) classifications and allows for lateral technological vision

f) clustering technologies at various levels of resolution, identifying sub-classes

g) detecing technology trends by mapping the field over time and identifying changes in density of the hyperlink network that my be indicators of promising technology, or signals of technology

Table 5
Clustering of Industry 4.0 technologies. Clusters are manually labeled. The table also shows the top 15 technologies of the cluster in terms of the in-degree.

#	Label of the cluster	Constituent technologies
1	Big Data	Virtual machine, Data mining, User interface, Algorithm, Computer vision, Cryptography, Printed circuit board, Middleware, Real-time computing, Virtual reality, Augmented reality, Human–computer interaction, Multiprocessing, Decision support system, Supervised learning
2	Transactions, digital certification, digital currency	Bitcoin, Cryptocurrency, Bitcoin network, Cryptocurrency tumbler, Digital currency exchanger, Alternative currency, Dogecoin, Ethereum, Litecoin, Monero (cryptocurrency), Namecoin, Peercoin, Virtual currency, Auroracoin, Blockchain, Lisk, Primecoin, Ripple (payment protocol), Titcoin, Zerocoin
3	Program ming languages	Python (programming language), Database, Computing platform, Ruby (programming language), C Sharp (programming language), HTML, Perl, Hypertext Transfer Protocol, XML, Java (software platform), Haskell (programming language), .NET Framework, Lua (programming language), Sun Microsystems, BASIC
4	Computing	MacOS, IOS, Mainframe computer, Graphical user interface, Cloud computing, Home computer, Laptop, Solaris (operating system), Microcomputer, Personal digital assistant, QNX, Read-only memory, Tablet computer, ASCII, DOS
5	Embedded Systems	Programmable logic controller, Zilog Z80, CMOS, Zilog Z8, Toshiba TLCS, Zilog eZ80, NEC µPD780C, MOS Technology 6502, R800 (CPU), U880, Zilog Z180, Zilog Z800, Zilog Z8000, KP1858BM1, Hitachi HD64180
6	Intel	3D XPoint, Intel ADX, Intel Clear Video, Intel SHA extensions, Intel System Development Kit, Intel 1103, Intel AZ210, Intel Cluster Ready, Intel Compute Stick, Intel Display Power Saving Technology, Intel Mobile Communications, Intel Modular Server System, Intel PRO/Wireless, Intel Quick Sync Video
7	Internet of Things	Wireless sensor network, Near field communication, Arduino, NetSim, Z-Wave, OPNET, Telemetry, RIOT (operating system), Routing protocol, TinyOS, Internet of things, NesC, MiWi, Nano-RK, LinuxMCE
8	Protocols & Architec tures	1-Wire, Profibus, Smart meter, ×10 (industry standard), Modbus, Local Interconnect Network, TTEthernet, Fleet Management System, Keyword Protocol 2000, Meter-Bus, MTConnect, OPC Unified Architecture, PROFINET, RAPIEnet, SAE J1587
9	Communica tion Network and Infrastruc tures	Wi-Fi, Cellular network, Router (computing), Internet Protocol, Radiotelephone, ARPANET, Radio frequency, Digital subscriber line, General Packet Radio Service, Global Positioning System, CYCLADES, Beacon, Wireless, High Speed Packet Access, Evolution-Data Optimized
10	Production	Laser, 3D printing, Home automation, Agricultural robot, Nanorobotics, Semantic Web, Machine vision, Nanotechnology, Robotics, Information and communications technology, Speech recognition, Smart grid, Memristor, OLED, Computer-generated holography
11	Identification	Barcode, RFID, QR code, MaxiCode, Mobile tagging, Code 128, GS1 DataBar, High Capacity Color Barcode, Aztec Code, Barcode printer, Bokode, Codabar, CPC Binary Barcode, Interleaved 2 of 5, ITF-14

