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

Report about the State-of-the-Art

17 July 2015


technopolis
group
oxford
internet institute
university of oxford
CE
PS
CENTRE FOR
EUROPEAN
POLICY
STUDIES

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

Report about the State-of-the-Art

technopolis |group|

Oxford Internet Institute

Centre for European Policy Studies

Martijn Poel[a], Ralph Schroeder[b], Jérôme Treperman[a], Mor Rubinstein[b], Eric Meyer[b], Bea Mahieu[a], Chiel Scholten[a] and Marina Svetachova[a]

(a: Technopolis Group, b: Oxford Internet Institute)

The authors would like to thank the 100+ experts and practitioners that have provided input on relevant trends, research projects and operational initiatives via interviews, email, Twitter, LinkedIn, the European Commission's internal Data for Policy Network and the Policy-Making in the Big Data Era conference, 15-17 June 2015 in Cambridge. We would like to thank Colin Blackman (CEPS), Prabhat Agarwal and Diana Vlad-Calcic (European Commission, DG CONNECT) for all their valuable suggestions.

17 July 2015

Table of Contents

Table of Figures

Summary

This study, undertaken by the Technopolis Group, the Oxford Internet Institute and the Centre for European Policy Studies for the European Commission, has investigated **data-driven approaches for evidence-informed policymaking**. The aim was to identify existing practices and emerging opportunities for evidence-based policymaking. The study examined a purposive sample of relevant initiatives and stakeholders and has undertaken interviews with key stakeholders.

Two main types of data: The discussion of data-driven approaches to support policymaking should distinguish between two main types of data. The first is the use of **public datasets** (administrative (open) data and statistics about populations, economic indicators, education, etc.) that typically contain descriptive statistics, which are now used on a larger scale, used more intensively, and linked. Open data is widely promoted in the public sector and among NGOs, and there are now analyses of the quality of freely available open government datasets that finds that heterogeneity is still an issue even if quality is often good. Half the datasets among the initiatives examined were open or semi-open, with some kind of registration needed. The rest of the datasets are restricted for private or non-commercial use, e.g. use by public authorities in a specific policy area rather than all public authorities, stakeholders or the general public.

The second main type of data is from **social media, sensors and mobile phones** that are typically new data sources for policymaking, analysed with novel methods such as sentiment analysis, location mapping or advanced social network analysis. The importance of these new types of data is increasing. Data linking is becoming the norm, e.g. linking new data sources with established data sources such as official statistics, surveys and commercial/business databases.

Privately held data is, currently, of less relevance. Private actors (market research, data brokers, data analytics, etc.) benefit from public datasets, while still investing in private datasets and bespoke data collection. As such, a nuanced picture emerges with respect to 'data as the new oil.' Only some data are scarce. A more accurate metaphor would be to say that data offers new fields for harvesting, but some will bear better fruit than others, especially since when for so many datasets, the main task is cleaning and making data available. The picture with respect to openness of data is nuanced too.

Users and uses in the policy cycle: National governments and national and international agencies are becoming common users of innovative data-driven approaches. Based on our inventory of 58 initiatives, the most common uses include utilising data 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.

A number of pilots, included in the inventory of 58 initiatives, and a quick scan of relevant European research projects illustrate the opportunities for full-scale implementation of data-driven approaches across all stages of the policy cycle, including policy design, evaluation and impact assessment (ex ante and ex post). The same picture emerges from research programmes managed by the UN (Global Pulse) and World Bank. Another source of inspiration for national and international policy makers is the progress being made at regional and local policy level (cf. smart cities).

In terms of policy areas, a diverse picture emerges. No policy area dominates although the inventory includes several initiatives in transport and mobility, environment, information society, budget, financial and economic affairs. 11 out of 58 initiatives concern the cross-cutting policy theme of transparency and accountability (e.g. data about procurement, policy procedures and funding of political parties).

Data analytics and visualisation: There is a contrast between analytic techniques performed on conventional PCs with widely available (often open source) tools and those demanding high-performance computing tools and centralized databases. The most novel analytics tools are in the area of social media analysis, network analysis and visualisation tools for mapping mobile phone locations. More important than the tools are ways in which they can be combined and refined and applied, and more important than new visualisations are the skills needed to use them and not be misled by them. The opportunities for policymakers are substantial, as a range of tools has become available. Still, many initiatives focus on descriptive statistics. Among the initiatives examined, there is little use of advanced analytics or visualization techniques, which is mostly confined to academic research, public-private partnerships and NGO initiatives.

Privacy and inclusion: The concerns over privacy from large-scale data analysis have been much discussed, but this does not concern all datasets. To be sure, new mechanisms for more transparency in data collection are needed, as are new mechanisms for obtaining consent. The European Union's new proposed legislation will be a much-watched model as to whether data protection that is adequate to the novel big data environment can be provided. The new data protection proposals are aimed preventing harms and at the same time promote the Digital Single Market, and the main impact in relation to data for policy making can be anticipated to be reassuring the public that data are being handled appropriately. In the initiatives examined for this project, privacy and data protection played a role in only a small number of initiatives.

In the meantime, inclusion concerns arise. Certain parts of the population with no digital traces or uncaptured environments (e.g. less sensors) might be left out of data collection procedures. If so, their data and interests are not taken into account when using data across the policy cycle. The initiatives examined for this project appeared to provide little attention to inclusion concerns and mitigation strategies.

Known unknowns: There are advanced uses of data-driven approaches in the areas of crime, fraud and terrorism (although there are concerns that hierarchical, centralised organisations may not explore decentralised and innovative data-driven approaches). However, little is knowable about these largely secret initiatives, at least by means of a study of this type. Another 'known unknown' is whether the data collection strategies and data analytics that are developed in research projects can be implemented successfully, taking into account costs, benefits, privacy, etc. Many tools are at the stage of technical tests and operational pilots. For example, there are a number of initiatives that gauge prices, monitor public opinion, and try to detect emerging crises.

Outlook: While consulting experts and discussing the draft report, 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 data that is most crucial, given a policy issue, relevant factors and indicators, and using data that is readily available.
- Acknowledge that 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.

1. Introduction: about the study and this report

1.1 Objectives of the study

Big data and other innovative data-driven approaches for policymaking create opportunities for evidence-informed policy and modernization of public sectors. The nature and magnitude of these opportunities are difficult to assess for a number of reasons. Opportunities may differ between policy areas and between stages of the policy cycle, depending on the relevance of data sources, data technologies (such as data analytics and visualisation), privacy issues, skills, enabling ICT infrastructures, etc. Gradually, information is becoming available about data-driven approaches for policymaking at local and regional level and, to some extent, at national and international level. Still, more has been written about data-driven approaches in private sectors such as retail, media and manufacturing firms.

