Scientometrics (2021) 126:1745–1774
https://doi.org/10.1007/s11192-020-03803-z

Impact for whom? Mapping the users of public research with lexicon-based text mining

Andrea Bonaccorsi[1] · **Filippo Chiarello**[1] · **Gualtiero Fantoni**[1]

Received: 29 July 2020 / Published online: 17 December 2020
© The Author(s) 2020

Abstract
We contribute to the debate on societal impact of SSH by developing a methodology that allows a fine-grained observation of social groups that make use, directly or indirectly, of the results of research. We develop a lexicon of users with 76,857 entries, which saturates the semantic field of social groups of users and allows normalization. We use the lexicon in order to filter text structures in the 6637 impact case studies collected under the Research Excellence Framework in the UK. We then follow the steps recommended by Börner et al. (Annu Rev Inf Sci Technol 37:179–255, 2003) to build up visual maps of science, using co-occurrence of words describing users of research. We explore the properties of this novel kind of maps, in which science is seen from the perspective of research users.

Keywords Impact assessment · Societal impact · Science map · Supervised text mining · Research users · Lexicon

Introduction

As discussed by several authors, societal impact has become one of the criteria of ex ante project selection in many institutions and countries (Kanninen and Lemola 2006; Donovan 2011; Dance 2013; Atkinson 2014; Penfield et al. 2014). Some authors advocate impact analysis as a way to examine the effects of research agendas on the societal priorities and on distributional issues (Cozzens et al. 2002; Langfeldt and Scordato 2015).

It is also a crucial chapter in the ex post research assessment in some countries, such as United Kingdom. Within the UK Research Excellence Framework (REF) the assessment of impact has been responsible for 20% of the total score. The next planned exercise (2021 REF Exercise) will "assess the 'reach and significance' of impacts on the economy, society, culture, public policy or services, health, the environment or quality of life" with an increased weighting at 25% of the total score (REF 2019).

We greatly benefited from comments on an early draft of this paper from Pierre-Benoit Joly, Jordi Molas Gallart, Steven Hill, Gabi Lombardo and Jack Spaapen. All remaining errors are ours.

✉ Andrea Bonaccorsi
a.bonaccorsi@mail.com; andrea.bonaccorsi@unipi.it

[1] DESTEC, School of Engineering, University of Pisa, Largo Lucio Lazzarino 2, 56122 Pisa, Italy

The publication of REF case studies of impact has fueled a field of analysis (Derrick, Meijer and van Wijk 2014; Samuel and Derrick 2015; King's College and Digital Science 2015; Khazragui and Hudson 2015). In particular, as we will see below, the study by King's College and Digital Science (2015) has made use of advanced Text Mining techniques to investigate the structure and content of documents reporting the impact of research of UK departments.[1]

This surge of policy interest, however, comes in a period in which the scientific analysis of the concept of societal impact and of the potential and limits of existing methodologies has not yet come to a general agreement (Bozeman and Sarewitz 2011). As succinctly stated by Lutz Bornmann, impact evaluation is "still in the infant stage" (Bornmann 2013). This state of the art is confirmed by several reviews of the literature (Greenhalgh et al. 2016; Reale et al. 2018; Pedersen et al. 2020).

This paper is a contribution to the substantive and methodological work on the assessment of societal impact of research. From the substantive point of view, it develops the notion of target group, or *group of potential users of research*, as a necessary component of the design and implementation of research projects and of their evaluation.

From the methodological point of view, the paper strongly supports the idea, already advanced in the literature, that text mining techniques are promising in the field of impact assessment (King's College and Digital Science 2015; Bornmann et al. 2016). We build upon this pioneering literature by introducing a new methodology aimed at detecting and classifying all cases in which the authors of research documents mention a group of potential users of their research. The methodology is based on a dedicated lexicon, as a methodology adopted in the text mining literature.

We develop a full scale, replicable and scalable methodology to identify the user groups mentioned in research-based texts, such as research proposals (ex ante), impact case studies (ex post), or publications. We test the methodology on the collection of case studies developed under the Research Excellence Framework (REF) in the United Kingdom.

We build upon the work on impact assessment and try to develop a quantitative methodology. The notion of users, as we will see, is not entirely new in the literature. However, little efforts have been done to examine it systematically and to approach it in quantitative terms. In this paper we give a contribution to the user perspective on research impact assessment in terms of: (a) saturation of the semantic field; (b) normalization; (c) mapping at various levels of granularity.

In this way we open a new direction for the large literature dealing with science mapping. We develop a mapping exercise with a Text-mining, bottom up approach, generating a complete classification with the support of a dedicated lexicon. The result is a global map of impact of research, as described by the social groups of users of the results of research.

We must recognize that the term "users" may convey a narrow meaning, suggesting that we limit our analysis to social groups actively engaged into searching the results of research and applying it to their domain of interest. In reality, given the saturation of the semantic field, our definition of users includes all social groups that are affected, directly or indirectly, by the research activity. Our only requisite is that they are mentioned in a text, in this case the REF impact case study. This seems an acceptable restriction. In this sense our

[1] See the initial press releases at https://www.digital-science.com/press-releases/ref-impact-case-studies-to-be-analysed/; https://www.digital-science.com/blog/news/ref-impact-case-studies-to-be-analysed-as-digital-science-grows-its-consultancy-division/. Accessed December 17, 2019.

use of the word "users" is compatible with other approaches in the literature that prefer to use different notions, such as beneficiaries, or stakeholders, or target groups.

After the survey of literature (Sect. 2 and 3) we describe the methodology and the data following the workflow for the construction of maps suggested by Börner et al. (2003) (Sect. 4). We then develop a global map of research impact of all UK universities, build up clustering indicators, and discuss their properties (Sect. 5). The final section comments on the main findings and calls for more research on user target groups.

Impact for whom. Substantive and methodological challenges in the assessment of societal impact

Impact for whom

A promising perspective to address the issue of societal impact is opened by asking "impact for whom", or trying to define which are the social groups that are potentially interested by the research.

The issue of users of research is certainly not new. For example, among many others, the ISRIA guidelines recommend the definition of research users within the broader definition of stakeholders (Adam et al. 2018) and Rowe and Frewer (2005) classify several mechanisms to engage users into the research process. On the basis of previous research on societal impact and systematic reviews of international practices (Grant et al. 2010; Morgan and Grant 2013) a joint undertaking by the King's College and Digital Science has extensively examined the case studies of the REF using a text mining approach (King's College and Digital Science 2015), with a large follow-up of studies (Derrick 2014; Hinrichs and Grant 2015; Digital Science 2015, 2016). Adams et al. (2015) have used text similarity in REF impact case studies to illustrate the landscape of research activities of leading universities. One of the key findings of these studies has been the identification of research beneficiaries and the mapping between research projects, topics, and research beneficiaries. The total number of beneficiaries, or users of research, is in the order of dozens.

These initial suggestions point to the need to develop a full scale analysis of research users, with the final aim to provide tools for semi-automatic extraction of knowledge from documents. However, this will require a very large number of items in the definition of users and a high level of granularity. This goal is beyond the current state of the art and is the main object of this paper.

Why the identification of potential users of research is difficult

In academic research evaluation it is clear that users of research are, by definition, other researchers. The quality of research is defined as a function of the use of published research by other researchers.

Conceptually, the possibility to identify precisely the social groups of researchers and to define their boundaries (for example by compiling lists of journals that researchers regularly read and cite and in which they publish) is a requisite for the use of bibliometric indicators for research evaluation. In those fields in which bibliometric indicators are not used, the practice of peer review follows exactly the same general logic- asking other researchers, as actual or potential users of published research, to formulate a judgment.

When coming to the societal impact, the question "impact on whom" becomes much more problematic. It is useful to review the literature on societal impact of research from this angle, before advancing formal definitions and a methodology for data extraction and measurement.

First, potential users are heterogeneous. It is largely recognized that the ways in which research has an influence on society are multiple and specific to scientific disciplines (Bornmann2014; Bornmann and Marx 2014; Jacobsson and Perez 2010; Jacobsson et al. 2014). Let us follow the use of "impact pathways" to describe this heterogeneity. Miettinen et al. (2015) develop the epistemic rationale for such a multiplicity, arguing that "science (is) a heterogeneous social activity where different disciplines possess dissimilar methodologies, ontologies and forms of interaction with society" (Miettinen et al. 2015, p. 258). Research in political science is different from research in oncology not only because their scientific foundations, methods, objects and cognitive styles are different, but also because they talk to different user groups. Muhonen et al. (2020) inductively derive as many as twelve different types of impact pathways.

Second, potential users have different time scales (Adam et al. 2018). The time scale of societal impact is not always known in advance, is not fixed, and varies greatly across disciplines, technologies, and institutional and social systems (Martin 2011). In some cases it goes well beyond the time horizon of actors themselves (researchers, funding agencies, policy makers, stakeholders). This implies that in many cases what will be observed will not be a specific product, or a discrete event in time (e.g. a policy document, a legislation, a regulation) but a process, whose start and end dates might be unknown and whose boundaries might be difficult to trace.[2] This is another major difference with respect to the impact on researchers: in the latter case the time window for observing the impact on citations can be known with a certain precision in most scientific disciplines (with the notable exception of sleeping beauties). The standardization of the time window of citations used in bibliometrics is therefore acceptable. Using multiple time windows is common practice, but their duration is standardized.

Third, potential users interact with researchers in a variety of ways. The final impact on society does not depend only on the research side, but on the societal side, that is, its institutions, actors, formal and informal rules, culture and values. The analysis of societal impact, therefore, requires a theory of research utilization, which in turn is based on theories of information processing, diffusion of innovations and decision making in various user contexts (Leckie et al. 1996; Sarewitz and Pielke 2007; Mohammadi et al. 2015). This also means that the final impact of research on users may come from a variety of contributions, often from several sources, among which it is often impossible to establish the authorship.

Fourth, early interaction with potential users enhances the impact of research. There is certain agreement in the literature on the observation that the impact of research is greatly magnified if researchers involve the potential users in the research process at an early stage (Nutley et al. 2003, 2007; Meyer 2011). Potential users are not passive recipients of useful information, but have their own active information search and processing strategies and use information for a variety of uses. It is recognized that passive processes of knowledge

[2] In the case studies examined in the ASIRPA project, for example, the time scale of impact of agricultural research on farmers, environmental authorities and regulators varies between few years and 30 years (with two ouliers at 50 and 80 years), with an average delay of 14 years for the intermediate impact and additional 6 years for the impact in terms of diffusion (Colinet et al. 2014; Matt et al. 2017).

dissemination are ineffective. Researchers should target audiences purposefully and precisely (Lavis et al. 2003; Krücken, Meier and Müller 2009). Expected and intended impact should be explicitly included in research proposals (Holbrook and Frodeman 2011). This kind of early interaction with potential users is not requested for academic impact, although the social interaction with peers before the publication of results is common practice.

