technopolis |group|

oxford internet institute university of oxford

CENTRE FOR EUROPEAN POLICY STUDIES

Data for Policy: A study of big data and other innovative data-driven approaches for evidence-informed policymaking

Final study report

Summary

May 2016

Introduction

- The European Commission has invited the Technopolis Group, the Oxford Internet Institute (OII) and the Centre for European Policy Studies (CEPS) to conduct an international study on innovative data-driven approaches to inform policymaking. In short: data for policy.

- **The main objective of the study was to explore the opportunities that innovative data-driven approaches offer for evidence-informed policy making, including the relevant data sources and technologies.**

- The focus of the study was on (big) data for policy opportunities for national and international policy makers, including the European Commission itself.

- While exploring big data, linked data and innovative uses of small data across the policy cycle, the study touched on open data, citizen science, participative policymaking and advances in system dynamics.

- The study was action-driven and aimed at the development of an Agenda for action for practitioners and other stakeholders (policymakers, public agencies, NGOs, companies that provide tools, collect data, etc.). To this end, it engaged with interested parties and contributed to creating or linking relevant communities in the field. For instance, the draft report about the State-of-the-Art was published on the study website www.data4policy.eu allowing stakeholders to provide comments and suggest relevant initiatives. Moreover, the draft report about the State-of-the-Art was presented and discussed at the Big Data for Policy conference that was held 15-17 June 2015 in Cambridge. The study workshop took place 22 September 2016 in Brussels. The workshop was designed as an interactive event that attracted over 90 practitioners and other stakeholders.

- This final study report summarises the activities and deliverables of the Data for Policy study. The final deliverables are available at the study website. This concerns:

    - State-of-the-Art report (Appendix A to the Final Study Report)

    - Workshop report (Appendix B)

    - Ten use cases of innovative data-driven approaches for policymaking at EU level (Appendix C)

    - Bee health: demonstrator of (big) data innovations at EU level (online demonstrator of data linking and data visualization for policymaking)

State-of-the-Art report

- Based on a literature review, interviews with thought leaders and an inventory of 58 innovative data for policy initiatives in the EU, selected non-EU countries and international organisations, the State-of-the-Art report concluded:

    - The use of big data and other innovative data-driven approaches for policymaking is being researched and piloted in a **broad range of policy areas** and in the context of systemic challenges such as the interaction between mobility, health and environment, or the various factors that influence economic development of regions.

    - So far, the emphasis is on **pilots/experiment**, with very few implementations and scaling up of innovative approaches. There are exceptions in a small number of countries (including initiatives by statistical offices), at the EC's Joint Research Centers and at NGOs.

    - The emphasis is on using innovative data-driven approaches for **agenda setting** and **problem analysis** (e.g. measuring global

priorities via Twitter and tracking traffic via sensors and mobile phone data), the use of open data for **transparency, accountability and enhancing participation** (with initiatives by policy makers and NGOs) and using administrative data and statistical data for **implementation of policy** and **monitoring** the output policy. There are few initiatives that focus on policy evaluation and impact assessment.

- The majority of 58 initiatives studied concerns **linking of established datasets** (such as national statistics and open, administrative data) rather than using new data sources such as social media data, sensor data and mobile phone data.

- The main types of **data analytics** deployed are descriptive statistics and trend analysis, with some experiments using (advanced) sentiment mining, profiling, predictive analytics and other recent tools.

- **Privacy concerns** depend, to a large extent, on the types of data sources that are used and linked. For instance, concerns are more substantial for linking different administrative datasets and using new data sources such as mobile phone data (e.g. Call Detail Records), compared to few concerns about using sensor data.

- A number of topics emerged, as the field of data-driven approaches for policymaking is developing rapidly:

  - Concerns about the availability of relevant **skills** in public sector organisations, e.g. skills related to data collection, data analytics and interpretation of (visual) data.