The European Commission has commissioned 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.

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

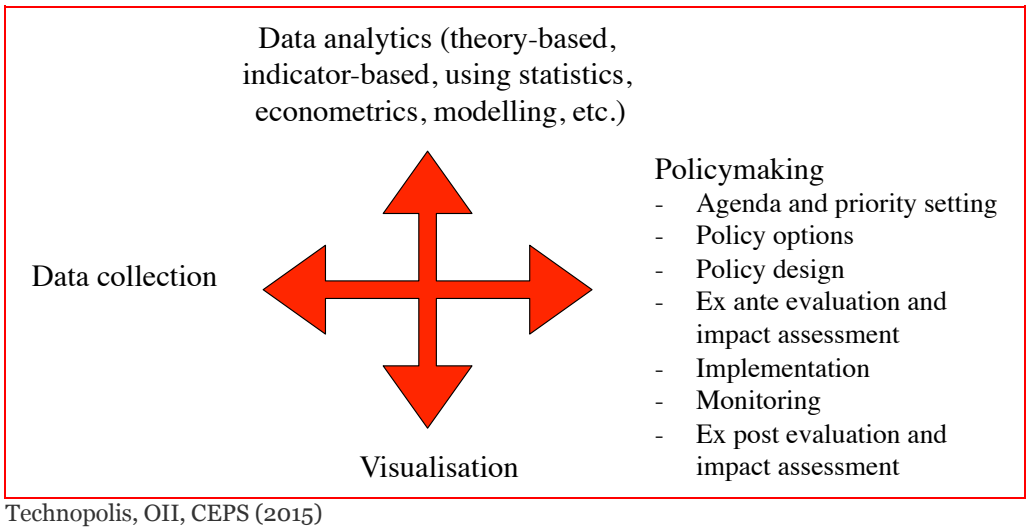
The study is action-driven and aims at the development of an Agenda for action for practitioners and other stakeholders (policymakers, public agencies, NGOs, companies that provide tools, collect data, etc.). To this end, it engages with interested parties and contributes to creating or linking relevant communities in the field.

1.2 Scope of the study

This study by definition explores a dynamic phenomenon. As a result, decisions were needed on the precise scope of the study in order to ensure its relevance and feasibility, as well as to build on its complementarity to existing studies.

The study covers four dimensions that constitute the focal areas for the analysis, i.e. data collection and sources, data analytics, the use of data in subsequent stages in the policymaking cycle and, to a lesser extent, visualisation tools (Figure 1).

Figure 1 Focal points of the study

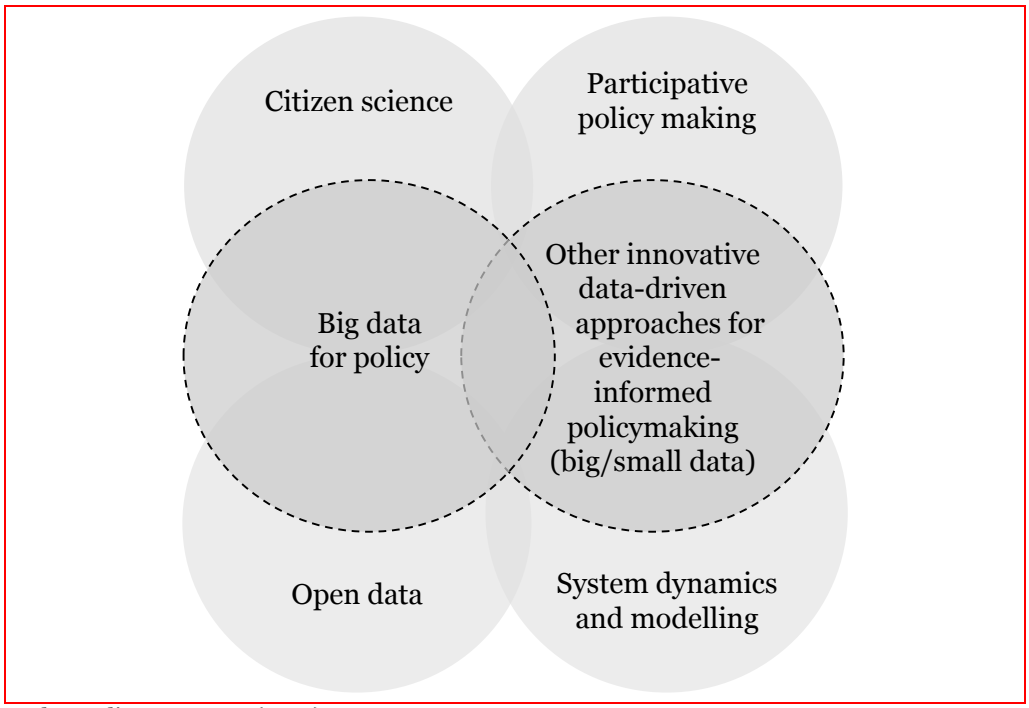


Technopolis, OII, CEPS (2015)

The study covers a broad range of policy areas because big data and other innovative data-driven approaches are relevant for all policy areas (transport, energy, education, security, health, research, taxation, etc.). Data-driven approaches also frequently cut across policy areas. For example, policymakers may combine datasets to explore the interactions between education, research, innovation, economic growth, employment and health.

Figure 2 illustrates the scope of the study. The study will focus on big data for policy but it also addresses innovative data-driven approaches that benefit from linking or scaling-up small data or that combine big data and small data (the definition of big data is discussed in Chapter 2. To some extent, the study addresses four developments or building blocks that are related to the big data phenomenon: open data, system dynamics and policy modelling (e.g. complexity science, complex adaptive systems and agent-based modelling), participative policymaking and citizen science. The study only addressed these four developments when they are an important part of data-driven approaches for policymaking. For example, open data initiatives that start adding data analytics and where the results are used in policymaking can be relevant for the study. Open data initiatives that remain a relatively specialized element of 'data ecosystems', instead, are less relevant for this study (but of course may be highly relevant in other contexts). Along the same lines, some citizen science and participative policymaking is not data-driven, while system dynamics and modelling need not always be linked to policy-making.

Figure 2 Scope of the study



Technopolis, OII, CEPS (2015)

The study will address policymaking at national and international level, with particular attention for policymaking by the European Commission. As such, the study is complementary to studies on 'big' data used for or by smart cities, local police, fire departments, crowd control, etc.

Finally, the study covers operational pilots, demonstrators and full implementations of data-driven approaches for policymaking (58 examples are included in the inventory of relevant initiatives). However, it also looks at research projects in which new types of data are used and relevant tools are developed (the emphasis is on Europe's Seventh Framework Programme).