Table 1 summarizes the main differences between research evaluation and societal impact assessment from the perspective of potential users. Faced with these differences, it is clear why the methodological foundations and the assessment practices are different in the two cases.

In the case of research evaluation the clear identification of a single category of potential users makes it possible to identify its boundaries, to define a measurement process, to compare and standardize the measures. Normalization is possible. There is an assumption of a one-to-one mapping between the activity of researchers, their observed output (publications) and their impact (citations). The formal notion of authorship reinforces this assumption (Cronin 1984). The notion of authorship ensures that any given evidence of impact (citation) can be attributed to a formal entity (publication), which is in turn unambiguously credited to one or more authors (Cronin 2005). This makes it possible to adopt a form of attribution approach.

By attribution is meant a causal allocation of a demonstrated impact upstream to the research activity. By causal it is meant a relation that controls, to the best possible degree, all other factors that may impinge upon the relevant observed variables.

As it has been argued, a strict attribution approach is highly problematic in impact assessment, due to multiple influences on potential users, coming from several research fields, often combined together in unplanned and unexpected ways, with a number of indirect effects over an extended and uncertain time scale (Martin 2011). As an alternative, several authors propose the notion of *contribution*, or partial, empirically observed, participation in a dynamic process whose effect can be demonstrated but in which independent causal factors cannot be controlled with reasonable approximation (Spaapen and Van Drooge 2011; de Jong et al. 2011, 2014; Bell et al. 2011; Morton 2015). The notion of contribution is at the core of the ASIRPA methodology, based on standardized case studies and developed by Pierre-Benoit Joly and co-authors for the French Institute of Agricultural Research (INRA) (Colinet et al. 2014; Joly et al. 2015; Matt et al. 2017). It is also central to the notion of *productive interaction*, an explicit recognition that potential users have a variety of ways in which they can use research results. The SIAMPI project has developed a framework for the identification and analysis of productive interactions (Spaapen and Van Drooge 2011; Molas-Gallart and Tang 2011; De Jong et al. 2014).

On the basis of this analysis we suggest to operationalize the concepts in an appropriate way.

Definition 1(a) *Potential users* are individual entities that might be influenced by the research activity and/or research results. This definition covers all possible entities that engage an active or passive relation with the research activity.[3]

[3] Strictly speaking, these definitions would cover also researchers (they are by definition influenced by their own research), funders (they fund research and look for impact), universities (they receive money from research), administrators, auditors, regulators and so on. While they are clearly out of scope in the current analysis, we keep the definition deliberately broad.

Table 1 Role of potential users of research in two types of research assessment. Research evaluation vs. societal impact assessment

	Research evaluation	Assesment of societal impact
Nature of potential users	Homogeneuous (researchers)	Heterogeneous (many social groups)
Boundaries of group of potential users	Well defined (authors of scientific publications by Subject Category of journals)	Ill defined
Time scale of impact	Well defined (time window for citations)	Ill defined or unknown
Forms of interaction between researchers and potential users	Mostly unidirectional (publications, citations) Some interactive (seminars, conferences)	Mostly interactive
Early involvement of potential users	Some (but not mandatory) social interaction with peers before publication	Early social interaction crucial for societal impact
Main epistemological and methodological approach for assessment	Causal attribution	Contribution

Source: our elaboration

Definition 1(b) *Groups of potential users* are recognizable social or collective entities that might be influenced by the research activity and/or research results.

Definition 2 *Target and target groups* are entities or groups of entities (potential users) on which researchers claim to have an effect.

With these definitions at hand it will be possible to engage in a large scale mapping exercise.

Mapping science from the perspective of users

In recent years, a fascinating new field of science representation has been reopened, building on the pioneering co-word analyses of Callon and co-authors (Callon 1983; Callon et al. 1986, Callon and Courtial 1989) but using more advanced graph-theoretic algorithms and powerful visualization techniques. In the more recent literature the potential of co-word analysis has been clearly shown (Leydesdorff 1989; Leydesdorff and Nerghes 2017). These maps allow detailed representations of disciplines and/or topics and their evolution over time at aggregate level. Large scale maps of science have been produced on the basis of co-occurrence of words and co-citations (Moya-Anegon et al. 2004, 2007; Boyack et al. 2005) or on the basis of views of articles in digital platforms (Bollen et al. 2009).

There is a large agreement on the structural properties of the world map of science (Klavans and Boyack 2009). More recently, overlay maps that allow the interpretation of distance between nodes have been introduced for mapping science, using publication data (Rafols et al. 2010; Leydesdorff and Rafols 2009, 2012; Carley et al. 2017) and for mapping technology, using patent data (Leydesdorff et al. 2014; Kay et al. 2014). Overlay maps position individual entities, such as universities, companies, regions or countries, in the global world map of science or technology.

In all these cases the maps represent science or technology from the perspective of production of knowledge. Would it be possible to build up maps of science from the perspective of users of knowledge? Maps that are not supply-side, but user-side? A science map from the perspective of users would not have nodes representing scientific disciplines, journals, patents, or topics. It would have nodes representing social groups that benefit from the research of a country, or a region, or a single university.

It would be a useful complement to the map of science, offering a different perspective. It might be used by universities as an input to the definition of long term strategy, or by the government to get a summary view of the impact of research funded, or to give account to the public opinion of the scope and depth of social groups positively affected by public research.

There is an obvious difficulty here. Maps of science and technology make use of existing classifications of Subject categories of journals, or Patent classes of patents. Or they make use of keywords associated to papers or abstract of patents. In all these cases there are authoritative sources of classification that can be used to normalize the data and define precisely the distance between nodes in the map. Nothing similar does exist for the users of research.

Here comes the lexicon approach to text mining. We advocate the use of text mining because the level of codification of social groups of users of research is extremely low in established statistical systems. At the same time, the conventional approach to text mining

fails to build up the conditions for standardization and normalization. We turn to these issues below.

Improving text mining for societal impact mapping and assessment

Text mining as a promising approach for impact assessment

Given the issues of heterogeneity, uncertainty of time scale, multiple influences and weak attribution it would be perhaps natural to adopt a qualitative approach, based on in-depth case studies. This is not the only option, however. As discussed by Joly et al. (2015), most impact assessment studies are indeed based on case studies, but this methodology does not ensure, if not subject to standardization, the requirements for comparability and scalability.

We suggest that recent methodological developments allow the exploration of quantitative analyses that are able to cope with high levels of diversity and variability. In particular, text mining offers a menu of tools that give full justice to the multifaceted and complex nature of the problem of research assessment, while allowing some a certain level of comparability and measurability.

The use of text mining for impact assessment has been recommended by Bornmann et al. (2016). Hecking and Leydesdorff (2019) compare the Latent Dirichlet Allocation (LDA) technique (a Topic modeling approach within text mining) to Principal Component Analysis as tools for mapping and conclude that LDA generates reproducible and consistent results, although it is vulnerable to small changes in the corpus and/or in the number of different topics.

A pioneering application has been done by King's College and Digital Science (2015) in the analysis of the collection of REF impact case studies, which we will also examine below.

The authors of that study have examined the ways in which the authors of the REF reports describe the impact of research on specific groups of users. The technique used is Topic Modeling, the most largely used tool in text mining to retrieve and classify semantic content from large corpora of texts. The main result is a map in which 65 categories are listed, from "business" to "citizens", from "teachers" to "administrators". A remarkable part of the analysis is the clear demonstration that the impact of research does not follow a linear path, from clearly identified products of research to clearly observable effects. Rather, the impact is the result of a multiplicity of contributions, often from distant disciplines.

Our approach differs from the one followed by King's College and Digital Science (2015). First, we use a full scale lexicon that is able to extract *all* words that represent users, saturating the semantic space and allowing the normalization of entities. Second, we are able to build up quantitative indicators with desired properties. Finally, we demonstrate applications that exploit various levels of granularity.

Beyond topic modeling: the lexicon approach to text mining

Conventional text mining leaves unsolved an important issue. Being based on bottom up analysis of texts, it generates findings that are not necessarily associated to a clear semantic content, or meaning. In particular, the most largely used technique, i.e. Topic modeling, delivers collections of words, or topics, whose semantic meaning is described in statistical

terms. This well known limitation is largely discussed in the technical literature (Blei and Lafferty 2006; Lee et al. 2009; Blei 2012; Chen et al. 2015). This is particularly worrysome in the context of impact assessment of research, due to the intrinsic and large heterogeneity of words used in research texts.

From the perspective of scientometrics, this is a fatal limitation, insofar as it inhibits the normalization of measures, which is a precondition for the development of indicators and metrics.

Therefore we suggest to integrate text mining techniques with a lexicon approach (Zhang et al. 2011). This is a top down, or supervised approach, based on domain knowledge that allows the filtering of words according to a predefined dictionary of words that saturate a certain semantic field. In our case the lexicon is associated to a full scale development of definitions, so that it is also labeled Enriched dictionary.

Lexicons are a fundamental tool in text mining. There are two main types of lexicons: word lexicons and domain specific lexicons. Word lexicons include general word lexicons (Turney and Littman 2003; Hu and Liu 2004a), that are used as a universal text mining tool. Word lexicons are also largely used in one of the most diffused application of text mining, i.e. sentiment analysis (Esuli and Sebastiani 2006, 2010; Baccianella and Esuli 2010; Tan et al. 2012; Jang et al. 2013; Mustafa 2013; Mohammadi et al. 2015). In this application lexicons are developed in order to automatically classify words in terms of subjectivity and polarity (positive, negative, or neutral) (Barbosa and Feng 2010; Hemmatian and Sohrabi 2019).

On the contrary, domain specific lexicons are built by processing the text of corpora that refer to narrow fields of experience. These lexicons embed domain-specific knowledge and terminology in order to automatically classify words. The literature offers a large menu of applications, from products such as mobile phone, tablet or transport (Rathan et al. 2017; Zhou et al. 2017) to services such as hotel, restaurant, customer service or movie (Molina-González et al. 2015;Jiménez-Zafra et al. 2015; Chinsha and Joseph 2015; Chao and Yang 2018). In these cases the texts are taken from online customer reviews.