  - The risk that data-driven approaches can reduce **transparency** of the policy process when data collection and data analytics (e.g. algorithms, machine learning) are not fully understood and explained by policymakers and other stakeholders.

  - The balance between collecting the most **relevant data** (given a policy issue, relevant factors and indicators) and using data that is readily available.

  - Stakeholders influence the selection of data sources and tools. Big data doesn't reduce the incentives for stakeholders to present **policy-based evidence**.

  - The continuous or even increasing importance of **international collaboration** on data standardisation/harmonisation related to skills, company data, air quality, etc.

  - Opportunities to combine **policy experiments** (e.g. in specific countries or for specific target groups) and data-driven approaches for impact assessment.

  - The need for (better) strategies to ensure that policymakers are informed about the tools that are being developed and piloted successfully in research projects. This should stimulate knowledge exchange and **implementation**.

Workshop report

- The study workshop took place 22 September 2015 in Brussels. Participants discussed the **expectations** of using big data and other data-driven approaches in policymaking. New data sources, data linking and new analytical and visualisation tools are expected to address the shortcomings of current approaches for answering existing questions. New approaches are expected to also provide new perspectives and opportunities, e.g. real-time policy monitoring, moving from descriptive to explanatory models and studying digital phenomena such as online political engagement of citizens.

- However, the use of new data sources (social media, web crawling, sensor data, etc.) for policymaking is still in an **exploratory phase**. This applies to the data sources as such, with concerns on data quality and representatives, and the tools for analyzing and visualizing data. Another challenge lies in the skills that are needed to analyze data, interpret data and to translate the results (and the uncertainties) into policy recommendations or policy changes.

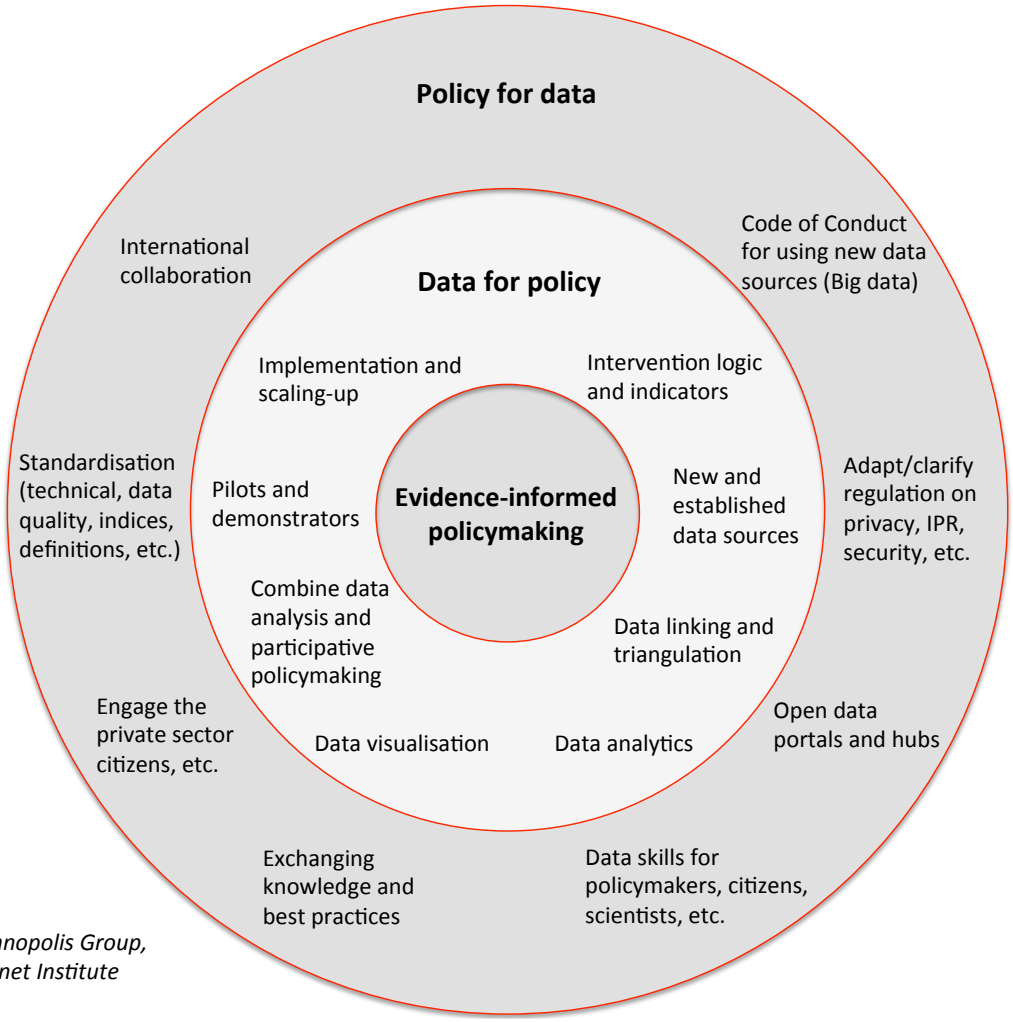
- A number of **best practices** are emerging. Of particular interest are the costs and benefits of using new data sources, compared to using well-established approaches such as large-scale surveys. This information can be used to further increase awareness among policymakers.