1.3 This report and the next steps

This report about the state-of-the-art in data-driven approaches for evidence-informed policy is the first result of the study. It is based on:

- Literature review.
- Stakeholder analysis.
- Interviews with ten key selected thought leaders.
- An inventory of innovative data for policy initiatives in European Union Member States; in Canada, India, Singapore, South Korea and the US; and by international organisations (EC, UN, etc.). This concerns operational pilots, demonstrators and full implementations that are supported - on a structural basis - by policymakers.

The study team contacted between 5 and 20 persons in each EU Member State, to identify relevant initiatives. In addition, experts and stakeholders were contacted via a kick-off meeting with 40 EC officials, via Twitter (linked to the project website), LinkedIn, webinars and at conferences. Nevertheless, given the explorative nature and the broad scope of the study, relevant initiatives may have been overlooked.

For this reason it was decided to open the draft report to the broader community by publishing it on the project website (www.data4policy.eu) and allowing for online and offline feedback. Moreover, the draft report was presented at the Policy-Making in the Big Data Era conference, 15-17 June 2015 in Cambridge. Suggestions were taken into account when preparing the final version of the report.

The next steps of the study are:

- Studies of six to ten case studies (use cases) where data-driven approaches can be used for evidence-informed policymaking for the European Commission.
- One of these use cases will be developed into a pilot online demonstrator.
- Workshop on 22 September 2015 in Brussels.
- Workshop report.
- Final project report.

1.4 Structure of the report

The report is structured as follows:

- In Chapter 2 we present the literature review.
- Chapters 3-6 report on the results of the stakeholder analysis, inventory of relevant initiatives and the interviews. Chapter 3 covers the progress and expectations in specific policy areas; Chapter 4 looks into data sources; Chapter 5 presents the findings on data analytics and data visualisation; Chapter 6 covers the issues of privacy and inclusion.
- Chapter 7 summarises the main points and provides an outlook.

There are six appendices:

- Appendix A: List of interviewees. Thank you!
- Appendix B: Summary of the stakeholder analysis.
- Appendix C: List and summaries of 58 relevant initiatives (the inventory).
- Appendix D: The template that was used for doing the inventory.
- Appendix E: List and summaries of relevant research projects in the European Commission's Seventh Framework Programme.
- Appendix F: References Chapter 2 (Literature review).

2. Literature review

2.1 Introduction

What follows is a concise, rather than exhaustive, literature review designed to present the main challenges and insights related to using data for policy. The literature review also allows us to introduce a number of key topics related to data-driven approaches for policy-making. The four main topics discussed are: 1) scientific issues related to data-driven approaches; 2) existing and emerging research approaches; 3) ethical and legal issues; and 4) the rhetoric and reality of evidence-informed policy processes. Where previous empirical work has been conducted, this Chapter contains concrete examples of these empirical studies.

2.2 Prediction, replication, and limits of big data approaches

The debate about the **validity** of big data approaches, and thus with how useful big data are, has been shaped decisively by the paper by **Lazer et al**. entitled 'The Parable of Google Flu Trends' (2014).[1] The paper discussed a particularly policy-relevant area that is also of widespread public interest: namely how big data can help with prediction, in particular in this case for predicting disease outbreak. Google Flu Trends was launched in 2008 as an effort by a small team at Google to use queries in the Google search engine to accurately predict the prevalence of flu, and to do so more quickly than official sources such as the CDC. The authors of the 2014 paper make a number of methodological criticisms of the Google Flu Trends project which, they argue, undermine its results. A key criticism is that the source of data is such that the research cannot be **replicated**, which is an essential requirement for scientific validity. This is because researchers are unable to look inside the 'black box' of how Google Trends works since the data are proprietary and held by Google. It is interesting however that this criticism does not apply to a subsequent study that was done about predicting disease outbreaks using Wikipedia.[2] This study was able to **predict** a number of disease outbreaks with greater accuracy than the Google study, and it can be built upon and replicated since the data are freely accessible. The problem which the Wikipedia study highlighted is that not all diseases are equally susceptible to this type of analysis. Ebola, for example, yields only misleading results since the disease is mainly searched for on Wikipedia in the rich North, not in the poor countries of West Africa where the disease is most prevalent but where there are few internet connections.

Bengtsson's (2011)[3] research on how disease spread after a Haitian earthquake using mobile geolocation data highlights the **validity issue** in a different way. In this case there are few internet connections and survey data are hard to come by. But there is widespread mobile phone use and these data from mobile phones in this case could be accessed (another set of examples like this one comes from the Orange challenge of population movements in Ivory Coast, again using mobile phone records; see Taylor and Schroeder 2015)[4]. In Bengtsson's case, he was able to validate the findings from mobile phone records by comparing them with findings of officials who were gathering more traditional statistics, a comparison which found mobile phone data to be more accurate. Again, however, access to data is key, as work with mobile phone records obviously

[1] Lazer, D., Kennedy, R., King, G., and Vespignani, V. (2014). The Parable of Google Flu: Traps in Big Data Analysis. Science. 343, no. 14 March: 1203-1205.

[2] Generous, N., Fairchild, G., Deshpande, A., Del Valle, S.Y., Priedhorsky, R. (2014). Global Disease Monitoring and Forecasting with Wikipedia. *PLoS Comput Biol,* 10(11): e1003892.

[3] Bengtsson, L., Lu, X., Thorson, A., Garfield, R. andVon Schreeb, J. (2011). Improved Response to Disasters and Outbreaks by Tracking Population Movements with Mobile Phone Network Data: A Post-Earthquake Geospatial Study in Haiti. *PLoS Med*, 8(8): e1001083.

[4] Taylor, L. & Schroeder, R. (2015). Is bigger better? The emergence of big data as a tool for international development policy. *GeoJournal.*

requires a relationship with the mobile operator. In some cases, as with the collaboration between Jana (a mobile phone company) and Twitter and the UN Global Pulse initiative (http://www.unglobalpulse.org/), phone data can be made available on the basis of serving the public good.

2.3 Big data research approaches

Research about **Wikipedia** has become an important field because it can be replicated and the data are freely available, and this research has a number of implications for policymaking, including for example the incentives for collaboration and the possibilities for predicting economic phenomena.[5]

Einav and Levin (2013)[6] have surveyed the various ways in which **economists** have recently begun using datasets that were previously unavailable. Economists have used big data in many ways, such as gathering information about prices by scraping them from websites. The advantages (cost, real-time data and the like) must be set against the disadvantages (such as potential inaccuracies due to data quality, bias against items that are less visible online[7]).