The methodology followed in the construction of the User lexicon adopted for this study is discussed at lenght in "Appendix", on the basis of Chiarello et al. (2018) and of the examples of applications in Chiarello et al. (2017) and Bonaccorsi et al. (2017).

Lexicons are a peculiar type of written text, characterized by authoritativeness, saturation and update. They share the properties of well established institutions in natural language, i.e. dictionaries. In fact, a dictionary must be composed of entries established by some authority, most often an academic one and/or an authority established since long time by reputation (e.g. editorial initiatives of prestigious publishers). Saturation means that all words that are related to the domain of the dictionary must be included. It is a major flaw of a dictionary the lack of important entries. A dictionary is characterized by a property of semantic saturation: all words that have a meaning associated to a given field are included in the dictionary. In the computational linguistic world, lexicons are built with the same requirements, but without a board of editors of human experts (Zhang et al. 2011).

Construction of the user lexicon

These formal requisites, that used to be appropriate only for established dictionaries, are currently satisfied by a larger variety of sources. In particular, the huge power of text mining techniques has made it possible to automatize at least some of the steps needed to create a formal lexicon. Chiarello et al. (2018) illustrates the steps undertaken in order to build

up a lexicon of users, while Chiarello et al. (2020) illustrates a lexicon of advantages and disadvantages. The user lexicon currently includes 76.857 entries, that have been shown to saturate the semantic field of users. It includes, among others, all jobs, work positions, professions, hobbies, patient roles, sports, creative and entertainment roles, political, institutional and organizational roles, social roles, that have been classified in hundreds of official sources. In particular, this includes all potential users and target groups, as defined above. A full-scale description of the methodology followed to build up the lexicon is available in "Appendix". In order to ensure transparency of the procedure, replicability and scalability, we publish the entire REF dataset tagged with the research user tags. The full collection is available at https://github.com/FilippoChiarello/REF_target_groups_data?files=1.

In the same collection we make it available the full-scale tagging of the single most frequently used word in impact assessment. i.e. "people".

Data extraction and text pre-processing

The corpus is composed of 6637 REF impact case studies. They generally follow a template illustrated in the REF criteria. The template has a Title and five main text sections, plus the name of the Submitting Institution and the Unit of Assessment. In addition to the Title of the case study, the text sections of the template and the indicative lengths, as recommended in the REF criteria are:

1. Summary of the impact, 100 words
2. Underpinning research, 500 words
3. References to the research, 6 references
4. Details of the impact, 750 words
5. Sources to corroborate the impact, 10 references

We take into consideration the sections *Summary of the impact* and *Details of the impact*.[4] It is common practice in computational linguistics to examine the length of documents to be included in a corpus in order to ensure comparability. Figure 1 shows that the limits established by the REF criteria are not always respected. Nevertheless, since the distribution of the length is almost normal and there are not outliers it is appropriate to include all documents in the corpus.

Within the REF repository projects are classified using three criteria.

- *Impact type* There are eight Summary Impact Types. These follow the PESTLE convention (Political, Economic, Societal, Technological, Legal, and Environmental) widely used in government policy development, with the addition of Health and Cultural impact types.
- *Units of assessment (UoA)* Institutions were invited to make REF submissions in 36 subject areas, called units of assessment (*UoA*s), each of which had a separate expert panel.

[4] We do not make use of the section "Underpinning of research" since it gives background information of research and publications, rather than information on potential users. In a future study the relation between this section and the various measure of Frequency, Diversity and Specificity will be examined.

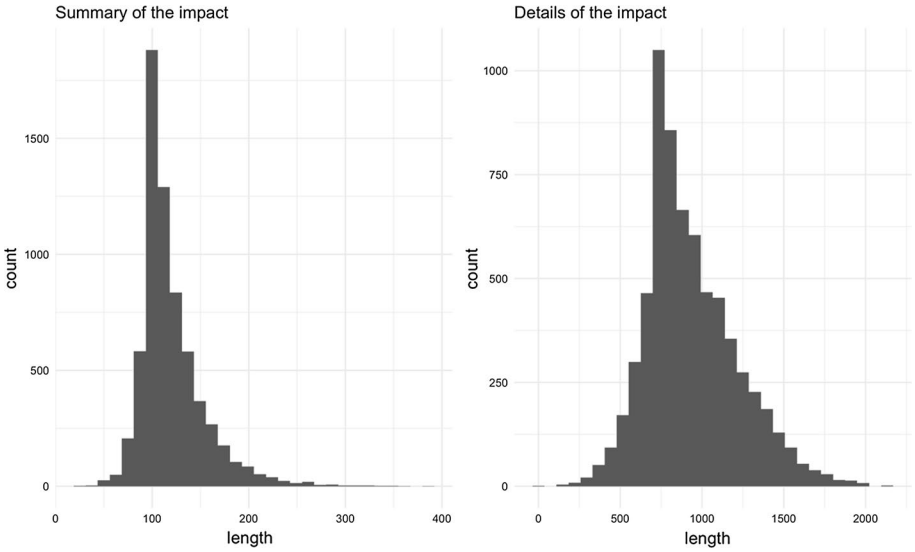

Fig. 1 Distribution of number of words in relevant sections of the REF impact case studies

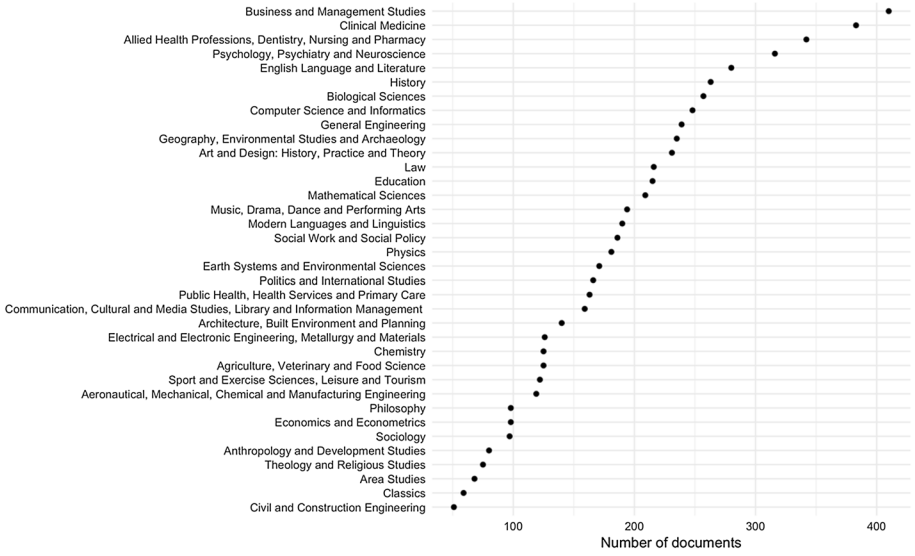

Fig. 2 Number of documents per unit of assessment (UoA) in REF impact case studies

- *Research subject areas* The REF Impact case studies are assigned to one or more Research Subject Areas (to a maximum of three) by text analysis of the 'Underpinning research' (Sect. 2 of the Impact case study template). This is a guide to text search that uses a disciplinary structure that is more fine-grained than the one in the 36 Units of assessment.

Figure 2 shows the number of documents per Unit of assessment. Before submitting the collection to the extraction of words there is a need for pre-processing the texts. Our Natural Language Processing (NLP) system follows the following typical steps (Manning et al. 1999): *sentence splitting and tokenization*; *POS tagging and lemmatization*; *target groups annotation*. As it is clarified in "Appendix", the procedure allows the recognition of users even when the verbal expression is indirect. For example, in the sentence "The new scanning method in this research can prevent cancer deaths among women by 5%" the term "cancer deaths" indirectly refers to patients. The system recognizes that the word "cancer" is associated to specific categories of users (for example, cancer patients) and recognizes that the term "cancer deaths" actually means "death of cancer patients".

Accessing the website https://github.com/FilippoChiarello/REF_target_groups_data?files=1 the readers may directly verify the procedure.

Construction of the maps

Following Börner et al. (2003) after the definition of the variables to be used (Data extraction), a number of steps should be followed to build up a map. They are: Unit of analysis, Measures, Similarity, Ordination, and Display. These steps addressed in the following sections.

Unit of analysis

The workflow we have developed allows two units of analysis: words and documents. In the construction of science maps we will use words as units of analysis. The map will be a network representation of co-occurrence of words filtered with the User lexicon. In this study we will build the map at the country level, showing the impact of all UK universities. In a companion paper the maps will be drawn for individual universities, with an aim to examine the social impact profile of institutions and to compare them with academic impact.

The same methodology, however, can be used to examine documents as units of analysis. This will require the construction of indicators that can be aggregated at document level. This will be done in future research, with an aim to compare indicators across broad disciplinary areas (for example, STEM vs SSH).

Measures

For the construction of the impact map we use a measure of occurrence of words that describe users of research.

Given that we do not have classification schemes or established lists of keywords to use, it is mandatory to establish the validity of the use of words extracted with the User lexicon. This amounts to discuss the issue of Recall and Precision of the measures obtained.

In fact, a collection of users from textual sources belongs to the class of Named Entity Recognition problems. There are several methods and algorithms to deal with the entity extraction task, but the most used ones can be divided in two groups: supervised methods and lexicon methods (Nadeau and Sekine 2007).

Supervised methods tackle the task by extracting relevant statistics from an annotated corpus. A portion of the corpus is annotated *manually* in order to identify a large-enough

set of examples for the trainining of a Machine Learning model. Lexicon approaches (like the one proposed in the present paper) automatically search for entities using a pre-collected list of entities. In this approach, the human effort (and knowledge) is applied to the more value-added task of searching for pre-existing lexicons describing the entities.

Both approaches have advantages and disadvantages (see "Appendix" for a discussion): here we focus on the fact that using an external lexicon makes it hard to compute the recall of the approach.

Recall is defined as the fraction of the total amount of relevant entities that were actually retrieved. Since in lexicon approaches the total amount of relevant entities is unkown, it is impossibile to compute this statistic without a manual review of the entire corpus. This manual review (which is similar to a manual annotation in the supervised approach) lowers the value added brought by the lexicon approach (i.e. minimizing the manual effort). Furthermore, the unbalance of the dataset (only 2% of total words are target groups) makes it really rare to find target groups.

For this reason, we rather give qualitative evidence of the coverage (recall) of the dictionary. We use two different approaches:

- We randomly examine 1% of words ($n = 82.306$ out of 8.230.598 in total) and check manually whether the lexicon missed important information about users.
- We publish the entire tagged dataset, in order to let the reader validate the results https://github.com/FilippoChiarello/REF_target_groups_data?files=1.

