DELIVERABLE

Can web scraped data inform innovation policy?

Experience from BIGPROD project

Summary

While there has been increased interest on the utilization of big data and data analytics in public policymaking, policy analytics has not been widely adopted. This has been the result of a mismatch with expectations and capabilities. This policy brief reports reflect on the findings of a 2.5 year project gathering micro-level data from approximately 96 000 companies using web scraping. The report highlights challenges and offers recommendations based on the learnings from the project.

Deliverable Information

Deliverable number and name:	**Policy Brief 3**
Due date:	28 Feb 2022
Deliverable:	D34
Work Package:	WP5
Lead Partner for the Deliverable:	UNU-MERIT
Author:	Arho Suominen
Reviewers:	Hugo Hollanders
Approved by:	Arho Suominen
Dissemination level:	Public
Version	12th April 2022

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870822

Disclaimer

This document contains a description of the **BIGPROD** project findings, work and products. Certain parts of it might be under partner Intellectual Property Right (IPR) rules so, prior to using its content please contact the consortium coordinator for approval.

In case you believe that this document harms in any way IPR held by you as a person or as a representative of an entity, please do notify us immediately.

The authors of this document have taken any available measure in order for its content to be accurate, consistent and lawful. However, neither the project consortium as a whole nor the individual partners that implicitly or explicitly participated in the creation and publication of this document hold any sort of responsibility that might occur as a result of using its content.

This publication has been produced with the assistance of the European Union. The content of this publication is the sole responsibility of BIGPROD consortium and can in no way be taken to reflect the views of the European Union.

The European Union is established in accordance with the Treaty on European Union (Maastricht). There are currently 27 Member States of the Union. It is based on the European Communities and the member states cooperation in the fields of Common Foreign and Security Policy and Justice and Home Affairs. The five main institutions of the European Union are the European Parliament, the Council of Ministers, the European Commission, the Court of Justice and the Court of Auditors (http://europa.eu.int/)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870822.

Introduction

As in many areas, big data and data analytics has become a near necessity. With the increasing volume of data being created each day, we can uncover a wealth of data on all types of human activity. In this, innovation is no different. While we have seen a significant upsurge on the use of big data and data analytics in many areas, such as research and industry (Zhou et al., 2014), public policy-making has not taken full use of the potential of policy analytics.

For adoption to happen advocates are dependent on the user seeing the usefulness of the technology and on them understanding that there is relative ease of use. (Venkatesh & Davis, 2000). We also know that there is or has been scepticism towards the use of policy analytics in public policymaking (Guenduez et al., 2020). This suggests that practitioners have not seen the utility of data analytics, or that there is a lack of capability among practitioners, or that the combination of both is present. Some of this lack of adoption can be the result of a mismatch with practice and expectations – this is to say that we do not see the perceived usefulness and needed capabilities the same way. Durrant et al. (2018) say that our aspirational motivation to use policy analytics is not reflected by its everyday utility.

We know that big data can provide tools for better decision making and particularly in ensuring that decisionmakers grow better informed. Research has suggested that we have been prone to select the easy-to-handle aspects of big data and that we should still focus more on in-depth empirical research (Chatfield et al., 2015), on building technical capacity (Höchtl et al., 2016; Wang et al., 2015), and on showing actual implementations of big data in public policy (Guenduez et al., 2020; Vydra & Klievink, 2019). Future efforts to support decision-making should potentially address both the perceived usefulness of big data for policymakers and the ease of use of such technologies.

Want to know more about the large dataset. Watch the project webinar online

Webinar: Indicators on firm level innovation activities from web scraped data - BIGPROD Data & Information (dataverse.nl)

The BIGPROD project strived to address all of the three aspects mentioned above -- empirical research, technical capacity, and actual reference implementations -- in the context of innovation policy. During the project there has been an emergence of a significant body of research on websites providing an interesting source of data on companies' innovation activities, particularly with regard to downstream innovation activities (Gök et al., 2015). The results delivered additional value to patenting and publication based analysis. Similar findings have been presented by for example Kinne & Axenbeck (2020), Arora et al. (2020) and Li et al. (2018). The BIGPROD project focused on empirical research in the problem framing of the productivity paradox. During the 2.5 year project we implemented a data platform to source novel data for approximately 96 000 companies, created and engaged stakeholders with open access data, webinars and code examples. Following highlights the key outcomes from the project.

Evidence and Analysis

The BIGPROD project set out to create a large-scale web scraping based exercise to create novel firm-level data on innovation. The backdrop of the project are the challenges entailed in measuring productivity, as well as the challenges in instrumenting the often poorly measured background processes of innovation. Solving both challenges requires the use of novel data sources. However, we envisioned that other aspects of measuring innovation which could be uncovered during the project would be of equal value.