Research has shown how what is at top of the agenda in the media is different for traditional media (newspapers, television) as compared with **new social media** (blogs and Twitter, see **Neuman et al 2014**)[8]. The implication for policymaking here is that it is possible to measure peoples' concerns without the need to survey them, and this is possible in a more comprehensive and timely and less costly way than with traditional media. Further, while traditional media measurements arguably reflect the concerns of journalists and media institutions, what people select and share via new social media may provide a more accurate and direct way of obtaining information of their interests and views. Along the same lines, much has been written about the uses of social media in during the 'Arab Spring'.

In terms of **scale**, **Eagle and Greene** (2014)[9] demonstrate that big data does not just apply to large populations: individual lives and interactions with others and with the environment, for example, can be measured in real time and using several types of measurement. They show how data then scales up, to neighbourhoods (for crime and traffic), to cities, then to nation-state, and ultimately to the global population. While all these scales are possible for big data, it will be clear that some are easier to use for policymaking than others.

Borgman (2014)[10] has unearthed the quite different ways in which researchers make use of various kinds of **digital data**. One of the policy implications for research policy in particular is that there are a many of challenges to how data can made available via data infrastructures, for example in terms of providing incentives to sharing data, providing the conditions for re-use of data including adequate metadata, and giving credit for data sources in the form of citations and the like.

---

5 Taylor, L. & Schroeder, R. (2015). Is bigger better? The emergence of big data as a tool for international development policy. *GeoJournal*.

6 Einav, L. and Levin, J. D. (2013). The Data Revolution and Economic Analysis. NBER Working Paper Series No. 19035: http://www.nber.org/papers/w19035

7 Taylor L., Schroeder R. and Meyer E. T. (2014). Emerging practices and perspectives on big data analysis in economics: Bigger and better, or more of the same? *Big Data and Society*. July-December: 1-10.

8 Neuman, R.W., Guggenheim, L., Mo Jang, S. and Bae, S. Y. (2014). The Dynamics of Public Attention: Agenda-Setting Theory Meets Big Data. *Journal of Communication*. 64: 193–214.

9 Eagle, N. and Greene, K. (2014). *Reality Mining: Using Big Data to Engineer a Better World*. Cambridge MA: MIT Press.

10 Borgman, C. (2014). *Big Data, Little Data and Beyond*. Cambridge: MIT Press.

Savage and Burrows' influential article (2007)[11] argued that the social sciences would be outperformed in terms of being able to do research about phenomena of social science and wider interests by companies with more powerful **access to data**. This argument can be faulted for its zero-sum approach to knowledge production and overlooks that data is also widely available to social scientists and other academic researchers, even if it is true that companies often have data that social scientists would dearly like to get their hands on (it can pointed that in many cases, commercial data quickly loses its value and could therefore be made available to academic social scientists, producing a win-win for both apart from the effort that is involved in making the data available in a suitable and legally unproblematic form). This point, and related points, are also addressed in a good **review** of the 'opportunities and challenges' of digital data for social research.[12]

The use of **Google for analytics** purposes has been described by **Aiden and Michel** (2013) and the use for uncovering patterns that emerge only from large-scale data of this type or that reveal facets of human behaviour that cannot be obtained from surveys (for example, about sexual behaviour) using search engine behaviour has also been demonstrated by Seth **Stephens-Davidowitz.**[13]

2.4 Ethical and legal issues

Boyd and Crawford (2012)[14] were among the first to sketch the epistemological, ethical and social challenges of big data. These challenges are also addressed in **Mayer-Schoenberger and Cukier's** (2013)[15] influential book about big data. It raises the major ethical and legal issues of big data, and especially the ethical issue of constraining **freedom** if certain behaviours (like crime) can be predicted and the legal issue of how an appeal about services offered on the basis of big data analysis (like credit scores or health and car insurance) is possible given that ordinary people and sometimes even the analytics experts themselves cannot understand how the analysis works. The solution they propose, a kind of neutral ombudsman who must have access to the data and be able to make a judgement about how it works, is one that continues to be put forward.[16] Along similar lines, it has been suggested that there should be a scientific ombudsman when scientific issues are at stake in EU policymaking.[17] More in general, there is now a transitional period when the hard work of implementing big data approaches is replacing the early phase of hype and enthusiasm.

The issues around **privacy and consent**, as well as ensuring the accuracy of big data analyses, are only starting to be tackled by researchers.[18]

Greenleaf (2012)[19] provides an inventory of countries that have data **privacy laws**, though he points out that these laws are aspirational and compare poorly with the

---

[11] Savage, M. and Burrows, R. (2007). The Coming Crisis of Empirical Sociology. *Sociology*. 41(5): 885-899

[12] Golder, S. and Macy, M. (2014). Digital Footprints: Opportunities and Challenges for Online Social Research. *Annual Review of Sociology*. 40:6.1–6.24

[13] http://sethsd.com/

[14] Boyd, D. and Crawford, K. (2012). Critical Questions for big data: Provocations for a cultural, technological and scholarly phenomenon. *Information, Communication and Society*. 15(5), 662-79.

[15] Mayer-Schonberger, V. and Cukier, K. (2013). *Big Data: A Revolution That Will Transform How We Live, Work and Think*. London: John Murray.

[16] Pasquale, F. (2015). The Black Box Society: *The Secret Algorithms that Control Money and Information*. Cambridge MA: Harvard University Press.

[17] http://www.sciencebusiness.net/news/77010/How-to-fix-the-EU-science-advice-muddle.

[18] Lane, J., Stodden, V. (2014). *Privacy, Big Data, and the Public Good: Frameworks for Engagement*. Cambridge: Cambridge University Press.

[19] Greenleaf, G. (2012) Global data privacy laws: 89 countries, and accelerating. Queen Mary University of London School of Law. *Legal Studies Research Paper*. No. 98/2012.

implementation of these laws. A related matter is **(non-)transparent uses** of data. Big data increasingly determine many facets of our lives (such as insurance and loans, but also online prices and targeted advertisements). **Pasquale** (2015: 146)[16] argues that 'we should require data controllers to keep records of the original source of their data, noting how it was collected, purchased, or bartered'.

The threat of big data is often associated with the 'big brother' scenario of Orwell's novel 1984, and the **surveillance** that the state and companies can exercise over populations.[20] What is overlooked in this dominant discourse about big data is a different troubling aspect, which is that people may get used to the convenience and benefits that big data provides, and in this way **adapt** to how their behaviour may be shaped by this technology (as when, for example, certain consumption offerings are tailored to individuals and not so much chosen as conformed to). This is the 'Brave New World' scenario depicted in Huxley's novel.[21] The flipside is that for certain purposes – changing peoples' behaviour with regard to energy uses, for example – this shaping of behaviour will be necessary and useful ('nudging' is one example).