The set (even if not statistically representative) is selected in order to identify potential biases of the method that can lower the recall of the output. Two main cases were notable. First, the procedure missed some nouns that must be considered general and abstract, but that in some cases might, at least in principle, point to a concrete group of users. This is the case of words such as "management", "policy", "service", "training", "region", "media" or "business". We checked whether in the same document these words appeared and were correctly tagged. In most cases the classification was correct, since the word indeed had a different meaning (e.g. "clinical management", "pharmacological management", "fluid management", "management of difficult cases" and the like). In most other cases the word was correctly tagged when it pointed to a concrete user (e.g. "business company"). Therefore, these cases cannot be considered source of poor recall. The second case refers to proper names of organizations and companies (e.g. Roche, Astra Zeneca, Glaxo Smith Kline, or IBM), media and newspapers (e.g. Daily Telegraph, Guardian), and charities (e.g. Prostate Cancer Charity). In this case the name may point to a concrete user of research, following various pathways. In order to improve the recall, however, a full scale treatment of proper names would be needed, by including open archives such as GRID[5] and other sources for the ID definition. This is left to future research.

With respect to precision, we extracted all cases that refer to the most used word in the REF cases, i.e. "people", which appears in more than 30% of documents and shows up in as many as 1410 different versions. This is the most generic word, so it is reasonable to assume it might be affected by lack of precision.

The full list is available at https://github.com/FilippoChiarello/REF_target_groups_data?files=1.

[5] See https://grid.ac/ for the ID of research organizations, or http://org-id.guide/results for the (complex) issue of ID of companies.

Table 2 Tokenization, lemmatization and annotation of a sentence in the corpus

Doc_id	Sentence_id	Token_id	Token	Lemma	Xpos	Full_target_group
1855	1	1	Each	Each	DT	NA
1855	1	2	Year	Year	NN	NA
1855	1	3	Year	Year	NN	NA
1855	1	4	IN	IN	IN	NA
1855	1	5	England	England	NN	NA
1855	1	6	Alone	Alone	RB	NA
1855	1	7	Alone	Alone	RB	NA
1855	1	8	Approximately	Approximately	RB	NA
1855	1	9	152,000	152,000	CD	NA
1855	1	10	People	People	NN	People
1855	1	11	Suffer	Suffer	VBP	NA
1855	1	12	a	a	DT	NA
1855	1	13	Stroke	Stroke	NN	NA
1855	1	14	Stroke	Stroke	NN	NA

We inspected manually the tagged words and identified two sources of concern with respect to precision, occurring in 113 cases, or 8% of cases in which the word "people" appears (thus giving a precision of at least 92%).

The first and largest source of noise is the separate tagging of expressions in which the correct word is associated to quantifiers (e.g. "many people", "most people", "few people" and the like). We examined them carefully and decided there was no reason to treat them as separate entities. We therefore developed software code to eliminate all quantifiers from the tagging procedure across all words. The second problem has to do with the close association of the correct word with a verb, for which the word is an object (e.g. "touching people", "moving people"). Here the tagging procedure interprets the verbs as they were adjectives. In this case the disambiguation is more complex and in some cases there is no reason to eliminate these bigrams. We then decided to leave these expressions as separate. Summing up, we conservatively estimate the revised tagging procedure to have a precision in excess of 96%. Furthermore, the inclusion of separate n-grams does not influence the total number of user groups, but only their diversity and specificity (see below). The overall metrics, as discussed below, will not be significantly influenced by the remaining lack of precision.

Table 2 shows the output of the NLP procedure after the improvements discussed above. It shows the procedure for a sentence contained in the corpus ("Each year, in England alone, approximately 152,000 people suffer a stroke."). As it can be seen, the automatic annotation system isolates the only word ("people") that may be part of a target group.

Similarity

After extraction of words representing users of research we build up the map by calculating the co-occurrence between words in the same document.

The corpus contains 8,230,598 words in total and 141,705 different words. By annotating the entire corpus with the entries of the lexicon we find that the total number of words referring to target groups is 169,037, while the number of different target groups is

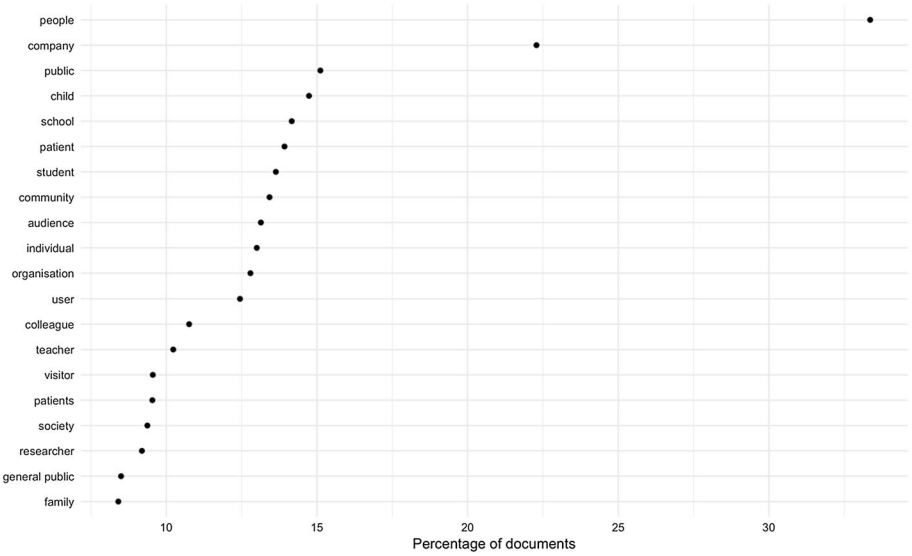

Fig. 3 Top 20 occorrences of words referring to target groups in the corpus of REF impact case studies

1830, or 1.3% of different words. The number of documents that contain at least one target group is 6628, or 99.9% of the total. Only for nine documents we were unable to locate any word referring to a target group. As already stated, the full collection of REF documents tagged with the user groups is available for inspection at https://github.com/FilippoChiarello/REF_target_groups_data?files=1.

By examining the frequency distribution of words representing users it is clear that a number of them have a broad semantic content, i.e. are generic terms.

Figure 3 offers a vivid demonstration of this issue. As many as 37% of all projects include *people*, and as many as 25% mention *company* as isolated words. Among the top 20 occurrences we find extremely generic words such as *public*,[6] *community*, *individual*, *organization*, *user*, or *society*. Slightly more specific are the words referring to the school or youth context (*child*, *school*, *student*, *teacher*) or the health context (*patient*, *patients*). In order to find more specific words we have to go much further down the ranking. Please note that in all these cases these words do *not* appear in combination with other that might increase the specificity, but in isolation. Should the same word appear in combination with other more semantically connotated words, they would form a separate target group. As an example, the word people is considered part of a separate expression in the following examples: people with cystic fibrosis, people with primordial dwarfism, people with rheumatoid arthritis, ordinary people in extraordinary situation, people in senior management, people from different background, key policy people in UK government, specific community of people, young people in deprived community in Glasgow. Each of these expressions

[6] Note that "public" in this context is a noun, not an adjective, referring to a generic audience of listeners. The word "general public" is a more specific noun, referring to the audience of listeners characterized by lack of specialization in the topic. It appears as a separate expression than "public", although it may be considered an instantiation of the more generic term.

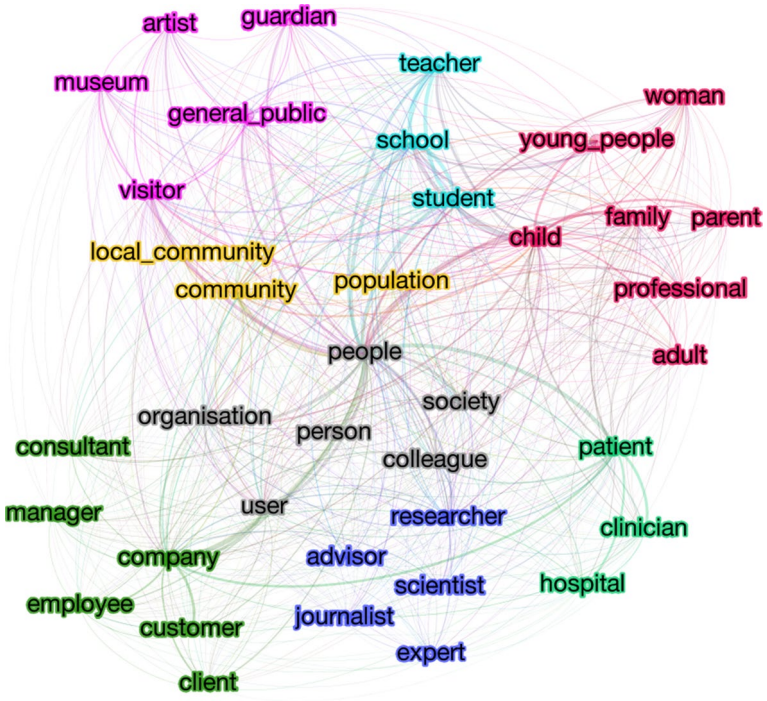

Fig. 4 Coarse-grain map of users of university research in UK. Modularity level 1.0

is considered as a separate target group. Nevertheless, generic words still appear after extraction.

In order to build the graph of co-occurrences the presence of generic words is a serious obstacle, since these words will create large connected components that will obscure the presence of clusters of semantically delineated words. In order to cope with this issue we first compute the degree of each word representing a user group. As expected the distribution is highly skewed, with few words having extremely large degree. After experimenting with several thresholds, we eliminate all words with a degree larger than 1000.

Ordination

Following Börner et al. (2003) after the construction of similarity measures a crucial step is the dimensionality reduction of resulting matrices. We perform a clustering procedure based on the co-occurrence in the same document of words representing different user groups. We follow the community detection algorithm developed by Blondel et al. (2008).

We apply the clustering algorithm in two versions. The first one is carried out by fixing the modularity value at 1.0. This generates a coarse-grained map, including 8 clusters. The second map is generated by taking modularity at 0.2, obtaining 46 clusters.