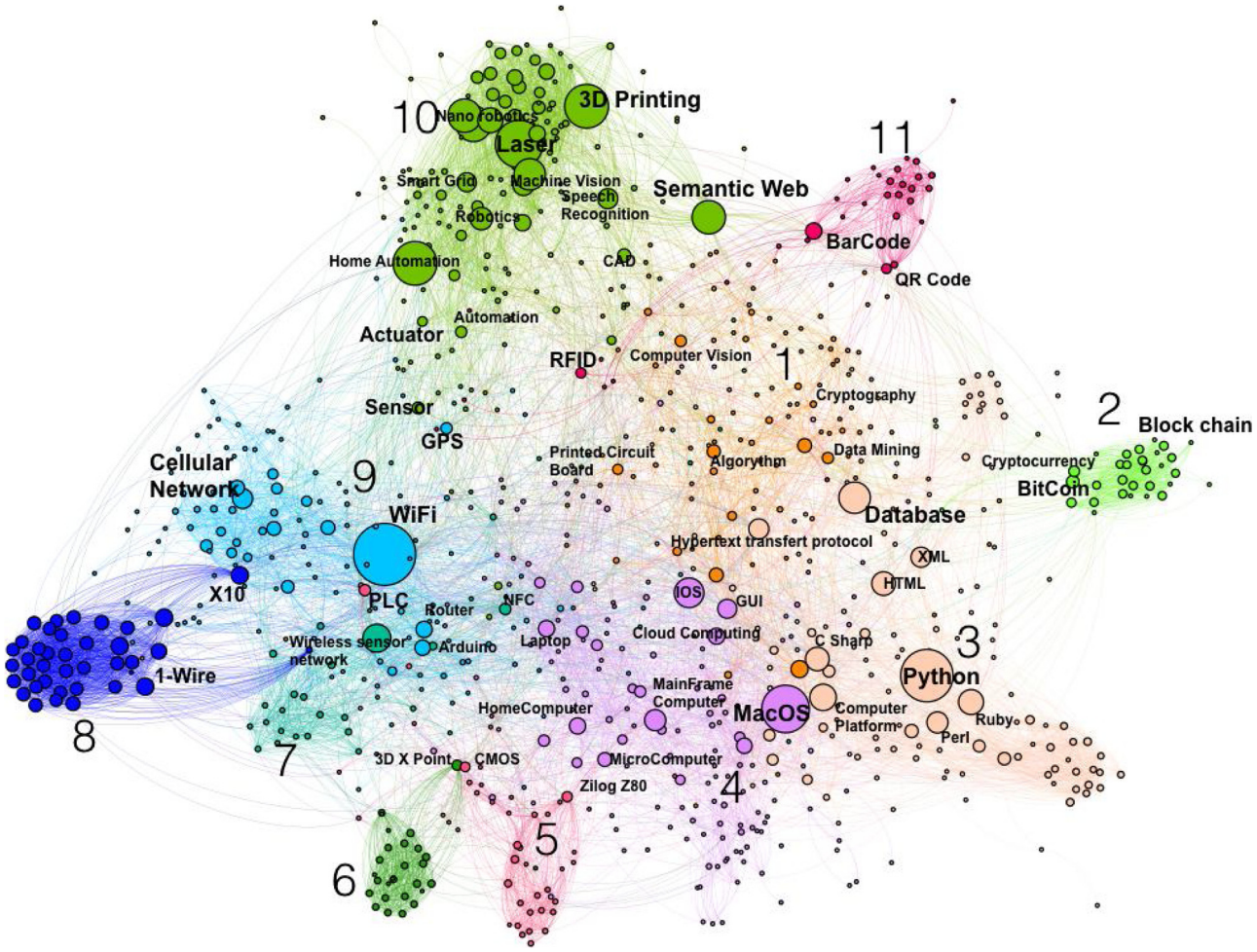

Fig. 6. Representation of the graph of 4.0 technologies and of the clusters in wich them are arranged.

Fig. 7. Wordcloud of the words belonging to the class 11 labelled as *Identification*. In this figure the size is proportional to the logarithm of the in-degree of each node that represents a word.

maturity or convergence

h) detecting vacuum regions in the hyperlink network that may be of interest for future entry and growth of technologies.

Most of these applications will require only incremental work with respect to the propodes methodology.

References

[1] A. Riel, C. Kreiner, G. Macher, R. Messnarz, Integrated Design for Tackling Safety and Security Challenges of Smart Products and Digital Manufacturing, CIRP Annals-Manufacturing Technology, 2017.

[2] J. Smit, S. Kreutzer, C. Moeller, M. Carlberg, Industry 4.0 – Study for the ITRE Committee, Policy Department A: Economic and Scientific Policy, European Parliament, Brussels, 2016 Available at http://www.europarl.europa.eu/RegData/etudes/STUD/2016/57000/IPOL_STU(2016)570007_EN.pdf (Accessed 15 September 15 2017).

