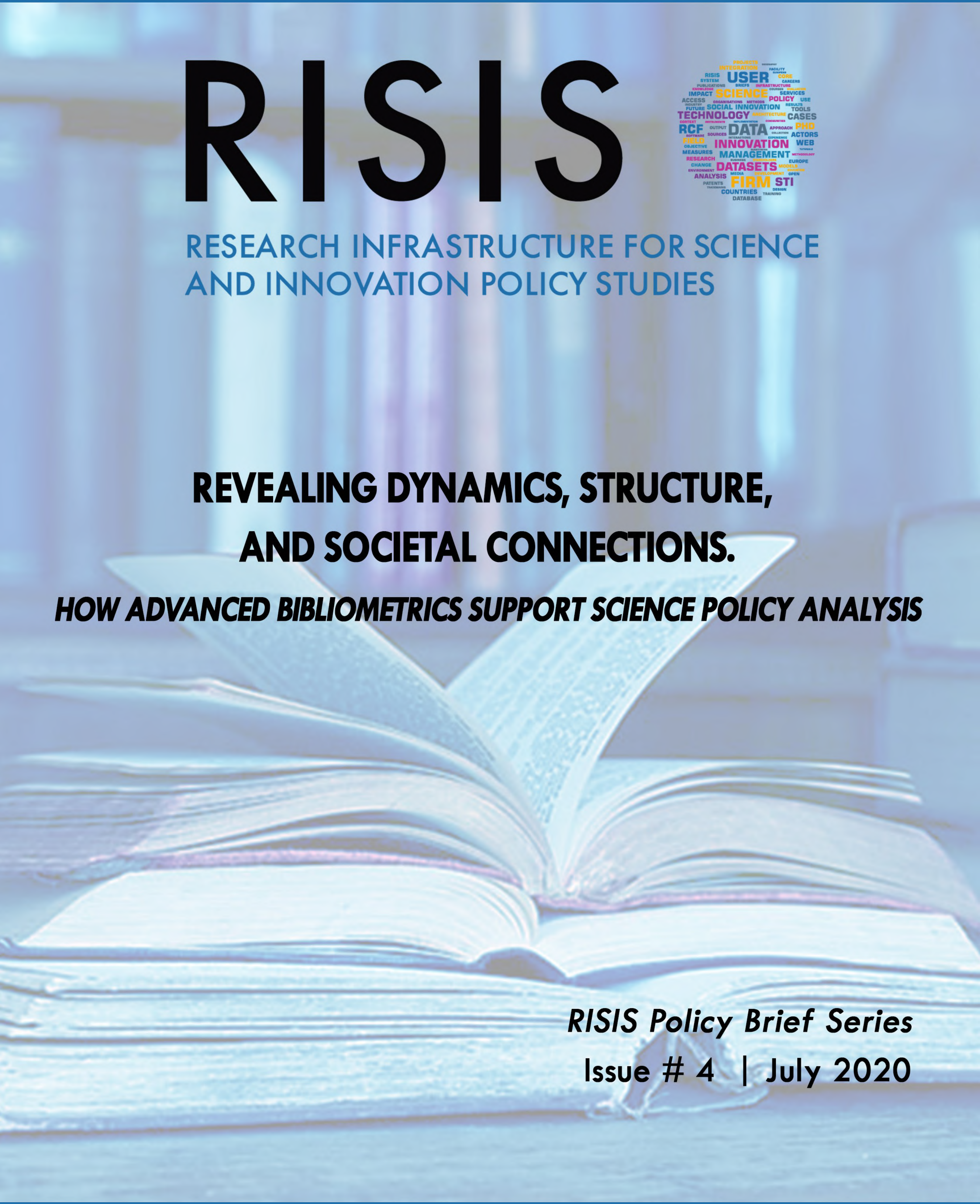


RISIS

RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES

REVEALING DYNAMICS, STRUCTURE, AND SOCIETAL CONNECTIONS.