Another ethical issue, related to inclusion, is that the relations between the data sources used (search engines, Wikipedia, mobile phone records) and the phenomena under investigation skews the actionable knowledge towards **those about whom the data are captured**. This is not just a question of the representativeness of users, but also potentially opens up a divide between those with access to the technology as against those who do not. This issue may arise in highly developed economies (e.g. urban vs. rural, rich vs. poor, high vs. low use of internet and mobile apps) as well as in emerging economies.[22] A critical concern here is the power asymmetries whereby those with access to the data (for example, mobile phone operators and governments who set the rules for use of the data) can exercise control over populations without scrutiny or possibilities for rights violations.

2.5 Evidence-based policy or policy-based evidence?

Policymaking is increasingly being done by means of evidence-based approaches. **Pentland** (2014)[23] has brought a big data approach to bear on many **social problems**, from health and transportation to enhancing productivity in organizations by improving collaboration. **Clarke and Margetts** (2014)[24] review big data and open data in efforts at **government reform**, primarily in the UK, US and Canada but also around the world. They find that while there are many initiatives holding much promise, these efforts fall short because the limited feedback loop, whereby these initiatives could obtain responses from citizens and their efficacy could be evaluated.

Here it is interesting to reflect briefly on recent trends: among researchers who study policymaking, one view that has recently come to the foreground is how the evidence-based policy process has become highly politicized, leading some researchers to speak of 'policy-based evidence'.[25] In other words, Strassheim and Kettunen turn the conventional wisdom in this area upside down, arguing that policy drives nature of the evidence rather than the other way around.

---

[20] Schneier, B. (2015). *Data and Goliath*. New York: W.W. Norton.

[21] Schroeder, R. (2014). Big Data and the brave new world of social media research. *Big Data and Society*. July-December: 1-11.

[22] Schroeder, R. and Taylor, L. (n.d.). Big Data and Wikipedia Research: Social Science Knowledge across Disciplinary Divides, forthcoming in *Information, Communication and Society*.

[23] Pentland, S. (2014). *Social Physics: How Good Ideas Spread: The Lessons from a New Science*. London: Penguin.

[24] Clarke, A. and Margetts, H. (2014). Governments and Citizens Getting to Know Each Other? Open, Closed, and Big Data in Public Management Reform. *Policy and the Internet*, 6(4), 393–417.

[25] Strassheim, H. and Kettunen, P. (2014). When does evidence-based policy turn into policy-based evidence? Configurations, contexts and mechanisms. *Evidence and Policy*. 10(2): 259-77.

The reason for mentioning this is that this idea opens up a range of issues in relation to data and policy.[26] We can ask, for example, is the point of data-driven policymaking a way of underpinning policymaking with scientific or expert evidence, in order to make it more effective, regardless of the political preferences of policymakers or other interested groups? In this respect, what the research literature suggests about this attempted 'scienticization' is that this is not possible. Data are never neutral, and the more likely scenario is that the marshalling and selection of evidence (the data) takes place in order to promote the policies of a particular group, or that these data will be selected to be skewed towards a desired outcome, and that the very aim of promoting evidence-based methods is to push a particular agenda in the first place which will not be influenced by the analysis of data regardless of which policy they support or point to **(Kitchin 2014)**[27]. A particular aspect of this is that actors have different levels of resources to collect and analysis data. The idea that there is a competitive situation of 'whoever has the best data will benefit most for policy' is exemplified in several studies (see **Deetjen, Meyer and Schroeder 2015**[28], in the context of dementia research).

The reason for this brief discussion is twofold: the first is to acknowledge ongoing debates about the role of data-driven policymaking. Indeed, in some ways, there are merits in this understanding: overall, it is indeed plausible that policymakers are interested in deflecting decision-making onto impersonal procedures while being cognizant of the ways in which they can make these procedures work for their own political ends (see note 25). These debates harken back to much earlier debates in the 1960s, including **Habermas** (1968)[29] and others about the growing role of expertise in technocracies, and how in the end, more democratic policymaking can only be achieved not by means of scientific knowledge, but rather via building a normative consensus and that these inputs should guide decision-making.

However, a second reason for this discussion is to take issue with this view of the policymaking process and of data. Of course, the view of data in science as providing external evidence is an idealized one. And of course, many elements play a part in the policymaking process beyond data analysis and using this analysis as evidence. Further, policymaking has the aspirations of being scientific, but it is in fact oriented to particular goals rather than to contributing towards cumulative knowledge. Yet at the same time, the role of evidence (data) in policymaking, and how data and policies are linked, should not be dismissed.

To give some examples: to fix potholes, the link between reporting a pothole and fixing it is rather straightforward, even if here too there are shortcomings in the chain which links evidence to action.[30] For the International Panel on Climate Change (IPCC), on the other hand, the link may be rather indirect, since the determination of climate change is the product of many groups taking many considerations into account. At the same time, on an ideal-typical view, the link between data (temperature increase) and what to do about it (taking steps to decrease global temperature) is equally straightforward, or perhaps even more so, since there is single scale for temperature while potholes may come in many different and indeterminate shapes and sizes.

All this is simply to make the point that the data-driven nature of policymaking raises fundamental questions about the role of scientific knowledge in shaping society. This point, however, must be taken conjointly with the larger trends that society depends ever

---

[26] See also the report by the National Academy of Sciences (2015) "Using Science as Evidence in Public Policy", http://www.nap.edu/catalog.php?record_id=13460

[27] Kitchin, R. (2014). *The Data Revolution*. London: SAGE.

[28] Deetjen, U., Meyer, E., and Schroeder, R. (2015). Big Data for Advancing Dementia Research. Report for the OECD. http://bigdatadementia.oii.ox.ac.uk

[29] Habermas, J. (1968). *Technik und Wissenschaft als "Ideologie"*. Frankfurt: Suhrkamp.

[30] Mayer-Schonberger, V. and Cukier, K. (2013). *Big Data: A Revolution That Will Transform How We Live, Work and Think*. London: John Murray.

more on science, and that peoples' lives are being suffused ever more by the uses of data (most directly via the quantified self movement, but also in many areas such credit scores, online hotel rankings and much more). This report attempts an agnostic view on the issue of the scientific and data-driven nature of policymaking, even if there are fundamental debates of which we are aware but that fall outside the scope of our aims here.