[3] H. Lasi, P. Fettke, H.-G. Kemper, T.M. Feld Hoffmann, Industry 4.0, Bus. Inf. Syst. Eng. 6 (2) (2014) 239–242.

[4] A.J.C. Trappey, C.V. Trappey, U.H. Govindarajan, J.J. Sun, A.C. Chuang, A review of technology standards and patent portfolios for enabling cyber-physical systems in advanced manufacturing, Adv. Eng. Inf. 4 (2016).

[5] R. Apreda, A. Bonaccorsi, F. Dell'Orletta, G. Fantoni, Functional technology foresight. A novel methodology to identify emerging technologies, Eur. J. Futur. Res. (2016) 4–13.

[6] V. Roblek, M. Meško, A. Krapež, A complex view of industry 4.0, SAGE Open 6 (No.2) (2016) 1–12.

[7] National Research Council, Visionary Manufacturing Challenges for 2020, The National Academies Press, Washington DC, 1998.

[8] D. O'Halloran, E. Kvochko, Industrial internet of things: unleashing the potential of connected products and services, World Economic Forum, (2015) (Davos-Klosters, Switzerland).

[10] D. Rotolo, D. Hicks, B.R. Martin, What is an emerging technology? Res. Policy 44 (10) (2015) 1827–1843.

[11] H. Small, Tracking and predicting growth areas in science, Scientometrics 68 (3) (2006) 595–610.

[12] Y.H. Tseng, C.J. Lin, Y.I. Lin, Text mining techniques for patent analysis, Inform. Process. Manage. 43 (2007) 1216–1247.

[13] H. Noh, Y. Jo, S. Lee, Keyword selection and processing strategy for applying text mining to patent analysis, Expert Syst. Appl. 42 (2015) 4348–4360.

[14] A. Bonaccorsi, R. Apreda, G. Fantoni, Cognitive and Motivational Biases in Technology Foresight, (2017) (Under review).

[15] A. Bonaccorsi, G. Fantoni, R. Apreda, D. Gabelloni, Functional patent classification,