While we did not anticipate that these novel approaches would reduce the importance of existing statistical office data or survey data, the project is expected to offer a novel vantage point on innovation. This is a perspective which might be more flexible or broaden what can be measured. Now, as the project ends, it is important to reflect on the most policy-relevant findings from the project.

More on the challenges of Policy Analytics, read the project 's recent publication.

https://onlinelibrary.wiley.com/doi/pdfdirect/10.1002/poi3.258

1. Leverage the data platform. The project created a data platform which enables the extraction of novel micro-level data at scale. This is an important outcome as we have demonstrated the possibility to create a data platform that can create meaningful data and pre-process it automatically to policy-relevant variables. For policymakers this suggests that-web mined data is becoming a routine, operational tool rather than something researchers create and use on a case-by-case basis.

2. Expect data loss. The project envisioned, based on an analysis during the project planning phase, roughly 200 000 companies being web scraped. This analysis was based on calculating how many companies would be available for our sample from the Orbis database. In our efforts to web scrape the websites of our sample, data was only retrieved for roughly 96 000 and from those good financial data from Orbis was only available for roughly 47 000.

   The first loss of data was the result of missing websites or the retrieval of otherwise erroneous data from the websites. The second loss of data was mainly due to heterogenous availability of financial data from different countries. For policymakers our results suggest that while web-scraped data has the potential to create significant volumes of new and interesting policy-relevant data, gaps still remain. This in our case can lead to important aspects of innovation remaining invisible. Transparently reporting any invisible subpopulations is essential to ensure validity of the results.

3. Embrace the novel vantage points to innovation. The project runs several different pilots from the data. Most interestingly the analysis focused on three cases: 1) the structure of economic activity, 2) the measurement of digitalization and 3) instrumenting collaboration in economic activity. For all the cases, the web-scraping based analysis was shown to offer a novel view to innovation that expanded the vantage point of existing measures. For example on collaboration, BIGPROD data was able to identify roughly two times more collaboration than existing bibliometric measures reported. These findings challenge our existing indicators. Nonetheless we should avoid a one-to-one comparison with existing measures. In the collaboration example, we were able to identify that the web-scraped data reports on a much broader scope of activity than for example co-patenting. Thus, a one-

to-one comparison between patenting and web indicators quickly leads to comparing apples to oranges.

4. Leverage open access material. With over ten stakeholder events throughout the two-year project, several data and code releases and webinar recordings, the project has made an effort to up-skill stakeholders on the use of big data in policymaking. However we see that policy analytics of this type is still in nascent form and therefore more support is needed to stakeholders to better understand the potential, the limitations and the capabilities needed to utilize the data and methods used in the BIGPROD project. For this purpose, the BIGPROD project leaves a significant record of open access material for interested innovation policy practitioners.

BIGPROD data and descriptive analysis was published open access.

https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3938767

Policy Implications and Recommendations

The BIGPROD project's key policy implication is that meaningful innovation policy indicators can be created from unstructured web sources. This requires that there is a robust data platform and data analytics process in place and that policymakers are transparently informed on what is and is not in the data and subsequent analysis. We further found that you should expect data loss and that the process is sensitive to the creation of invisible subpopulations, which further call for a rigorous and transparent analysis process.

We also found that validating the data is challenging to a significant extent. Our results suggest that the measures offered by the web scraped data report on a broader behavior than we can measure with existing indicators. For example, our analysis of the website data identified seven different categories of information to be available on the websites. The breadth of these extends outside of innovation outcomes to e.g. processes and investor relations.

Despite all its limitations, the potential of web-scraped data for innovation policy analysis is clear. This was also indicated across multiple of our stakeholder events. The existence of a robust data platform, releasing data publicly and creating transparent indicators, offers an important vehicle to develop longitudinal analyses with the web- scraped data. This would allow analysts to promptly see important changes in innovative behavior. The possibility to rapidly reveal such changes may not exist in other types of data.

How does the BIGPROD data platform work? Find the platform description from the project website

https://www.bigprod.eu/wp-content/reporting/D8_Interim-Platform-Operation-Report-v3-22-12-2020.pdf

For this to happen, more work needs to be done to validate the indicators created thus far, and to also incorporate these indicators to the policy process. This requires more hands-on work with policy practitioners to ensure that the resultant developed indicators are tried and tested so as to be useful in the policy process.