HOW ADVANCED BIBLIOMETRICS SUPPORT SCIENCE POLICY ANALYSIS

RISIS Policy Brief Series

Issue # 4 | July 2020

www.risis2.eu
risis.communication@ircres.cnr.it

CONTENTS

CWTS, Leiden University's study aims to demonstrate the added value of advanced bibliometrics analyses in the context of **science policy**, beyond its conventional uses in **research evaluation.**

The analyses takes into consideration the example of cancer research to illustrate the potential of bibliometric data to reveal information beyond counting publications and citations, and specifically to gain insights into the dynamics of a research field or organisation, its structure and its different ways in which research is connected to societal processes. The bibliometric analyses have been conducted using the **RISIS CWTS publication Dataset.**

The power of advanced bibliometrics is its capacity to analyse and contextualize a research group, a university, a discipline, regions, countries or group of countries. Thus, the study encourages science policy analysists to make use of advanced scientometric techniques to support decision making.

In addition, it will show how bibliometrics can support benchmarking of countries and types of cancer across a variety of properties – not only number of publications and citations, but also the extent to which these publications are mentioned in social media, are cited in patents, are produced in local hospitals, in local languages (non-English), in collaboration with industry, etc. In the area of cancer research, this multidimensional perspective could help science policy analysts to identify specific research areas which might be candidates to be funded and/or to better understand and reflect on how the cancer mission evolves across Europe.

1. INTRODUCTION

In the last two decades, the success and institutionalisation of bibliometrics (i.e. methods based on the statistical analysis of scientific publications and citations) for the evaluation of research performance has resulted in some misuses and abuses which may have led to poor views of the analytical potential of bibliometrics. This is unfortunate, because while simplistic bibliometrics is sometimes problematic due to the use of mono-dimensional data (e.g. only citations) and lack of contextualisation (how should they be compared?), the power of advanced bibliometrics lies precisely in its capacity to contextualisation by using diverse types of information.

By contextualisation, it is meant that advanced bibliometrics allows to make comparison relevant to policy questions in terms of dynamics (i.e. over time), structure (e.g. over cognitive networks) and on connections (e.g. over collaborations or flows). By diverse data sources, it is meant that advanced scientometrics provides not only counts of publications and citations, but also a variety of meta-data associated with publications – including authors, organizations, locations, collaborations, and mentions of the publications in non-scientific spheres such as social media or policy documents, among others.

CWTS, Leiden University's study uses as illustration research on cancer. **Cancer** is one of the five areas selected by the European Commission as **mission-areas for Horizon Europe.** There are many potential goals to try achieve **progress in cancer research.** What type of research should be supported under the cancer mission-area? Cancer is a collection of diseases with different burden. Which cancers should be prioritised? Cancer research includes investigations on the molecular mechanisms, the socio-environmental determinants, prevention and diagnosis, and clinical interventions. Which type of research is more important **to improve health in the mid - and long-term?** Cancer research is carried out in universities, hospitals, small biotech and large pharmaceutical companies. To which extent should these organisations collaborate or take complementary roles?

Advanced bibliometrics does not provide direct answers to the questions posed above, but it provides the **contextual information for analysts to reflect on overall cancer research portfolios to seek a given strategy.** For example, it can show whether research on prevention is relatively underfunded as is often claimed, which types of cancers received relatively more **EC funding,** or whether in certain countries **industry** is more or less engaged on specific types of cancer. The data needs to be interpreted with caution because a large part of **health R&D** is carried out in companies or clinical research organizations where publishing is a secondary practice.