in: W. Glanzel, U. Schmoch, H.F. Moed (Eds.), Handbook of Science and Technology Indicators, Springer, Dordrecht, 2017.
[16] M. Zitt, E. Bassecoulard, Delineating complex scientific fields by an hybrid lexical-citation method: an application to nanosciences, Inform. Process. Manage. 42 (6) (2007) 1513–1531.
[17] A. Porter, J. Youtie, P. Shapira, Nanotechnology publications and citations by leading countries and blocs, J. Nanopart. Res. 10 (2008) 981–986.
[18] S. Ghazinoory, F. Amari, S. Farnoodi, An application of the text mining approch to select technology centers of excellence, Technol. Forecasting Social Change 80 (2013) 918–931.
[19] S. Ozcan, N. Islam, Patent information retrieval: approaching a method and analysing nanotechnology patent collaborations, Scientometrics 111 (2017) 941–970.
[20] K.W. Boyack, H. Small, R. Klavans, Improving the accuracy of co-citation clustering using full text, J. Ame. Soc. Inform. Sci. Technol. 64 (9) (2013) 1759–1767.
[21] J. Joung, K. Kim, Monitoring emerging technologies for technology planning using technical keyword base analysis from patent data, Technol. Forecasting Social Change 114 (2017) 281–292.
[22] V. Petrov, The laws of system evolution, TRIZ J. 3 (2002) 9–17. Available at https://triz-journal.com/laws-system-evolution/ (Accessed 15 September 15 2017).
[23] M.Y. Wang, D.S. Chang, C.H. Kao, Identifying technology trends for R&D planning using TRIZ and text mining, R&D Manage. 40 (2010) 491–509.
[24] J. Yoon, K. Kim, An automated method for identifying TRIZ trends from Patents, Expert Syst. Appl. 38 (12) (2011) 15540–15548.
[25] H. Park, J.J. Ree, K. Kim, Identification of promising patents for technology transfers using TRIZ evolution trends, Expert Syst. Appl. 40 (2013) 736–743.
[26] P.A. Verhaegen, J. D'hondt, J. Vertommen, S. Dewulf, J.R. Duflou, Relating properties and functions from patents to TRIZ trends, CIRP J. Manuf. Sci. Technol. 1 (2009) 126–130.
[27] G. Cascini, A. Fantechi, E. Spinicci, Natural language processing of patents and technical documentation, International Workshop on Document Analysis Systems DAS 2004, Dordrecht, Springer, Document Analysis Systems VI, 2004, pp. 508–520.
[28] G. Cascini, M. Zini, Measuring patent similarity by comparing inventions functional trees, Comput.-Aided Innov.(CAI) (2008) 31–42.
[29] J. Yoon, K. Kim, Identifying rapidly evolving technological trends for R&D planning using SAO-based semantic patent networks, Scientometrics 88 (2011) 213–228.
[30] J. Yoon, K. Kim, Detecting signals of new technological opportunities using semantic patent analysis and outlier detection, Scientometrics 90 (2012) 445–461.
[31] S. Choi, J. Yoon, K. Kim, J.Y. Lee, C.H. Kim, SAO network analysis of patents for technology trends identification: a case study of polymer electrolyte membrane technology in proton exchange membrane fuel cells, Scientometrics 88 (2011) 863–883.
[32] S. Choi, H. Park, D. Kang, J.Y. Lee, K. Kim, An SAO-based text mining approach to building a technology tree for technology planning, Expert Syst. Appl. 32 (13) (2012) 11443–11455.
[33] H. Park, J. Yoon, K. Kim, Identifying patent infringement using SAO based semantic technological similarities, Scientometrics 90 (2012) 515–529.
[34] X. Wang, P. Qiu, D. Zhu, L. Mitkova, M. Lei, A.L. Porter, Identification of technology development trends based on subject–action–object analysis: the case of dyesensitized solar cells, Technol. Forecasting Social Change 98 (2015) 24–46.
[35] G. Fantoni, R. Apreda, F. Dell'Orletta, M. Monge, Automatic extraction of function–behaviour–state information from patents, Adv. Eng. Inf. 27 (3) (2013) 317–334.
[36] H. You, M. Li, J. Jiang, B. Ge, X. Zhang, Evolution monitoring for innovation sources using patent cluster analysis, Scientometrics 111 (2017) 693–715.
[37] L. Leydesdorff, I. Hellsten, Measuring the meaning of words in contexts: an automated analysis of controversies about monarch butterflies, frankenfoods and stem cells, Scientometrics 67 (2) (2006) 231–258.
[38] S. Wang, R. Koopman, Clustering articles based on semantic similarity, Scientometrics 111 (2017) 1017–1031.
[39] A. Jaffe, M. Trajtenberg, R. Henderson, Geographic localization of knowledge Spillovers as evidenced by patent citations, Quart. J. Econ. 108 (3) (1993) 577–598.
[40] H.F. Moed, Citation Analysis in Research Evaluation, Springer, Dordrecht, 2005.
[41] B. Verspagen, Mapping technological trajectories as patent citation networks: a study on the history of fuel cell research, Adv. Complex Syst. 10 (1) (2007) 93–115.
[42] S. Lee, W. Kim, The knowledge network dynamics in a mobile ecosystem: a patent citation analysis, Scientometrics 111 (2017) 717–742.
[43] M. Callon, J.P. Courtial, W.A. Turner, S. Bauin, From translations to problematic networks–an introduction to co-word analysis, Social Sci. Inform. 22 (2) (1983) 191–235.
[44] A. Rip, J.P. Courtial, Co-word maps of biotechnology: an example of cognitive scientometrics, Scientometrics 6 (6) (1984) 381–400.
[45] L. Leydesdorff, Words and co-words as indicators of intellectual organization, Res. Policy 18 (4) (1989) 209–223.
[46] E.C. Engelsman, A.F.J. van Raan, A patent-based cartography of technology, Res. Policy 23 (1) (1992) 1–26.
[47] A.F.J. Van Raan, R.J.W. Tijssen, The neural net of neural network research: an exercise in bibliometric mapping, Scientometrics 26 (1993) 169–192.
[48] B. Yoon, Y. Park, A text-mining-based patent network: analytical tool for high-technology trend, J. High Technol. Manage. Res. 15 (2004) 37–50.
[49] W. Glanzel, H.J. Czerwon, A new methodological approach to bibliographic coupling and its application to the national, regional and institutional level, Scientometrics 37 (1996) 195–221.
[50] O. Kuusi, M. Meyer, Anticipating technological breakthroughs: using bibliographic coupling to explore the nanotubes paradigm, Scientometrics 70 (2007) 759–777.