Given the exploratory nature of our research, we find it premature to apply methods of optimal determination of the number of clusters, as well as Machine Learning techniques for the labeling of the clusters. As a matter of fact, the semantic content of clusters at coarse and fine-grained level of resolution is coherent and understandable to a surprising

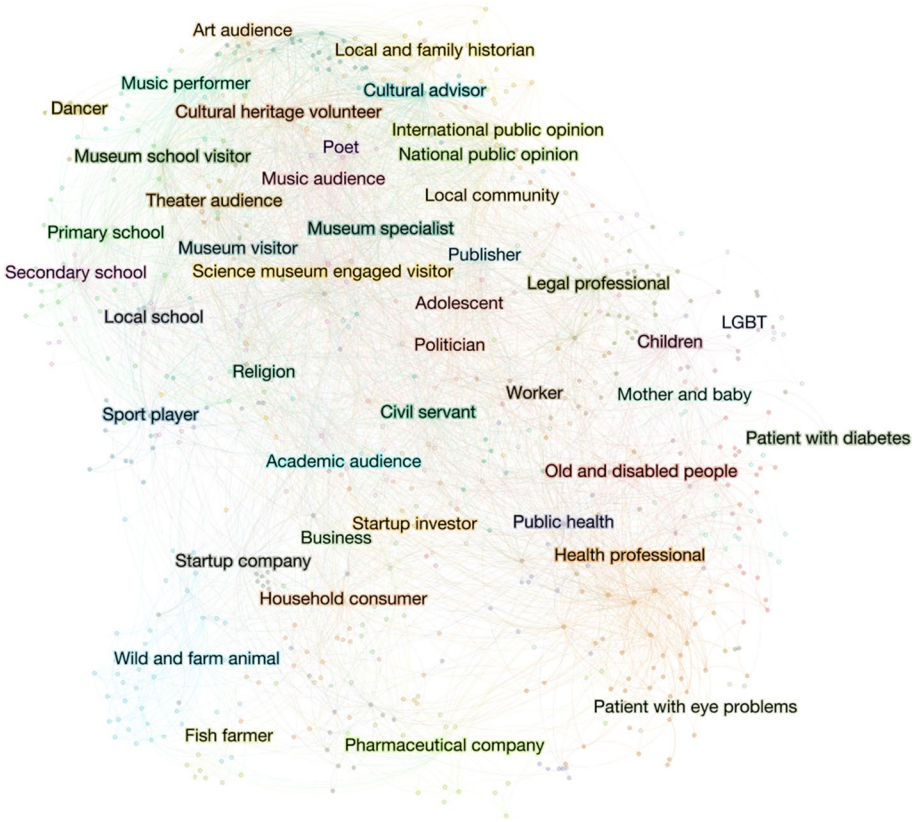

Fig. 5 Fine-grained map of users of university research in UK. Modularity level 0.2

degree. In future research it will be possible to compare alternative clustering results and apply the relevant metrics.

Display

Figure 4 shows the coarse-grained map, while Fig. 5 shows the fine-grained map. The display of the maps is based on R.

It is available and navigable at https://github.com/FilippoChiarello/REF_target_groups_data?files=1.

The findings from the mapping exercise are commented in Sect. 5 of the paper.

Findings and discussion

The maps give, first of all, an impressive view of the size and complexity of research in society. The density of the graph suggests that the impact of research follows a variety of pathways involving a huge diversity of societal actors. Interestingly, the clustering exercise delivers groups of words with a clear meaning in terms of impact and user groups.

The coarse-grained map allows to identify 8 clusters. The words with the largest degree in the cluster are coloured in different ways in Fig. 4. The clusters can be labelled as follows:

(a) Art and museum (museum, artist, guardian, visitor, general public)
(b) School (teacher, school, student)
(c) Family and welfare (child, young people, family, parent, woman, professional, adult)
(d) Community (local community, community, population)
(e) Health (patient, clinician, hospital)
(f) Expertise (researcher, expert, scientist, advisor, journalist)
(g) Economy (company, consultant, manager, employee, customer, client)
(h) Gatekeepers (people, person, society, user, colleague, organization).

With the exception of the cluster gatekeepers, which includes generic words that have survived the filtering procedure, all other clusters point to clearly delineated impact pathways, describing large sub-systems of modern societies on which universities have an impact. This map largely confirms the literature, already discussed in this paper, that has examined the variety of impact pathways, as well as the existing analyses of REF. At the same time it offers quantitative large scale evidence of social groups of users, as well as it shows their proximity, or distance, in the textual space.

By decreasing the modularity of the clustering algorithm is it possible, however, to go largely beyond the state of the art in the literature and offer a more fine-grained analysis of impact pathways. We explore a map with a modularity value of 0.2, which delivers 46 clusters. Table 3 gives a full description of the clusters, including the 5 words with the highest degree in the cluster (in descending order of degree). As stated above, the full list of words associated to the clusters is available at the github website.

Again, we find a clean structure, showing well delineated co-occurrences of words that describe groups of users. It must be remarked that the filtering of words obtained with the technical lexicon delivers words that correspond nicely to the definition of user groups. The map allows the identification of very precise impact pathways. These pathways take a sufficient density at national level to emerge as separate clusters. To make only a few examples, the support that universities offer to art can be finely examined with respect to dance, poetry, theater, art, music, publishing and media. The support to the world of museums takes the form of volunteer activity, support to visitors and to school visitors, and specialist advice for collections and for science museums. The impact of universities on social work and social needs is focused on maternity, child protection, adolescence, and civil rights. The role of universities as source of expert advice is visible in influencing the public opinion at national and international level, supporting policy makers, providing advice to various legal professionals, offering cultural consultancy, and studying local and family past history.

It is also possible to discover impact pathways that were not visible with a coarse-grained map: for example universities contribute to the social dialogue by interacting with religious communities. Or, in the field of business, it is remarkable that the only industry collaboration that takes sufficient density to emerge as a distinct cluster is with pharmaceutical industry.

For each of these clusters it is possible to examine the full list of words, obtaining further insights on the specific pathways. Furthermore, by choosing a smaller modularity value it would be possible to generate a larger number of clusters, increasing the granularity.

Table 3 Clusters of user groups of research of UK universities

No.	Cluster	Main words
1	Cultural heritage volunteer	Volunteer, wider society, citizen science, heritage organization, community centre
2	Secondary school	Secondary school, secondary school student, high school, Teacher, school leader
3	Media producer	Producer, guest, media professional, academic consultant, production company
4	Church community	Leader, church leader, international community, base company, fabian society
5	Civil society	Civil society, citizen, international organization, non-governmental organization, public service
6	Public health	Public health, European Society of Cardiology, Kcl researcher, elderly, hospital
7	Publisher	Publisher, everyone, anyone, many people, ordinary people
8	Museum visitor	Unique visitor, university student, register user, unique visitor per month
9	Museum school visitor	School child, visitor per year, school engagement, local museum
10	Music audience	Listener, non academic audience, ancestor, someone, non specialist audience
11	Worker	Worker, workforce, trade union, health worker, migrant worker
12	Art audience	New audience, diverse audience, global audience, total audience, art organisation
13	Health professional	Health professional, health care professional, nurse, doctor, physician
14	National public opinion	Public audience, wide public audience, average audience
15	Academic audience	Professor, lecturer, senior lecturer
16	Civil servant	Civil servant, senior manager, senior civil servant, specialist adviser, voluntary organization
17	Theater audience	Actor, reviewer, audience feedback, theater company, assistant director
18	International public opinion	International audience, general audience, large audience, public, lay audience
19	Local/family historian	Historian, archivist, family historian, local historian, librarian
20	Wild and farm animal	Animal, farmer, healthy animal, owner, bird
21	Business	Trade, enterprise, social enterprise, international trade, entrepreneur
22	Local school	Local school, wider community, involved teacher
23	Startup company	Engineer, designer, phd student, wider community, startup company
24	Startup investor	Investor, inventor, stem cell, skilled people, private investor
25	Children	Social worker, protected child, psychologist, psychiatrist, adolescent
26	Adolescent	Friend, youth, girl, teenager, academic community
27	Politician	Politician, economist, observer, public policy, commentator

Table 3 (continued)

No.	Cluster	Main words
28	Legal professional	Judge, victim, lawyer, applicant, solicitor
29	Local community	Community group, communities, community engagement, local people, community organization
30	Music performer	Musician, performer, composer, audience, choir
31	Science museum engaged visitor	Public engagement, royal society, new scientist, chief scientific advisor, royal society summer
32	Mother and baby	Mother, baby, infant, young child, midwife
33	Household consumer	Consumer, household, special adviser, retailer, expert advisor
34	Sport player	Plater, athlete, coach, olympics, elite athlete
35	Old and disabled people	Service user, older people, disabled people, people with dementia, user group
36	Poet	Poet, audience member, spectator, visual artist, male
37	Pharmaceutical company	Pharmaceutical company, major pharmaceutical company, commercial company, major company, local population
38	Primary school	Classroom, primary school, school student, school pupil, school teacher
39	Museum specialist	British museum, specialist, visitor experience, visitor number, tourist
40	Religion	Jews, christians, jewish community
41	Cultural advisor	Scholar, academic, administrator, imperial war museum, cultural organization
42	Fish farmer	Fish, salmon, farmed fish, farmed salmon, marine harvester
43	Dancer	Dancer, choreographer, professional dancer, young dancer
44	LGBT	Gay, lesbian, bisexual, LGBT community, LGBT people
45	Patient with eye problem	Ophtalmologist, optometrist, optician, moorfields eye hospital
46	Patient with diabetes	People with diabetes, people with type 2 diabetes

Future research and limitations

There is a large literature that offers a global view of science seen from the perspective of the knowledge produced, or the topics addressed by researchers (Leydesdorff et al. 2013; Moya-Anegon et al. 2007).

However, we know that there is no one-to-one relation between topics of research and societal impact. As shown clearly by the pioneering analysis of King's College and Digital Science (2015), the relation between scientific disciplines and societal impact is best represented by an alluvial diagram, with many-to-many flows, rather than by an ordered and patterned relationship.

It is therefore useful to explore the possibility to build up new types of maps, in which nodes are social groups of users of research, and arc are proportional to the co-occurrence of these users in the same document. This would give countries, regions, or individual institutions a view on the scope and depth of impact, at a granular level.

It turns out that such undertaking must address the challenge of developing a nomenclature and a classification of social groups at the same level of completeness and granularity than existing classifications in science and technology, in order to normalize the measures. Text mining is a promising approach, but it does not offer per se the ground for normalization. We suggest a specific approach to text mining based on dedicated lexicons, or dictionaries that saturate specific semantic fields. This approach does not ensure normalization in the statistical meaning (i.e. an official procedure by independent authorities based on an extensive survey or a census, or a procedure largely agreed by relevant communities), but is able to provide normalization in the semantic sense, that is, saturation of the field.