- Workshop participants stressed the need for **caution and transparency**. For example, the concept of triangulation is highly relevant when exploring the possibilities and limitations of using new data sources. This refers to the combination and confrontation of several data sources and of different analytical tools (econometrics, machine learning, qualitative approaches, etc.). Privacy is another theme in which caution and transparency are needed.

- The community of data scientists/practitioners should take the initiative to develop a **Manifesto**, **Code of Conduct** on using new data sources (especially big data) for policymaking. Among the framework conditions to be addressed by a Manifesto or Code of Conduct are: standardization of concepts; data quality; privacy; IPR; skills; and access to (privately owned) data for public interest.

- Figure 1 on the next page depicts the main topics that emerged from the workshop. A distinction is made between **data for policy**, which refers to the increased possibilities to use data for evidence-informed policymaking (new data sources, new analytical tools, etc.) and **policy for data**, which refers to the relevant framework conditions for making further progress in using data for policy.

Figure 1: Policy for data and data for policy



Source: Technopolis Group, Oxford Internet Institute and CEPS

Ten use cases of innovative data-driven approaches for policymaking at EU level

- The main purpose of the use cases was to inspire the European Commission and other organisations at EU or international level, when considering or implementing innovative data-driven approaches for policymaking. The use cases were developed in close collaboration with officials from several European Commission Directorates-General (DGs).
- The 10 use cases cover the following policy areas/themes:
    - **Using learning analytics systems for better educational policies.** This use case focuses on the opportunity that micro-data on learning processes (e.g. within universities) and the use of learning analytics provide for the design of educational strategies by policymakers at national and European levels.
    - **Nowcasting for economic policy and beyond**. Nowcasting is a forecasting methodology that is becoming increasingly popular in economics. The use case considers the potential use of nowcasting in the context of economic policy setting and sets the potential value of an extended use of nowcasting against different contexts.
    - **Ocean governance**. A use case that illustrates the opportunity that exploitation of vessel positioning data, linked to other data sets, represents for an improved conservation of marine resources and global ocean governance.
    - **New data technologies for trade policy.** An exploration of the possibilities for using new innovative data sources for the development and monitoring of trade policy. Data on the firms' global supply chain management, Global Trade Item numbers and sensor/GPS tracking at the level of firms would allow for an analysis of the impact of new trade agreements at a micro level.
    - **Big data for consumer policy**. A use case exploring the possibilities and potential benefit of online retail monitoring for the implementation and enforcement of consumer policy, specifically web-scraping for consumer price indexing and sales conditions monitoring.
    - **Data linking for bee health**. A use case looking into the opportunity to develop shared data and visualization tools related to the health of bees for improved policymaking in a variety of areas, including agricultural policy and science policy.
    - **Big health data in dementia**. This use case considers the opportunity to leverage big data for improved policymaking and science with relation both to dementia detection and treatment and to dementia care.
    - **Big data for crisis anticipation and crisis management**. Crisis data can help policymakers to save lives, re-build infrastructure and improve the environment. This use case describes the potential use of data, the collection processes and analytical tools for data related to natural crises, confrontation crises, crises of malevolence, and technological crises.
    - **Citizen science: big data for environmental policies**. This use case describes the potential value and processes for the collection, analysis and use of crowdsourced citizen science data for the development, monitoring and enforcement of environmental policies.
    - **Text and opinion mining for policymaking**. This use case covers two methods that can assist policymakers throughout all stages of the policy cycle. It explains the sources for these data and how the outputs can be used to gain understanding of stakeholders' and citizen's opinions on policies and strategies.
- The most common (intended) use of the innovative data systems is the collection of **contextual information** for the purpose of informing policymaking at the **very**