2. METHODOLOGY AND DATA: RISIS CWTS PUBLICATION DATASET

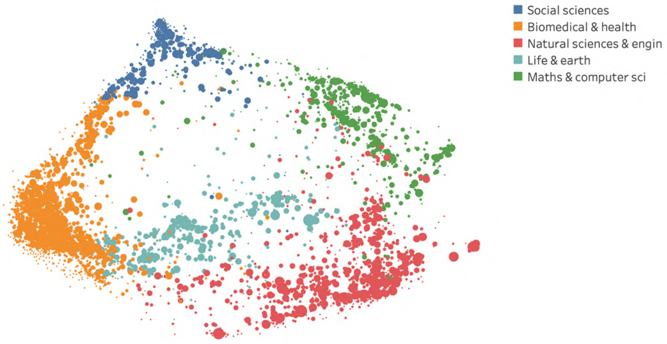
The bibliometric analyses have been conducted using the **RISIS CWTS Publication Dataset**. It is an enhanced version of the **Web of Science database**, developed for bibliometric purposes. The enhancement regards, among others, the following elements:

- **Author affiliation** – unified and harmonized names of organizations and a permanent identifier shared with other **RISIS data sources via OrgReg and FirmReg**;
- **Funding organizations** – unified and harmonized names of funding organizations mentioned in publication acknowledgements;
- **Publication-level classification** – individual publications are clustered into groups based on citation relations, irrespective of the journal in which they are published.

CWTS Publication level classification

The CWTS publication classification algorithm clusters individual publications into groups on the basis of direct citation relations. Around **30 million publications**, with over **500 million citation relations** among each other are the input for this in-house developed clustering method (Traag et al., 2019) to yield coherent clusters and do so at different levels of abstraction, from a clustering in **4013 micro-fields** to a clustering in just **5 large scientific domains.** Together these clusters form a self-organized structure of all sciences (as far as covered by Web of Science). It should be noted that publications are clustered regardless of the journal in which they are published. Hence, a publication in a multi-disciplinary journal like *Nature* may be clustered together with a publication in a specialized journal, such as the *Journal of Neurophysiology*.

Figure 1. Map of all science, Web of Science (2000-2019)



In this study, **CWTS, Leinden University** used the most fine-grained classification level that groups all publications into **4013 clusters.** A representation of this configuration of science can be found in Figure 1. The circles in the map visualize the **4013 publication clusters** in which the distance between clusters represent their cognitive relatedness in terms of citation traffic between them. The denser the traffic, the closer they are. The color-coding represents the very top-level structure of science organized in **5 scientific domains.** The size of the circles indicates the number of publications included in the cluster.

Measuring knowledge production at country level

For the research output CWTS, Leinden University measured a country's contribution by dividing each publication by the number of different countries involved. Thus, a publication involving two countries adds 0.5 for each to its total. In this way, it has been avoided a bias towards research fields where often many authors are involved, adding 1 to each country involved.

Selecting cancer publications

For this study **26 different types of cancer** with the highest burden of disease worldwide according to the World Health Organization have been selected. CWTS team collected all scientific publications related to each of these diseases from MEDLINE selecting the descriptor(s) (MeSH terms) that best represent the various types of cancer.

Burden of disease

The burden of disease is used to assess the relative **importance of diseases.** The researchers rely on disability-adjusted life-year (DALYs), which combines in a single measure the number of years of life lost from early death and years of life lived in less than full health. This information is interpreted as a proxy of unmet medical needs in society, which in some cases depend on scientific and technological developments to be alleviated. The researchers obtained disease burden statistics from the **World Health Organization.** For this study it has been used the burden of disease corresponding to the year 2015. Both the collection of publications and the link between publications and burden or disease is documented (Yegros Yegros et al., 2020).

Area based connectedness to society

Area-based Connectedness (ABC) is an approach to analyse the ways in which **research output is associated to societal processes beyond the science system** (Noyons, 2019). These dimensions of connectedness are based on signals between research output (publications) and society or traces within publications themselves. Examples of such signals are **a publication being cited** by a **policy document or in a newspaper** (it is a signal sent from policy or the general public/media). An example of traces in publications is co-authorship with industry authors or funding acknowledgements (it is a trace of industry involvement in the publication itself).