With this novel approach, we are able to extract all expressions in the REF impact case studies that refer to social groups of users and to examine them at the desired level of granularity. On the basis of co-occurrence at document level of these extracted items we can follow the procedure recommended by Börner et al. (2003) and build up maps.

What is the contribution of user mapping to policy making and the literature on S&T systems?

First, policy makers and funding agencies might be interested in producing maps of the societal impact of the research they fund. Instead of describing the impact in terms of goals or research results, it might be illuminating to describe the impact in terms of concrete, observable social groups. In due time, policy discussions about research priorities and responsibility of research with respect to the needs of social groups may benefit from the dynamic comparison of maps across years.

Second, the maps can be drawn at university level. Each university might be interested in visualizing the social groups that are affected more strongly from its research. Given that there is no one-to-one correspondence between disciplines and user groups, as discussed above, user maps may deliver pictures that to do not overlap with existing disciplinary specializations. External communication of universities, in terms of third mission or social responsibility, might benefit from a compact and friendly visual tool. These maps might be a nice counterpart of university scientific research profiles, addressing the "for whom" question of the institution. Maps can be generated at various levels of granularity.

Third, in this paper we have not made use of the REF scores to the impact case studies, so that our user groups are weighted only by the frequency they are mentioned, not the score assigned by experts. Weighting the clusters of users with the average score

assigned by the REF assessment might generate a visual representation not only of the scope, but also of the strength, of societal impact of universities.

Finally, there is an ongoing debate in S&T regarding the relation between research excellence and societal relevance. An interesting way to address this issue is to examine the matching, or mismatch, between indicators of excellence of scientific areas and indicators of impact on social groups.

We are currently working along all these directions. This approach has some limitations, however. It is based on self-declared reconstructions of the impact of research. The authors of the impact case studies might be researchers themselves or consultants hired by universities and departments. It might be that the description of users is overemphasized. From a technical point of view, a lexicon cannot be easily validated with classical recall and precision measures. To address this limitation we open the full collection of impact case studies to interested readers, in order to improve the procedure.

We believe there is an increasing disparity between the pressure for demonstration of impact from governments, funding agencies, and the public opinion, and the current state of the art of approaches and methodologies. We see the lexicon-based text mining approach as complementary to other methodologies and a useful contribution to the advancement of the field.

Funding Open access funding provided by Università di Pisa within the CRUI-CARE Agreement.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Appendix

The present section shows the approach used to extract the users of the invention described in a patent. The full scale NLP methodology is described in detail in Chiarello et al. (2018). The proposed process is shown in Fig. 1 and its phases are as follows.

1. List of users generation: generation of an input list of users with the largest possible coverage.
2. Patent set selection: selection of a patent set to process.
3. Patent text pre-processing: application of NLP tools on the patents with the aim of preparing them for the automatic analysis process.
4. Automatic patent set annotation 1: projection of the input list of users on the text to generate the Automatically Annotated Patent Set 1.
5. Relevant sentences extraction: selection of sentences containing at least one user to generate an informative training set.
6. Automatic patent set annotation 2: generation of a statistical model by a machine learning algorithm and automatic tagging the of patent set exploiting the generated statistical model to generate the Automatically Annotated Patent Set 2.

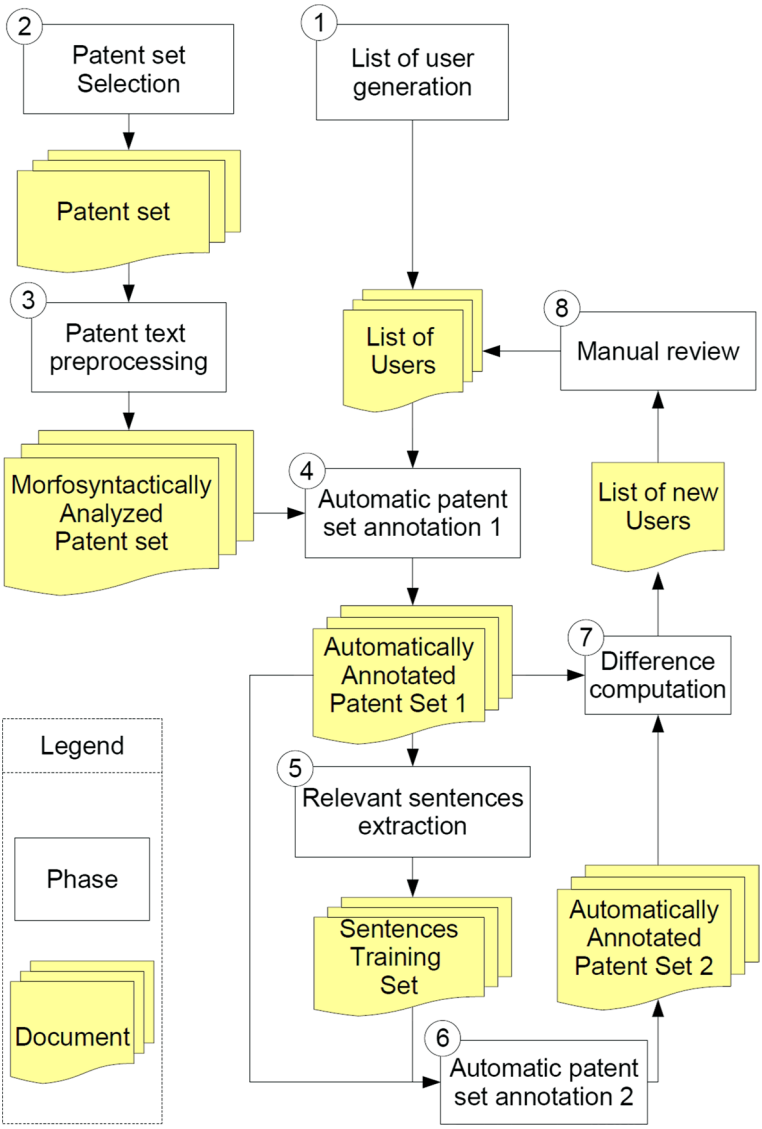

Fig. 6 Process flow diagram of the automatic user extraction system from patents. The diagram contains the representation of the documents and the operations performed on them. The process takes in input a patent set and a list of users and produces as output a list of new users

7. Difference computation: generation of the new list of users by computing the difference between the list of users found in the automatically annotated patent set 1 and the automatically annotated patent set 2.
8. Manual review: manual selection of the entities that, in the new list of users, are effectively users. This new list will enrich the original one.
9. Entity extraction tools used in patent analysis are largely based on NLP tools which can be applied to the analyzed text to extract entities that are important for the extraction

objective. The Named Entity Recognition is the task of identifying entity names like people, organizations, places, temporal expressions or numerical expressions.

There are several methods and algorithms to deal with the entity extraction task, but the most effective are the ones based on supervised methods. Supervised methods tackle this task by extracting relevant statistics from an annotated corpus. These statistics are collected from the computation of features values, which are strong indicators for the identification of entities in the analyzed text (Fig. 6).

Features used in NLP based entity recognition systems are divided in two main categories:

1. linguistically motivated features, such as n-grams of words, lemma and part of speech;
2. external resources features as, for example, external lists of entities that are candidates to be classified in the extraction process.

The annotation methods of a training corpus can be of two different kinds:

(a) human based, which is time expensive, but usually effective in the classification phase;
(b) automatically based, which can lead to annotation errors due to language ambiguity. For instance a driver can be classified both as a user (the operator of a motor vehicle), or not a user (a program that determines how a computer will communicate with a peripheral device).

Various training algorithms, such as hidden markov models (Eddy 1996), neural networks (Haykin 2009), Conditional Random Fields (CRF) (Lafferty et al. 2001) or Support Vector Machines (SVM) (Hearst et al. 1998) are used to build a statistical model based of features that are extracted from the analyzed documents in the training phase. In the recent yeas, the latest model of deep learning (i.e. Recurrent neural networks and Long short-term memory networks) has proven to outperform in the task of entity recognization and extraction (Hammerton 2003). Furthermore, new techniques of language representation such as contextualized vector representation (Peters et al. 2018) furtherly increased the accuracy of Named Entity Retrieval for standard entities (e.g. cities, dates, product names) in standard domains (social media, newspapers).

In the construction of external resources with respect to users there are two possible approaches. The first is to use existing classifications that are consistent with a definition of users as relevant for patent information, or more generally for technological information. A natural candidate is the classification of occupational categories. It is easy to see that the usefulness of inventions as described in patents may depend on the type of job. This approach has been followed by Pretiuc-Pietro et al. (2015) in classifying Twitter users according to the UK Standard Occupational Classification (SOC) and analyzing the content of their social communication in order to infer their income. The SOC classification has also been used by Sloan et al. (2015) and compared with Census and Twitter data. A crucial feature of occupational classification is that they are hierarchical: each person receives only one membership in a category and all categories are organised in a tree-like structure.

The second approach is to extract users from the text, in particular from social media. Beller and Van Durme (2014) tried to extract social roles from Twitter using heuristic methods. The authors identified all words preceded by constructions such as “I am” and

variations. This resulted in 63.858 unique roles identified, of which 44.260 appeared only once. It must be said that only a very small fraction of the extracted words corresponds to the definition of social roles. Some of the extracted entities are consistent with our definition (for example, doctor, teacher, mother or Christian). Overall, the procedure was considered too noisy.

Beller and Van Durme (2014) identify social roles in Twitter by assuming that they are associated to sets of verbs in social media communication. In order to clean the pool of identified users the authors crowd-sourced a manual verification procedure using the Mechanical Turk platform. Among the identified users we find artist, athlete, blogger, cheerleader, dee-jay and filmmaker.

We followed a third approach, which can be defined "hybrid". We developed a methodology that combines all classifications available in the open literature and in official statistics with a state-of-the-art computational linguistics procedure aimed at extracting user information from patents described in Chiarello et al. (2018).

The input list of users was obtained by collecting information from heterogeneous sources. Starting from the definition of user it is possible to elaborate its declinations.

To generate the list of users, we used two different approaches: the first bottom-up and second top-down. The bottom-up approach is based on merging together the following lists of entities:

- Lists of jobs: obtained by using U.S. Department of Labor (1981). Such list was merged with more recent lists[7] collecting a total of 11.142 users
- Lists of sports and hobbies: obtained by the union of lists[8] for a total of 9.660 users
- List of animals: obtained by parsing a web-page[9] for a total of 600 users
- Lists of patients: obtained by merging two web pages[10] for a total of 14.609 users
- List of generic words: manually generated. It contains users with an higher level of abstraction (such as person or human being), 56 users.