beginning of the policy cycle (foresight, problem analysis) or for **monitoring** purposes, at times with the intention of enforcing legislation.

- For each use case we indicated the **'readiness level'**, i.e. the extent to which the idea is still in its embryonic phase or has already reached the level of a working prototype, close to being used in the policymaking process. The majority of use cases scored low to medium readiness/maturity (Table 1). In some cases the **data sources are not (yet) available at the scale needed** (administrative and transaction data for learning analytics and sensor-based data for bee health) or are commercial data that are **proprietary** (data for trade policy). In other cases, technical problems or the current lack of maturity and exploratory nature of the analytical methods hinder the deployment of the use case. An example of the former is the interoperability of the data systems on dementia; examples of the latter are the predictive analytics for trade policy and the advanced semantics for webscraping for consumer policy. More in general, **descriptive analysis** is expected to remain the basic yet crucial analytical approach in most cases. Increasingly, explanatory modeling and predictive analytics will be added.

Table 1 Use cases: readiness, main data sources and data analytical methods

Nr	Use case	Readiness (scale 1-5)	Main data sources	Data analytical methods
1	Data mining of learning analytics systems for better educational policies	2	Administrative/student data, transaction data	Data mining
2	Nowcasting for economic policy and beyond	2	Sensor data, social media data	Statistical modelling
3	Ocean governance	4	Satellite data, administrative data, mobile (location) data	Behavioural classification models
4	New data technologies for trade policy	2	Transaction data, administrative data, sensor data	Trade models, predictive analytics
5	Big data for consumer policy	3	Website data, social media data	Webscraping algorithms and advanced semantics
6	Data linking for bee health	2	Sensor-based data, crowd sourced data, satellite data, transaction data	Integrated data platform for monitoring, trend analysis and modelling
7	Big health data in dementia	2	Administrative/medical data, sensor data and mobile (location) data (e.g. personal physical activity data)	Integrated data platform for monitoring and trends analysis
8	Big data for crisis anticipation and crisis management	4	Satellite data, geospatial data, sensor-based data, mobile data (location and Call Detail Records), social media, crowd sourcing	Monitoring, early warning, modelling
9	Citizen science: big data for environmental policies	3	Sensor-based data, crowd sourcing, satellite data	Environmental modelling
10	Text and opinion mining for policymaking	3	Social media data, text/documents	Webscraping algorithms and advanced semantics

- Five topics emerged in nearly all use cases:
    - **The challenge to actually use data analysis in the policy process**, i.e. establish a solid and timely link between on the one hand data collection and data analytics and the other hand the policy process (e.g. politics and stakeholder engagement). To establish this link requires collaboration, or a mutual understanding, between data scientists and policy makers. For instance, data should be visualised in ways that are relevant for policy makers, while policy makers should increase their data literacy.
    - **Privacy concerns**, e.g. when aggregating and using data about students or patients for policymaking, or when using social media data to anticipate crisis.
    - **Skills**, e.g. experts in specific policy areas and experts in data analytics that are investing in understanding each other, and that collaborate when developing analytical (causal) models about the impact of policy interventions.