In the ABC approach the object of analysis is not individual papers or authors but rather **research areas,** identified through the above clustering method. Publications are thus part of a research area and their connectedness is measured across the entire area since knowledge production is a collective process of scientific communities. A citation to a single publication is thus recognized as a signal that the research area is of policy importance but not that single publication is more policy-relevant than publications closely related to it.

The connectedness of a research area is defined by the proportion of publications within it that show particular signals or traces. Thus, **each area has its own connectedness-level** normalized by its size. If an area has 1000 publications and 20 of them are cited in a policy document, this area has an ABC-policy score of 0.02 (or 2%).

Measuring the connectedness to society, for instance links with policy and industry, for each research area in this way has two important implications for science policy and research evaluation. First, the ABC analysis shows that different research **areas are differently connected to society.** Knowledge in a less connected areas might primarily be of influence in society through being cited in publications in more connected areas. Second, for research evaluation and new science policy initiatives, it is crucial **to understand and evaluate knowledge production and knowledge producers** within the context of the research area they are part off. It might be unrealistic to expect particular forms of connectedness from researchers and institutes that contribute to areas that are not in this form connected to societal processes.

In the cognitive map below, we visualize the ABC-policy score of all research areas in the map of all science. The configuration is the same as the map in Figure 1, while the color-coding indicates in which areas the policy connectedness is denser (dark Blue). The map shows that **high policy connectedness is primarily found in the social sciences, biomedical sciences and life and earth sciences.** Figure 2, depicts industry co-authorship, showing the connectedness primarily in **engineering and computer science and pharma.**

Figure 2a. Distribution of Area Based Connectedness policy score across map of all science

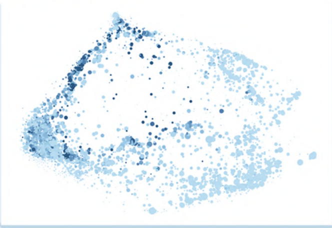
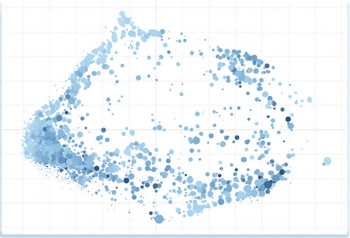


Figure 2b. Distribution of Area Based Connectedness to industry as shown by co-authorship across map of all science



Covering several ABC dimensions, CWTS Leiden University's study calculated various measures and used them as proxies of different connectedness pathways between **scientific research and the broader society.**

3. INSIGHTS FOR RESEARCH AND INNOVATION POLICY

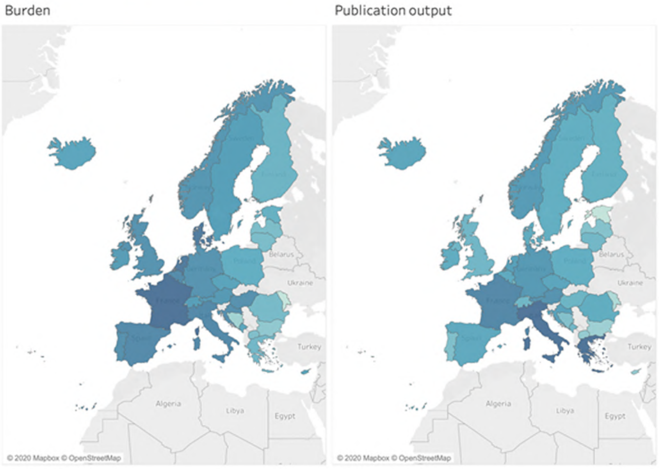
RISIS CWTS publication Dataset can be used to provide valuable insights for **Research and Innovation policy.** This Policy brief takes into consideration two options.

Disease burden vs research efforts

A first option in thinking about research priorities in cancer is to compare the **relative disease burden** caused by cancer(s) to the relative research efforts in cancer(s). One should not expect a linear relationship between health needs (burden) and research efforts – but the comparison may help think experts if more research in needed in some countries or cancers.