The top-down approach was then applied. Starting from the 35.767 users generated from the lists shown above, we then looked for alternative methods to indicate a user, finding defined word patterns. The most relevant are:

- Patterns like "hobby term + practitioner" for the hobbies
- Patterns like "person who has + disease term" or "suffering from + disease term" for the diseases
- Patterns like "practitioner of + sport term" for sports.

In the end of this process, a total of 76,857 users formed the knowledge base for the system, and gave us a reasonable number of terms representing potential users to be used in the next step of the process.

Obviously our lists have a limited coverage with respect to the entities that can be considered users. For instance, the lists miss some users of the classes mentioned above (e.g.

[7] http://www.careerplanner.com/DOTindex.cfm.

[8] http://www.notsoboringlife.com/list-of-hobbies/, http://discoverahobby.com/listofhobbies.

[9] http://a-z-animals.com/animals/.

[10] http://www.medicinenet.com/diseases, http://www.conditions.com/alphaa.html; http://www.cdc.gov/DiseasesConditions/az/a.html.

new jobs emerged in the last years) and all the alternative ways of referring to a user we do not spotted in the top-down approach. For example our lists miss jobs like *prostitute*, *lap dancer*, *undertaker*, *mortician* and *thief* or patients like *work-alcoholic* and *web-addicted*. Such terms have not been introduced in the input list because we considered these terms as candidates to be extracted by the process in our case study.

References

Adam, P., Ovseiko, P. V., Grant, J., et al. (2018). ISRIA statement: ten-point guidelines for an effective process of research impact assessment. *Health Reserch and Policy Systems, 16,* 8.

Adams, J., Loach, T., & Szomszor, M. (2015). *The diversity of UK research and knowledge. Analyses from the REF impact case studies.* Digital Research Reports.

Atkinson, P. M. (2014). Assess the real cost of research assessment: the research excellence framework keeps UK science sharp, but the process is overly burdensome for institutions, says Peter M. Atkinson. *Nature, 516*(7530), 145–146.

Baccianella, S., Esuli, A., & Sebastiani F. (2010). SentiWordNet 3.0. An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the 7th international conference on language resources and evaluation (LREC-10)*, 2200-2204.

Barbosa, L., & Feng J. (2010). Robust sentiment detection on Twitter from biased and noisy data. *Coling 2010 poster volume* 36–44. Bejing, August 2010.

Bell, S., et al. (2011). Real-world approaches to assessing the impact of environmental research on policy. *Research Evaluation, 20*(3), 227–237.

Beller, C. Harman, & Van Durme, B. (2014). Predicting fine-grained social roles with selectional preferences. *Association of Computational Linguistics, 2*(2014), 50.

Blei, D. M. (2012). Probabilistic topic models. *Communications of ACM, 55*(4), 77–84.

Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd international conference on machine learning*.

Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *8*(10), 10008.

Bollen, J., Van de Sompel, H., Hagberg, A., Bettencourt, L., Chute, R., Rodriguez, M. A., et al. (2009). Clickstream data yields high-resolution maps of science. *PLoS ONE, 4*(3), e4803.

Bonaccorsi, A., Chiarello, F., Fantoni, G., & D'Amico, L. (2017). Mapping users in patents. Steps towards a new methodology and the definition of a research agenda. In *Paper presented to the EPIP conference, Bordeaux, September*.

Börner, K., Chen, C., & Boyack, K. W. (2003). Visualizing knowledge domains. *Annual Review of Information Science and Technology, 37,* 179–255.

Bornmann, L. (2013). What is societal impact of research and how can it be assessed? A literature survey. *Journal of the Association for Information Science and Technology, 64*(2), 217–233.

Bornmann, L. (2014). Do altmetrics point to the broader impact of research? An overview of benefits and disadvantages of altmetrics. *Journal of informetrics, 8*(4), 895–903.

Bornmann, L., Haunschild, R., & Marx, W. (2016). Policy documents as sources for measuring societal impact: how often is climate change research mentioned in policy-related documents? *Scientometrics, 109*(3), 1477–1495.

Bornmann, L., & Marx, W. (2014). How should the societal impact of research be generated and measured? A proposal for a simple and practicable approach to allow interdisciplinary comparisons. *Scientometrics, 98*(1), 211–219.

Boyack, K. W., Klavans, R., & Börner, K. (2005). Mapping the backbone of science. *Scientometrics, 64,* 351–374.

Bozeman, B., & Sarewitz, D. (2011). Public value mapping and science policy evaluation. *Minerva, 49*(1), 1–23.

Callon, M. (1983). From translations to problematic networks: An introduction to co-word analysis. *Social Science Information, 22*(2), 191–235.

Callon, M., & Courtial, J.-P. (1989). *Co-word analysis: A tool for the evaluation of public research policy*. Paris: Ecole Nationale Superieure des Mines.

Callon, M., Law, J., & Rip, A. (Eds.). (1986). *Mapping the dynamics of science and technology*. London: Macmillan.

Carley, S., Porter, A. L., Rafols, I., & Leydesdorff, L. (2017). Visualization of disciplinary profiles. Enhanced science overlay maps. *Journal of Data and Information Science, 2*(3), 68–111.
Chao, A. F. Y., & Yang, H. (2018). Using Chinese radical parts for sentiment analysis and domain-dependent seed set extraction. *Computer Speech & Language, 47,* 194–213.
Chen, H., Zhang, Y., Zhang, G., Zhu, D., & Lu, J. (2015). Modeling technological topic changes in patent claims. In *Proceedings of the PICMET'15 conference*.
Chiarello, F., Cimino, A., Fantoni, G., & Dell'Orletta, F. (2018). Automatic users extraction from patents. *World Patent Information, 54,* 28–38.
Chiarello, F., Fantoni, G., & Bonaccorsi, A. (2017). Product description in terms of advantages and drawbacks. Exploiting patent information in novel ways. In *Paper presented to the ICED conference 2017*.
Chiarello, F., Bonaccorsi, A., & Fantoni, G. (2020). Technical sentiment analysis. Measuring advantages and drawbacks of new products using social media. *Computers in Industry*, *123*, 103299.
Chinsha, T. C., & Joseph, S. (2015). A syntactic approach for aspect based opinion mining. In *IEEE international conference on semantic computing (ICSC)* (pp 24–33).
Colinet, L., Joly, P.-B., Gaunand, A., Matt, M., Larédo, P., Lemarié, S. (2014). *ASIRPA. Analyse des impact de la recherche publique agronomique*. Rapport final, Rapport préparé pour l'Inra, Paris, France.
Cozzens, S. E., Bobb, K., & Bortagaray, I. (2002). Evaluating the distributional consequences of science and technology policies and programs. *Research Evaluation, 11*(2), 101–107.
Cronin, B. (1984). *The citation process. The role and significance of citations in scientific communication*. Oxford: Taylor Graham.
Cronin, B. (2005). *The hand of science. Academic writing and its rewards*. Lanham: The Scarecorw Press.
Dance, A. (2013). Impact: Pack a punch. *Nature, 502*(7471), 397–398.
De Jong, S., Barker, K., Cox, D., Sveinsdottir, T., & Van den Besselaar, P. (2014). Understanding societal impact through productive interactions: ICT research as a case. *Research Evaluation, 23*(2), 89–102.
De Jong, S. P., Van Arensbergen, P., Daemen, F., Van Der Meulen, B., & Van Den Besselaar, P. (2011). Evaluation of research in context: an approach and two cases. *Research Evaluation, 20*(1), 61–72.
Derrick, G. E. (2014). Intentions and strategies for evaluating the societal impact of research. Insights from REF 2014 evaluators. In *Proceedings of the ISSTI conference* (pp. 136–144).
Derrick, G. E., Meijer, I., & van Wijk, E. (2014). Unwrapping "impact" for evaluation: A co-word analysis of the UK REF2014 policy documents using VOSviewer. In *Proceedings of the science and technology indicators conference* (pp. 145–154).
Digital Science. (2015). *REF 2014 impact case studies and the BBSRC*. www.bbsrc.ac.uk/documents/1507-ref-impact-case-studies-pdf/. Accessed December 3, 2019.
Digital Science. (2016). *The societal and economic impacts of academic research. International perspectives on good practice and managing evidence*. Digital Research Reports, March.
Donovan, C. (2011). State of the art in assessing research impact: introduction to a special issue. *Research Evaluation, 20*(3), 175–179.
Eddy, S. R. (1996). Hidden markov models. *Current Opinion in Structural Biology, 6*(3), 361–365.
Esuli, A., & Sebastiani, F. (2006). SentiWordNet: A publicly available lexical resource for opinion mining. In *Proceedings of language resources and evaluation (LREC) conference.*
Esuli, A., & Sebastiani, F. (2010). Sentiment quantification. *IEEE Intelligent Systems, 25*(4), 72–75.
Grant, J., Brutscher, P. B., Kirk, S., Butler, L., & Wooding, S. (2010). *Capturing research impacts. A review of international practice.* Report prepared for the Higher Education Funding Council, Cambridge, Rand Europe.
Greenhalgh, T., Raftery, J., Hanney, S., & Glover, M. (2016). Research impact: A narrative review. *BMC Medicine, 14,* 78.
Hammerton, J. (2003). Named entity recognition with long short-term memory. In *Proceedings of the seventh conference on natural language learning at HLT-NAACL 2003-volume 4* (pp. 172–175). Association for Computational Linguistics.
Haykin, S. (2009). *Neural networks. A comprehensive foundation*. New York, Prentice-Hall.
Hearst, M. A., Dumais, S. T., Osman, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and their Applications, 13*(4), 18–28.
Hecking, T., & Leydesdorff, L. (2019). Can topic models be used in research evaluations? Reproducibility, validity, and reliability when compared to semantic maps. *Research Evaluation, 28*(3), 263–272.
Hemmatian, F., & Sohrabi, M. K. (2019). A survey on classification techniques for opinion mining and sentiment analysis. *Artificial Intelligence Review, 52*, 1495–1545.
Hinrichs, S., & Grant, J. (2015). A new resource for identifying and assessing the impacts of research. *BMC Medicine, 13,* 148.
Holbrook, J. B., & Frodeman, R. (2011). Peer review and the ex ante assessment of societal impacts. *Research Evaluation, 20*(3), 239–246.

Hu, M., & Liu, B. (2004a). Mining and summarizing customer reviews. In: *KDD'04 conference proceedings, Seattle, Washington, August 22–25*.

Jacobsson, S., & Perez, Vico E. (2010). Towards a systemic framework for capturing and explaining the effects of academic R&D. *Technology Analysis & Strategic Management, 22,* 765–787.