- **Public-private collaboration**, e.g. convincing companies in financial, retail, health and international trade, to share data that allow government agencies to monitor economic trends, explore which consumption patterns and health issues might indicate dementia, and analyse how global value chains evolve.
- **Financial resources**, e.g. the resources that are required to develop or scale up big data applications for policymaking, e.g. installing sensors, engaging citizens in crowd sourcing approaches and developing algorithms for web scraping of e-commerce websites (to replace mystery shopping by automated approaches).

Bee health: online demonstrator

- One of the goals of this project was to demonstrate the potential for improved support for policy decisions at the regional, national and EU level by building a tool that shows how disparate data can be linked in new ways, relying on both established and new data sources. The *Bee Health Demonstrator (online workbench)* was designed to do just that: to combine real and simulated data from public and private sources in novel ways to demonstrate the potential such a system could have if developed into a production-scale tool.
- Bee health has gained significant attention from many interested parties because of bees' essential role in crop pollination combined with rapid colony declines worldwide. Over the past 10 to 15 years, beekeepers have been reporting unusual weakening of bee numbers and colony losses, particularly in Western and Southern European countries including France, Belgium, Switzerland, Germany, the UK, the Netherlands, Italy and Spain. A single cause for the decline has not been found. Instead, one of the hypotheses is that multiple factors (pollution, food shortage, pesticides, pathogens and parasites, beekeeping practices) affect bee health. To counter the bee population decline, the European Union actively intervenes in the area with regulations, directives, support measures and research support.
- The policy implications of bee health are enormous. From the economic point of view, both agricultural employment and food security more broadly are intricately tied to the ability of bees to pollenate crops, and the inability of bees to pollenate crops can result in price increases for the consumer in additional to job losses all across the food supply network. From an environmental point of view, bee health is not only a bellwether for changes in the environment which may also effect other organisms, but bee behaviour is highly predictable and can be used to measure, for instance, changes in seasonality over time, flora diversity in rural and urban settings, migration of invasive species, and other topics.
- To be able to monitor policy effects, or at least be as close to the monitoring data as possible, a visual tool that combines data intuitively for exploration and investigation would be a valuable contribution.
- The *Bee Health Demonstrator* is intended to illustrate how data drawn from many different sources, and at different levels of granularity, can be integrated to afford users such as beekeepers, scientists and policy-makers an integrated view of the bee health domain and the various stressors that may affect bee health.
- This data exploration can be carried out at several levels:
    - Hive-level (microscale): the real-time behaviour of single hives and apiaries, including information on hive internal environments.
    - Neighbourhood-level (mesoscale): information on local environmental conditions, including local weather, nearby crops and their flowering patterns, local use of pesticides and local pollution.
    - National and EU-level (macroscale): information about crop-use and other biodiversity indicators, nation-wide usage of pesticides, incidence of disease, etc. and their effects at national level, as well as information about regulatory events and their impacts.

- Please note that this is intended to be a demonstrator, not a production platform. It makes no pretence at comprehensiveness nor scientific robustness. However, it serves as an illustrative example for the construction of an "industrial strength" workbench for this and similar purposes.
- This concept should be applicable not only to the case of bees but to many other policy areas that could benefit from real-time, continuous sensor monitoring combined with a comprehensive set of environmental and policy data that can be linked at the local, regional, national and EU level. Showing such combined data in a unified visualisation and analysis interface allows explorations and inferences that may not be apparent by looking at individual data sets.

Figure 2: Sample Screens from the hive (micro), area (meso) and national (macro) levels of the *Bee Health Demonstrator*