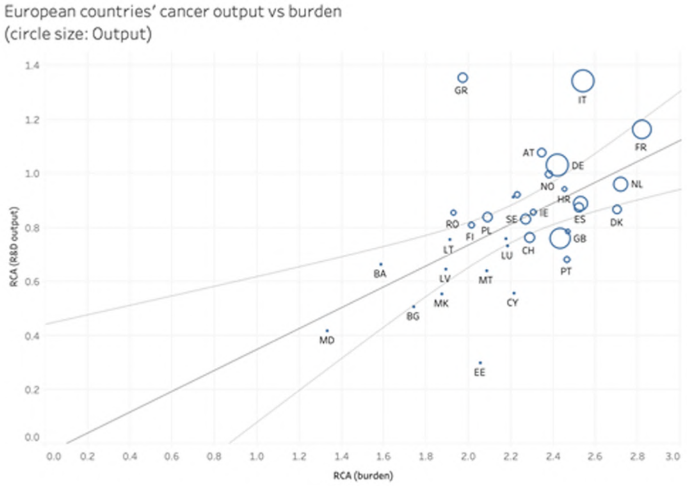
To connect knowledge production to health needs it has been linked the burden statistics of countries for a certain disease to the research output on that disease. For cancer it has been calculated the relative shares. Output contributions as well as burden per country are normalized by total output/burden of a country and total output/burden on a topic (similar to the Revealed Comparative Advantage (RCA)). All measures are calculated for the world and all entire biomedical and health topics. The results are discussed in the European context only. The results are visualized in a geographical map. The differences between **the maps provide a first overview** on the (mi-s)match between disease burden and knowledge production (Figure 3).

Figure 3. Cancer burden vs publication output in Europe



The study evaluated the **correlation between research efforts and burden of disease in European countries** to find out about potential mismatches, i.e., relatively low knowledge production with a high disease burden or vice versa. The results are depicted in Figure 4. This graph shows a strong correlation between burden and knowledge production at the country level. In general, **the higher the overall disease burden the more knowledge is produced regarding cancer**. However, some countries appear to show certain degree of mismatch but an underperformance in terms of output with regard to burden is hardly observed. The other way around, i.e., examples of over performance as compared to their burden include countries such as Italy and Greece.

Figure 4: Disease burden vs research output by country



By performing the same analysis for individual types of cancer it is possible to achieve a higher level of detail. This allows to observe **whether research efforts of countries are in line, above or below the level of burden corresponding to specific types of cancer.**

Exploring missions and social contexts of research

Another analytical option is to explore **how the publications are connected to social spheres,** which can particularly illuminate how research related to, for instance, local audiences and policy. To do this, the study analyses meta-data on publications as follows.

In the context of cancer research, CWTS, Leiden University analysed the extent to which **research funded by the EU is more oriented to support clinical research.** The researchers consider publications acknowledging funding from the EU as reporting research financially supported by the EU. While the involvement of non-academic hospitals has been considered a proxy of clinical research.

It appears that the EU tends to invest more in non-clinical research. This relationship is also found when analysing the relationship between EU funding and the extent to which research is directed at domestic or local audiences. Research published for a local audience reports less often having received funding from the EU.

Finally, CWTS, Leiden University measured the relationship between **EU funding and industry involvement in performing scientific research.** As for industry involvement it has been considered publications (co)produced by business companies a proxy of research performed by industry. In this case, there is a positive correlation, i.e., **the more EU funding the more industry involvement.** It should be noted that these three correlations stand by themselves. Increased industry involvement does not mean the research is more basic, nor is there any correlation between hospital involvement and industry involvement.

4. LEARNING FROM MULTIDIMENSIONAL ANALYSES OF PUBLICATIONS

Bibliometric analyses can provide a rich contextualization of knowledge production. This contextualisation can be very valuable in thinking about science policy, not only for ex-post evaluation but also for **strategic purposes,** such as on priority setting, in particular for **mission-oriented research programmes** in which societal goals are key policy objectives.