BIGPROD delivered a state-of-the-art platform for creating novel firm-level indicators for innovative activity. To leverage the work and potential of this new type of indicators we make several recommendations:

Interested on examples or pilots run during the project. Read the write-up and listen to the webinar recording.

Write-up: BIGPROD: Write-up of three pilot cases - BIGPROD Data & Information (dataverse.nl)

Recording: 2nd Webinar: Indicators on firm level innovation activities from web scraped data - BIGPROD Data & Information (dataverse.nl)

1. Enable the creation of longitudinal data. What has become clear during the project is that gathering longitudinal data is essential. This will allow us to use the indicators more on their own merit, tracking the relative change in the data. In addition, this will inform us on the robustness and overall volatility of the data. This recommendation is to have additional research conducted extending the BIGPROD project.

2. Engage stakeholders. To operationalize the indicators as an important policy tool for measuring innovation, more stakeholder involvement is needed. A potential avenue towards this is to use the data platform created in the project to create an innovation indicators dashboard allowing for stakeholders to see and interact with different web scraped data-based indicators. This would allow for practitioners to become aware of the data and indicators, which would ultimately increase adoption. This recommendation would require funding to facilitate interaction between stakeholders and researchers.

The above-mentioned policy recommendations address the main knowledge gaps remaining after the project. These are to better understand the value of the data through longitudinal data creation and having more stakeholder interaction to ensure that the measures created can have an impact to the policy cycle.

References

Arora, S.K. et al. (2020) “Measuring dynamic capabilities in new ventures: exploring strategic change in US green goods manufacturing using website data,” Journal of Technology Transfer, 45(5), pp. 1451–1480. doi:10.1007/S10961-019-09751-Y.

Chatfield, A., Reddick, C., & Al-Zubaidi, W. (2015). Capability challenges in transforming government through openand big data: Tales of two cities. In Proceedings of Thirty Sixth International Conference on Information Systems. Fort Worth, TX, USA

Durrant, H., Barnett, J., & Rempel, E. S. (2018). Realising the benefits of integrated data for local policymaking:Rhetoric versus reality. Politics and Governance, 6(4), 18–28.

Guenduez, A. A., Mettler, T., & Schedler, K. (2020). Technological frames in public administration: What do public managers think of big data? Government Information Quarterly, 37(1), 101406.

Gök, A., Waterworth, A. and Shapira, P. (2015) “Use of web mining in studying innovation,” Scientometrics, 102(1), pp. 653–671. doi:10.1007/S11192-014-1434-0/TABLES/5.

Höchtl, J., Parycek, P., & Schöllhammer, R. (2016). Big data in the policy cycle: Policy decision making in the digital era. Journal of Organizational Computing and Electronic Commerce, 26(1–2), 147–169.

Kinne, J. and Axenbeck, J. (2020) “Web mining for innovation ecosystem mapping: a framework and a large-scale pilot study,” Scientometrics, 125(3), pp. 2011–2041. doi:10.1007/S11192-020-03726-9.

Li, Y., Arora, S., Youtie, J., & Shapira, P. (2018). Using web mining to explore Triple Helix influences on growth in small and mid-size firms. Technovation, 76, 3-14.

Venkatesh, V., & Davis, F. D. (2000). A theoretical extension of the technology acceptance model: Four longitudinal field studies. Management Science, 46(2), 186–204.

Vydra, S., & Klievink, B. (2019). Techno-optimism and policy-pessimism in the public sector big data debate.Government Information Quarterly, 36(4), 101383.

Zhou, Z.-H., Chawla, N. V., Jin, Y., & Williams, G. J. (2014). Big data opportunities and challenges: Discussions from data analytics perspectives [discussion forum]. IEEE Computational Intelligence Magazine, 9(4), 62–74.

For more information, please contact

Dr. Arho Suominen (Consortium leader)
Tel. +358 50 5050 354
arho.suominen@vtt.fi

About BIGPROD

BIFPROD is a research project focusing on Big Data based analysis of productivity using webscraped data. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870822.

The project partners in the project are Quantitative Science and Technology Studies team, Foresight-driven Business Strategies, 1) VTT Technical Research Centre of Finland, Competence Center Innovation and Knowledge Economy (Coordinator), 2) Fraunhofer ISI, Economics of Knowledge and Innovation team, 3) UNU-MERIT, Maastricht University, 4) Public Policy and Management Institute, 5) Economics of Technology and Innovations, Faculty of Technology, Policy and Management, 6) Delft University of Technology, Economics of Technology and Innovations, 7) Faculty of Technology, Policy and Management, Delft University of Technology

www.bigprod.eu

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870822