To conclude, let us underline our agnostic and nuanced perspective by opting for a definition of big data that steers away from hypes or cynicism: "Big data is a step change in the scale and scope of the sources of materials (and tools for manipulating these sources) available in relation to a given object of interest."[31] This definition is more specific than, but compatible with META Group and Gartner descriptions that emphasise the four V's of big data: Volume (e.g. number of objects, time series), Variety (e.g. official statistics, sensor, social media), Velocity (e.g. real-time data) and Veracity (e.g. cleaning, noise, bias) and also allows the inclusion of other innovative approaches that may not be called "Big Data" as such, yet still can provide meaningful insights in the policy process.

[31] Schroeder, R. (2014). Big Data and the brave new world of social media research. *Big Data and Society*. July-December: 1-11.

3. Progress and expectations in specific policy areas

3.1 Introduction

This Chapter examines in which policy areas and in which stages of the policy cycle data-driven approaches are being used. It builds on the analysis of stakeholders (Appendix B), the inventory of relevant initiatives (summarised in Appendix C) and the interviews with thought leaders.

The inventory of relevant initiatives contains 58 initiatives. In short, the inventory analysis shows that either national agencies or national governments implement the majority of initiatives; however, also NGOs and other types of organisations play a significant role.

In relation to the policy areas concerned, a third of all initiatives in the inventory focus primarily on transport and mobility policy, environmental policy, information society policy, budget and economic and financial policy. The initiatives show that data-driven approaches are used also for cross-cutting and issues such as government transparency.

Almost half of all inventory initiatives focus primarily on the early stage of the policy cycle, i.e. agenda and priority setting. A significant number of initiatives use data in the later stages of the policy cycles, i.e. for the policy implementation and monitoring.

In this context, interviewees stressed the need to increase data literacy in the public sector.

3.2 Stakeholders

This study considers stakeholders at the level of organisations rather than individual persons. Policy stakeholders act at the global, European, national and the regional level. Policy stakeholders active at the global and European level include the United Nations (UN), World Bank and the European Commission (EC). The UN is involved in data for policy making via several departments and committees, notably with its Global Pulse initiative. The World Bank has also launched a number of experiments in big data. The main programme is 'Innovations in Big Data & Analytics for Development'. This programme includes the WBG Big Data Innovation Challenge. From 130 proposals, 14 winners received 1.5 million dollars in seed and growth funding in total.[32] The European Commission has increased its attention for data-driven policies in recent years, notably in the health, transport and environment areas. Policy makers at the global and European level dispose of significant financial means to support research projects and set up operational initiatives.

Several ministries and agencies at the national level are testing big data for policy and have, in some cases, already implemented initiatives. Important policy areas include transport, fraud detection and crime prevention. Data-driven methods are used also as a means to increase government accountability. National governments own large amounts of data as well as the financial means for implementing data for policy initiatives.

At the regional level, which is of inspiration to the national level, relevant policy stakeholders include municipalities, provinces and states as well as regional agencies concerned with traffic, resources and security. Besides associations of municipalities, networks such as EUROCITIES are exploring data for policy on the regional and local levels. Important policy areas on the regional level include traffic, water and waste management, security and public health. In contrast to the national policy stakeholders, regional initiatives focus more on the use of big data for policy implementation rather than agenda setting.

[32] For a full list of winners: http://blogs.worldbank.org/voices/meet-winners-and-finalists-first-wbg-big-data-innovation-challenge

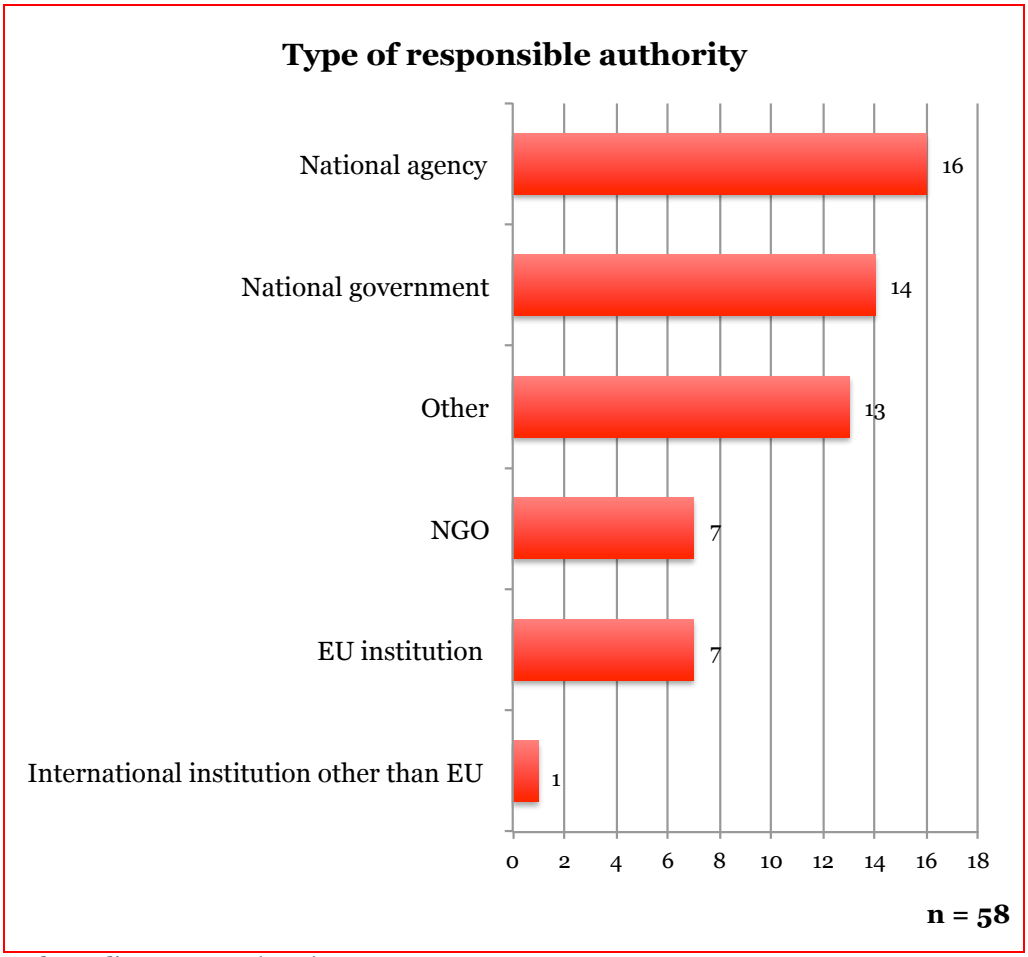
3.3 Inventory analysis

3.3.1 Types of authorities responsible for data for policy initiatives

National agencies and national government bodies (rather than international ones) implement the majority of initiatives in the inventory (Figure 1). National agencies involved are especially statistical offices but also geo-information agencies and education agencies. On the national government side (departments/ministries) the picture is more diverse. Departments involved in data for policy initiatives include economic affairs, telecommunications, transport and infrastructure, amongst others.