RISIS has the potential to link the publications with various datasets in various ways, including patents (CIB), EU-projects (EUPRO). These possibilities provide an even wider potential of policy relevant studies. The KNOWMAK project is an excellent example.

In this study CWTS, Leiden University have reported a few assets of an advanced publication data infrastructure to characterize and represent research output according to various relevant dimensions. The survey illustrates the added **value of publications meta-data that show some of the links to society.** Besides these linkages between research and society, the researchers consider the **area-based connectedness is a key conceptual contribution.** This approach allows to capture the research communities (rather than the individual actors) that develop **knowledge on a certain topic or theme that is linked to societal challenges, mission.** In this way a more contextual view can be obtained on, for instance, on the rationales and potential outcomes of **funding programs or research policies.**

It is relevant to notice that this approach to bibliometrics does not aim to provide performance analysis (success or failure) of programmes, but aims to characterise and provide **contextualised information on the research environment or potential goals of science policies.**

For example, by showing what types of cancer research are being conducted before and after a programme (e.g. more basic, more clinical?; in collaboration or funded by industry?) The study analysed cancer research as a whole. By differentiating types of cancer, it is possible to gain more insights of relevance to science policy. Moreover, additional dimensions could be considered such as gender information to the publication, author information, for instance, to address the issue of diversity, or analyse research done around cancer in relation to the gender dimension.

REFERENCES

LINK TO OTHER SOURCES

The publication dataset is linked to external data at the level of publications. By the harmonized list of author affiliations and permanent identifier (via OrgReg and FirmReg). However, RISIS has the potential to link these publications with various datasets in various ways, including patents (CIB), EU-projects (EUPRO). These possibilities provide an even wider potential of policy relevant studies. The KNOWMAK project is an excellent example.

REFERENCES

Noyons, E. (2019). Measuring Societal Impact Is as Complex as ABC. Journal of Data and Information Science, 4(3), 6–21. https://doi.org/10.2478/jdis-2019-0012

Traag, V. A., Waltman, L., & van Eck, N. J. (2019). From Louvain to Leiden: Guaranteeing well-connected communities. Scientific Reports, 9.
https://doi.org/10.1038/s41598-019-41695-z

Yegros-Yegros, A., van de Klippe, W., Abad-Garcia, M. F., & Rafols, I. (2020). Exploring why global health needs are unmet by research efforts: The potential influences of geography, industry and publication incentives. Health Research Policy and Systems, 18(1), 47.
https://doi.org/10.1186/s12961-020-00560-6

RISIS2 - European Research Infrastructure for Science, technology and Innovation policy Studies aims at building a data and services infrastructure supporting the development of a new generation of analyses and indicators on STI fields.

To develop a deeper understanding of knowledge dynamics and policy relevant evidence, RISIS goes beyond established quantitative indicators, developing positioning indicators, in order to reduce asymmetries in actors producing new knowledge, in places where knowledge is generated, and in themes addressed.

RISIS community is dealing with sensitive issues as social innovation, non-technological innovation, the role of PhDs in society, and portfolios of public funding instruments, studying both universities and firms.

This project is funded by the European Union
under Horizon2020 Research and Innovation
Programme Grant Agreement n°824091

***RISIS Policy Brief Series** aim at disseminating key results coming from RISIS2 to improve the use of data for evidence-based policy making. The outcomes are presented through short documents pointing out the main policy issues at stake, demonstrating the contribution provided by RISIS, and what new avenues for research are now open.*

AUTHORS OF THE CURRENT ISSUE:
Ed Noyons
Afredo Yegros-Yegros
Thomas Franssen
Ismael Rafol
CWTS, Leiden University, the Netherlands

GRAPHIC DESIGN:
Serena Fabrizio | CNR

www.risis2.eu