[51] P.L. Chang, C.C. Wu, H.J. Leu, Using patent analyses to monitor the technological trends in an emerging field of technology: a case of carbon nanotube field emission display, Scientometrics 82 (2010) 5–19.
[52] H. Small, Co-citation in the scientific literature: a new measure of the relationship between two documents, J. Am. Soc. Inf. Sci. 24 (1973) 265–269.
[53] H. Small, E. Sweeney, Clustering the Science Citation Index using co-citations/1. A comparison of methods, Scientometrics 7 (3–6) (1985) 391–409.
[54] H.D. White, B.C. Griffith, Author co-citation: a literature measure of intellectual structure, J. Am. Soc. Inf. Sci. 32 (1981) 163–171.
[55] D. Milne, I.H. Witten, Learning to link with Wikipedia, Proceedings of the 17th ACM Conference on Conference on Information and Knowledge Management, (CIKM-08) (2008) 233–242.
[56] P. Ferragina, U. Scaiella, Fast and accurate annotation of short texts with Wikipedia pages, IEEE Software 29 (1) (2012) 70–75.
[57] R.C. Bunescu, M. Pasca, Using encyclopedic knowledge for named entity disambiguation, Proceedings of the 11th Conference of the European Chapter of the Association of Computational Linguistics (EACL-06) vol. 6, (2006) 9–16.
[58] S. Dolan, Six Degrees of Wikipedia, (2008) Available at http://mu.netsoc.ie/wiki/ (Accessed 15 September 15 2017).
[59] A. Lih, Wikipedia as participatory journalism: reliable sources? Metrics for evaluating collaborative media as a news resource, Proceedings of the 5th International Symposium on Online Journalism, Austin, Texas, 2004.
[60] R. Mihalcea, A. Csomai, Wikify! linking documents to encyclopedic knowledge, Proceedings of the 16th ACM Conference on Conference on Information and Knowledge Management, (CIKM-07) (2007) 233–242.
[61] X. Cheng, D. Roth, Relational inference for wikification, Proceedings of EMNLP-2013 (2013).
[62] S.P. Ponzetto, M. Strube, Knowledge derived from wikipedia for computing semantic relatedness, J. Artif. Intell. Res. 30 (2007) 181–212.
[63] C. Bizer, J. Lehmann, S. Kobilarov, C. Becker, R. Cyganiak, S. Hellmann, DBpedia. A crystallization point for the web of data, Web Semantics 7 (3) (2009) 154–165.
[64] M. Hepp, K. Siorpaes, D. Bachlechner, Harvesting wiki consensus. using wikipedia entries as vocabulary for knowledge management, IEEE Internet Computing, September–October, 2007, pp. 54–65.
[65] S. Bryant, A. Forte, A. Bruckman, Becoming wikipedian: transformation of participation in a collaborative online encyclopedia, Proceedings of the ACM GROUP'05 Conference, November 6–9, 2005, Sanibel Island, Florida, USA, 2005.
[66] P. Resnik, Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language, J. Artif. Intell. Res. 11 (1999) 95–130.
[67] J.J. Jiang, D.W. Conrath, Semantic similarity based on corpus statistics and lexical taxonomy, Proceedings of the 10th International Conference on Research in Computational Linguistics, ROCLING'97, Taiwan, 1997.
[68] G. Gabrilovich, S. Markovitch, Computing semantic relatedness using wikipedia-based explicit semantic analysis, In Proceedings of the 20th International Joint Conference on Artificial Intelligence, IJCAI'07, Hyderabad, India, January 2007, 2007, pp. 1606–1611.
[69] D. Milne, Computing semantic relatedness using wikipedia link structure, Proceedings of the New Zealand Computer Science Research Student Conference, NZ CSRSC'07, Hamilton, New Zealand, 2007.
[70] M. Strube, S.P. Ponzetto, WikiRelate! computing semantic relatedness using wikipedia, AAAI '06 (2006) 1419–1424.
[71] O. Medelyan, D. Milne, C. Legg, I.H. Witten, Mining meaning from wikipedia, Int. J. Hum. Comput. Interact. 67 (9) (2009) 716–754.
[72] The Government Office for Science, The Future of Manufacturing: a New Era of Opportunity and Challenge for the UK Project Report, The Government Office for Science, London, 2013https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/255922/13-809-future-manufacturing-project-report.pdf.
[73] National Intelligent Factories Cluster, Research and Innovation Roadmap, National Intelligent Factories Cluster, 2015 Available at http://www.fabbricaintelligente.it/wp-content/uploads/Booklet-Fabbrica-Intelligente-2015-PAGINE-SINGOLE.pdf (Accessed 15 September 2017).
[74] French Government, New Industrial France – Building France's Industrial Future. French Government, (2016) Available at https://www.economie.gouv.fr/files/files/PDF/web-dp-indus-ang.pdf (Accessed 15 September 2017).
[75] Siemens, On the Way to Industrie 4.0 – The Digital Enterprise, (2015) Available at https://www.siemens.com/press/pool/de/events/2015/digitalfactory/2015-04-hannovermesse/presentation-e.pdf (Accessed 15 September 2017).
[76] D. Wee, R. Kelly, J. Cattel, M. Breunig, Industry 4.0 ? How to Navigate Digitization of the Manufacturing Sector, McKinsey & Company, 2015 Available at https://www.mckinsey.de/files/mck_industry_40_report.pdf (Accessed 15 September 2017).
[77] M. Rüßmann, M. Lorenz, P. Gerbert, M. Waldner, J. Justus, P. Engel, M. Harnisch, Industry 4.0 – The Future of Productivity and Growth in Manufacturing Industries, Boston Consulting Group, 2015 Available at https://www.bcgperspectives.com/content/articles/engineered_products_project_business_industry_40_future_productivity_growth_manufacturing_industries / (Accessed 15 September 2017).
[78] S. Heng, Industry – 4.0 Upgrading of Germany's Industrial Capabilities on the Horizon, Deutsche Bank Research, (2014) Available at https://www.dbresearch.com/PROD/DBR_INTERNET_EN-PROD/PROD0000000000333571.pdf (Accessed 15 September 2017).
[79] R. Geissbauer, J. Vedso, S. Schrauf, Industry 4.0 – Building the Digital Enterprise, PWC, 2016, https://www.pwc.com/gx/en/industries/industries-4.0/landing-page/industry-4.0-building-your-digital-enterprise-april-2016. pdf.
[80] J. Posada, C. Toro, I. Barandiaran, D. Stricker, R. de Amicis, E.B. Pinto, P. Eisert, J. Döllner, I. Vallarino, Visual computing as a key enabling technology for industry 4.0 and industrial internet, IEEE Comput. Graph. Appl. 35 (2) (2015) 26–40.
[81] D. Gorecky, M. Schmitt, M. Loskyll, D. Zühlke, Human-machine-interaction in the