It should be noted that the inventory was collected with a clear focus on the national and international policy level, and a similar exercise focusing on a different level such as local policy setting would yield a different picture.

Figure 3 Type of responsible authority



Technopolis, OII, CEPS (2015)

Examples of usage are: the Spanish ministry for industry uses big data analytics to support its understanding of the structure and size of the ICT sector and to improve its evaluation process for awarding ICT R&D grants. The Dutch Statistical Office uses mobile phone data to gather information on mobility in order to support the design of disaster contingency plans or to improve commuter traffic, amongst others. The importance of transport and infrastructure as a policy area for data for policy initiatives is illustrated also by the German simTD project. This project aims at improving road safety and the efficiency of the traffic system on highways through the use of car-to-car and car-to-road communication.

In relation to the implementation status of the initiatives, it is interesting to note that around half of the initiatives implemented by national agencies are in the pilot or demonstrator phase (as agencies have some flexibility to experiment) while initiatives by national governments that we explored are mostly fully implemented as we favoured these more mature initiatives in our selection criteria.

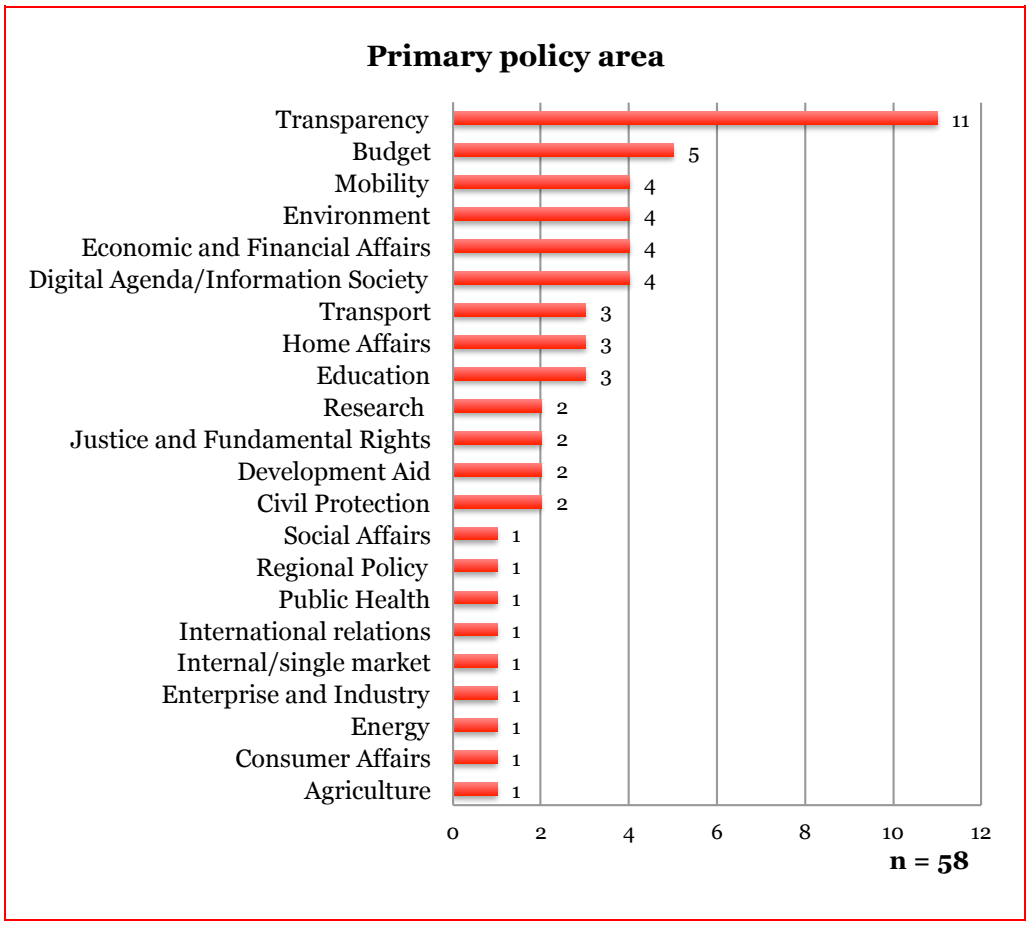
NGOs are also implementing innovative data for policy initiatives. Relevant NGOs include proponents of open government, who analyse available data in order to check on their governments' policies and processes in countries such as Slovakia, Hungary or India. NGOs in Canada and the UK focus on issues such as health and the general advancement of data science for the human good.

Data for policy initiatives are implemented also by other organisations that are not authorities or NGOs. These include state enterprises (such as the Centre of Registers in Lithuania), independent commissions (for example the Commission for the Prevention of Corruption in Slovenia), research consortia (like the German Smart Data for Mobility project), think tanks (Irish Fiscal Policy Research Centre), universities (TU Dresden), public-private partnerships (Digitale Delta in the Netherlands) as well as civic associations (such as the Italian Monithon initiative).

3.3.2 Policy areas

Figure 4 provides an overview of the primary policy areas, using a categorisation of policy areas as defined by the European Commission and adding transparency as a cross-cutting policy area/theme.

Figure 4 Primary policy area



Technopolis, OII, CEPS (2015)

The inventory includes initiatives in a broad range of policy areas. This illustrates the potential of data-drive approaches for, basically, all policy areas. Policy areas addressed by at least four initiatives: budget, transport (that is closely linked to mobility), environment, information society, economic and financial policy.