Jang, H. J., Sim, J., Lee, Y., & Kwon, O. (2013). Deep sentiment analysis. Mining the causality between personality-value-attitude for analysing business ads in social media. *Expert Systems with Applications, 40,* 7492–7503.

Jiménez-Zafra, S. M., Martín-Valdivia, M. T., Martínez-Cámara, E., & Ureña-López, L. A. (2015). Combining resources to improve unsupervised sentiment analysis at aspect-level. *Journal of Information Science, 42,* 213–229.

Joly, P. B., Gaunand, A., Colinet, L., Larédo, P., Lemarié, S., & Matt, M. (2015). ASIRPA: A comprehensive theory-based approach to assessing the societal impacts of a research organization. *Research Evaluation, 24,* 440–453.

Kanninen, S., & Lemola, T. (2006). Methods for evaluating the impact of basic research funding: An analysis of recent international evaluation activity. *Publications of the Academy of Finland*, *9*(06), 1–99.

Kay, L., Newman, N., Youtie, J., Porter, A. L., & Rafols, I. (2014). Patent overlay mapping. Visualizing technological distance. *Journal of the American Society for Information Science and Technology, 65*(12), 2432–2443.

Khazragui, H., & Hudson, J. (2015). Measuring the benefits of university research: impact and the REF in the UK. *Research Evaluation, 24*(1), 51–62.

King's College, Digital Science. (2015). *The nature, scale and beneficiaries of research impact. An initial analysis of REF (2014) impact case studies*. Research report 2015/01, London, HEFCE.

Klavans, R., & Boyack, K. W. (2009). Toward a consensus map of science. *Journal of the American Society for Information Science and Technology, 60*(3), 455–476.

Krücken, G., Meier, F., & Müller, A. (2009). Linkages to the civil society as 'leisure time activities'? Experiences at a German university. *Science and Public Policy, 36*(2), 139–144.

Langfeldt, L., & Scordato, L. (2015). *Assessing the broader impacts of research: A review of methods and practices.* NIFU working paper 8/2015.

Lavis, J. N., Robertson, D., Woodside, J. M., McLeod, C. B., & Abelson, J. (2003). How can research organizations more effectively transfer research knowledge to decision makers? *The Milbank Quarterly, 81*(2), 221–248.

Leckie, G. J., Pettigrew, K. E., & Sylvain, C. (1996). Modeling the information seeking of professionals. A general model derived from research on engineers, health care professionals, and lawyers. *Library Quarterly, 66*(2), 161–193.

Lee, S., Yoon, B., & Park, Y. (2009). An approach to discovering new technology opportunities: Keyword based patent map approach. *Technovation, 29*(6–7), 481–497.

Leydesdorff, L. (1989). Words and co-words as indicators of intellectual organization. *Research Policy, 18*(4), 209–223.

Leydesdorff, L., Carley, S., & Rafols, I. (2013). Global maps of science based on the new Web-of-Science categories. *Scientometrics, 94,* 589–593.

Leydesdorff, L., Kushnir, D., & Rafols, I. (2014). Interactive overlay maps for US patents (USPTO) data based on International Patent Classification (IPC). *Scientometrics, 98,* 1583–1599.

Leydesdorff, L., & Nerghes, A. (2017). Co-word maps and Topic Modeling: A comparison using small and medium-sized corpora (N < 1,000). *Journal of the American Association for Information Science and Technology, 68*(4), 1024–1035.

Leydesdorff, L., & Rafols, I. (2009). A global map of science based on the ISI subject categories. *Journal of the American Society for Information Science and Technology, 60*(2), 348–362.

Leydesdorff, L., & Rafols, I. (2012). Interactive overlays. A new method for generating global journal maps from Web-of-Science data. *Journal of Informetrics, 6,* 318–332.

Manning, C. D., Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, Mass.: The MIT press.

Martin, B. R. (2011). The Research Excellence Framework and the 'impact agenda': are we creating a Frankenstein monster? *Research Evaluation, 20*(3), 247–254.

Matt, M., Gaunand, A., Joly, P. B., & Colinet, L. (2017). Opening the black box of impact. Ideal-type impact pathways in a public agricultural research organization. *Research Policy, 46,* 207–218.

Lafferty J., McCallum, A., & Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th international conference on machine learning 2001 (ICML 2001)* (pp. 282–289).

Meyer, R. (2011). The public values failures of climate science in the US. *Minerva, 49*(1), 47–70.

Miettinen, R., Tuunainen, J., & Esko, T. (2015). Epistemological, artefactual and interactional. Institutional foundations of social impact of academic research. *Minerva, 53,* 257–277.
Mohammadi, E., Thelwall, M., Haustein, S., & Larivière, V. (2015). Who reads research articles? An altmetric analysis of Mendeley user categories. *Journal of the American Society for Information Science and Technology, 66*(9), 1832–1846.
Molas-Gallart, J., & Tang, P. (2011). Tracing 'productive interactions' to identify social impacts: an example from the social sciences. *Research Evaluation, 20*(3), 219–226.
Molina-González, M. D., Martínez-Cámara, E., Martín-Valdivia, M. T., & Ureña-López, L. A. (2015). A Spanish semantic orientation approach to domain adaptation for polarity classification. *Information Processing Management, 51*(4), 520–531.
Morgan, J. M., & Grant, J. (2013). Making the grade. Methodologies for assessing and evidencing research impact: 7 essays on impact. In J. Dean, et al. (Eds.), *DESCRIBE project report* (pp. 25–43). Exeter: University of Exeter Press.
Morton, S. (2015). Progressing research impact assessment: A 'contributions' approach. *Research Evaluation, 24*(4), 405–419.
Moya-Anegón, F., Vargas-Quesada, B., Chinchilla-Rodriguez, Z., Corera-Alvarez, E., Munoz-Fernández, F. J., & Herrero-Solana, V. (2007). Visualizing the marrow of science. *Journal of the American Society for Information Science and Technology, 58*(14), 2167–2179.
Moya-Anegon, F., Vargas-Quesada, B., Herrero-Solana, V., Chinchilla-Rodriguez, Z., Corera-Alvarez, E., & Munoz-Fernandez, F. J. (2004). A new technique for building maps of large scientific domains based on the cocitation of classes and categories. *Scientometrics, 61*(1), 129–145.
Muhonen, R., Benneworth, P., & Olmos-Peñuela, M. (2020). From productive interactions to impact pathways: Understanding the key dimensions in developing SSH research societal impact. *Research Evaluation, 29*(1), 1–14.
Mustafa, M. M. (2013). More than words. Social networks' text mining for consumer brand sentiments. *Expert Systems with Applications, 40,* 4241–4251.
Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes, 30*(1), 3–26.
Nutley, S., Walter, I., & Davies, H. T. (2003). From knowing to doing: a framework for understanding the evidence-into-practice agenda. *Evaluation, 9*(2), 125–148.
Nutley, S. M., Walter, I., & Davies, H. T. (2007). *Using evidence: How research can inform public services.* Bristol: Policy Press.
Pedersen, D. B., Grønvad, J., & Hvidtfeldt, R. (2020). Methods for mapping the impact of social sciences and humanities. A literature review. *Research Evaluation, 29*(1), 66–70.
Penfield, T., Baker, M. J., Scable, R., & Wykes, M. C. (2014). Assessment, evaluations, and definitions of research impact. A review. *Research Evaluation, 23*(1), 21–32.
Perez, Jacobsson S., Vico, E., & Hellsmark, H. (2014). The many ways of academic researchers. How is science made useful? *Science and Public Policy, 41*(5), 641–657.
Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., et al. (2018). *Deep contextualized word representations*. arXiv preprint arXiv:1802.05365.
Preţiuc-Pietro, D., Volkova, S., Lampos, V., Bachrach, Y., & Aletras, N. (2015). Studying user income through language, behaviour and affect in social media. *PLoS ONE, 10*(9), e0138717.
Rafols, I., Porter, A. L., & Leydesdorff, L. (2010). Science overlay maps. A new tool for research policy and library management. *Journal of the American Society for Information Science and Technology, 61*(9), 1871–1887.
Rathan, M., Hulipalled, V. R., Venugopal, K. R., & Patnaik, L. M. (2017). Consumer insight mining: aspect based Twitter opinion mining of mobile phone reviews. *Applied Soft Computing, 68,* 765–773.
Reale, E., et al. (2018). A review of the literature on evaluating the scientific, social and political impact of social sciences and humanities research. *Research Evaluation, 27*(4), 298–308.
REF. (2019). *Guidance on submission REF 2021*. https://www.ref.ac.uk/publications/guidance-on-submissions-201901/.
Rowe, G., & Frewer, L. (2005). A typology of public engagement mechanisms. *Science, Technology and Human Values, 30*(2), 251–290.
Samuel, G. N., & Derrick, G. E. (2015). Societal impact evaluation: Exploring evaluator perceptions of the characterization of impact under the REF2014. *Research Evaluation, 24*(3), 229–241.
Sarewitz, D., & Pielke, R. A. (2007). The neglected heart of science policy: reconciling supply and demand for science. *Environmental Science & Policy, 10,* 5–16.
Sloan, L., Morgan, J., Burnap, P., & Williams, M. (2015). Who tweets? Deriving the demographic characteristics of age, occupation and social class from twitter user meta-data. *PLoS ONE, 10*(3), e0115545.

Spaapen, J., & Van Drooge, L. (2011). Introducing 'productive interactions' in social impact assessment. *Research Evaluation, 20*(3), 211–218.

Tan, S., Li, Y., Sun, H., Guan, Z., Yan, X., Bu, J., et al. (2012). Interpreting the public sentiment variations on Twitter. *IEEE Transactions on Knowledge and Data Engineering, 6*(1), 1–14.

Turney, P. D., & Littman, M. L. (2003). Measuring praise and criticism. Inference of semantic orientation from association. *ACM Transactions on Information Systems, 21*(4), 315–346.

U.S. Department of Labor. (1981). *Check suffix codes for jobs defined in the dictionary of occupational titles* (3rd ed.). Washington: United States Employment Service, U.S. Dept. of Labor.

Zhang, L., Riddhiman, G., Dekhil, M., Hsu, M., & Liu, B. (2011). *Combining lexicon-based and learning-based methods for Twitter sentiment analysis*. HP laboratories working paper 2011-89.

Zhou, F., Jiao, J. R., Yang, X. J., & Lei, B. (2017). Augmenting feature model through customer preference mining by hybrid sentiment analysis. *Expert Systems with Applications, 89,* 306–317.