industry 4.0 era. In industrial informatics (INDIN), 2014, 12th IEEE International Conference on (pp. 289–294). IEEE (2014).
[82] M. Hermann, T. Pentek, B. Otto, Design principles for industrie 4.0 scenarios. In system sciences (HICSS), 2016, 49th Hawaii International Conference on (pp. 3928–3937). IEEE (2016).
[83] N. Jazdi, Cyber physical systems in the context of Industry 4.0, Automation, Quality and Testing, Robotics, 2014 IEEE International Conference on (pp. 1–4). IEEE (2014).
[84] J. Lee, B. Bagheri, H. Kao, A cyber-physical systems architecture for industry 4.0-based manufacturing systems, Manuf. Lett. 3 (2015) 18–23.
[85] J. Lee, H.A. Kao, S. Yang, Service innovation and smart analytics for industry 4.0 and big data environment, Procedia Cirp 16 (2014) 3–8.
[86] L. Monostori, Cyber-physical production systems: Roots, expectations and R&D challenges, Procedia CIRP 17 (2014) 9–13.
[87] F. Shrouf, J. Ordieres, G. Miragliotta, Smart factories in Industry 4.0: A review of the concept and of energy management approached in production based on the Internet of Things paradigm, Industrial Engineering and Engineering Management (IEEM), 2014 IEEE International Conference on (pp. 697–701). IEEE (2014).
[88] V.D. Blondel, J.L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks, J. Stat. Mech: Theory Exp. (2008) (P1000).