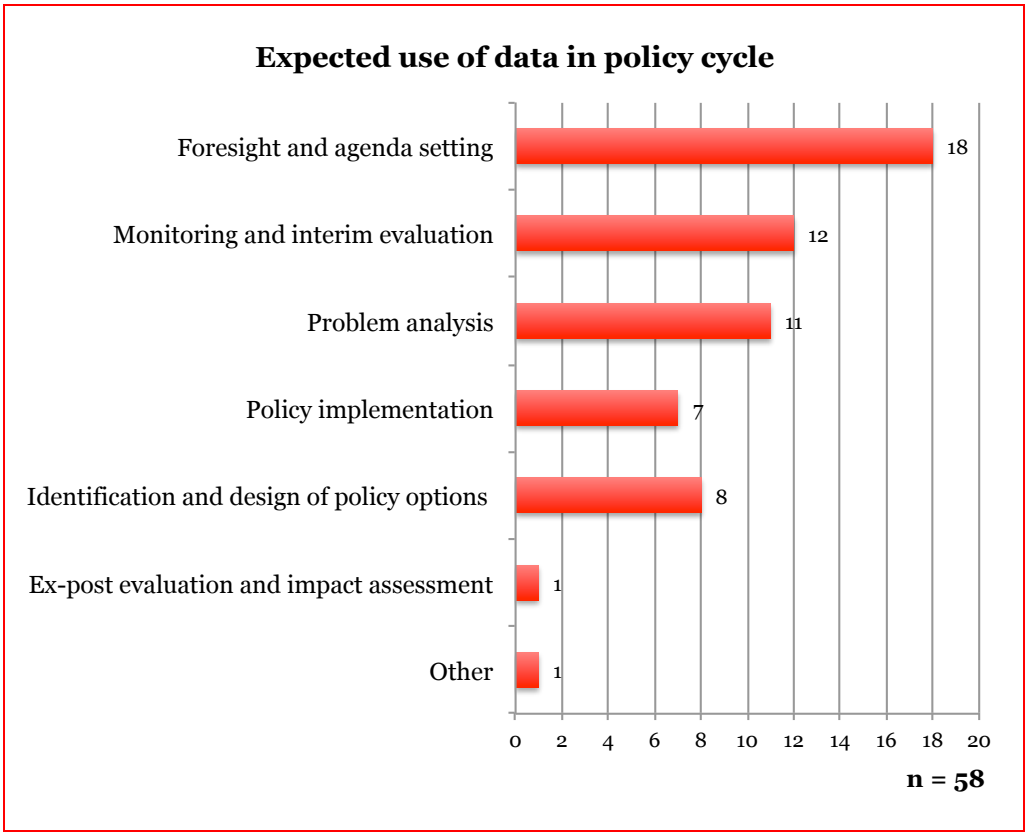
A substantial number of initiatives concerns government transparency, government accountability and citizen engagement. An example is the Tender Tracking Platform from Hungary, which offers a database of public procurement contracts in Hungary graded by a corruption risk scorecard. In Sweden, the Riksdag's facilitates public access to parliamentary documents and decisions, with the aim of increasing accountability. Public participation initiatives such as India's MyGov portal encourage citizen participation across all policy areas.

If one aggregates all policy areas addressed by the 58 initiatives in the inventory (including secondary and tertiary policy areas), economic and financial affairs as well as budget and transport are the three policy areas mentioned most often. A second observation is that nearly all initiatives are relevant for two or more policy areas. This indicates that data-driven initiatives cut across several policy areas, e.g. transport, mobility, environment and health (or environmental sustainability in general).

3.3.3 Policy cycle

Figure 5 shows the objectives and/or relevance of initiatives for the specific stages in the policy cycle.

Figure 5 Expected use of the data in the policy cycle



Technopolis, OII, CEPS (2015)

More than half of all inventory initiatives primarily focus on the early stage of the policy cycle, with 18 initiatives making use of data for foresight and agenda setting, 11 for problem analysis and 8 for the identification and design of policy options. An example is Singapore's Risk Assessment and Horizon Scanning programme, which explores

methods and tools that complement scenario planning for the anticipation of strategic issues. The DataKind UK project supports stakeholders in using their data to analyse areas in their work and to derive policies. Initiatives such as France's OpenFisca support policy makers in the identification and design of policy options by helping them to simulate the effects of tax reforms on their citizens.

Less than a third of the initiatives have a focus on the middle-stage policy cycles with seven initiatives using data for policy implementation and eight initiatives for monitoring and interim evaluation. For instance, the Dutch Ministry of Interior and Kingdom Relations uses open data and a web crawling of policy documents to monitor the transfer of tasks from the central government to municipalities and provinces.

Only one initiative in our inventory, the Greek PublicSpending.net portal, makes primarily use of data in the later stage of the policy cycle, where policy makers use the data for ex-post evaluation and impact assessment.

The overall picture looks similar also when considering the relevance of the initiatives for other parts of the policy cycle areas (a maximum of three, beyond the primary part of the policy cycle). However, there is a considerable increase in number of initiatives using data in the middle-stage policy cycle, such as for monitoring and interim evaluation purposes. This means that a large number of initiatives have a double use of the data: primarily in the early-stages of the policy cycle, but also for policy monitoring.

Whereas only a few initiatives primarily mention policy implementation, it is mentioned as second or third field by 58 initiatives. Initiatives aiming at policy implementation, such as the Star Metrics repository of data and tools to assess the impact of federal R&D investments, are implemented by national agencies (in this case: the US National Institute of Health). This observation is in line with interview statements (see below) indicating that specialised national agencies make a closer link between the design of policy and policy implementation (e.g. for laws, regulations and support programmes).

3.4 Insights from the interviews

Interviewees confirm that several types of authorities/organisations (including NGOs) are responsible for data for policy initiatives in their respective countries. On the one hand there are organisations that have traditionally been working with large datasets, such as statistical offices, financial and economic departments of national ministries, planning agencies as well as meteorological institutes. On the other hand, a number of interviewees mention that specialised, operational authorities are leading in the adoption of innovative data-driven methods (e.g. authorities in the field of transport and environment).

Interviewees stress, however, that the more innovative examples can be found at the city or regional levels (e.g. smart cities and living labs). A major part of the reason for this is that it is, often, easier to get, local and regional policymakers involved in new, small-scale initiatives, new collaborations (that require trust and flexibility) and new ways of making data available and using data for policymaking (that require risk-taking).

Concerning policy areas, interviewees repeatedly mention health as an important policy area, where potential is seen in the use of genomics, biomarkers and self-measurement. Another policy area that was mentioned by several interviewees is security, crime prevention and education. In South Korea, for instance, policy makers focus on safety and disasters as priority policy areas. With respect to security policy, it was also mentioned that hierarchical and centralised approaches, that are common, might be less effective and resilient than collaborative and decentralised approaches (involving citizens, local agencies, international collaboration, de-central data storage, etc.).

Moreover, it was explained that data-driven methods especially benefit policy areas with large volumes of data, such as tax, health and transport policy. However, as one interviewee believes, departments and agencies concerned with these policy areas are usually more traditional and follow strict hierarchies so that smaller departments may

move quicker albeit with less volume and policy impact. Potential for larger use of data is seen in the prevention of crime and terrorism.

Regarding the relevance of data-driven approaches for specific parts of the policy cycle, a number of interviewees point to policy implementation as an important field of application. Besides policy implementation, agenda setting, developing and choosing between policy options benefits from big data and policy analysis (by statistical offices and specialised agencies). As one interviewee points out, the fact that data becomes available more quickly also leads to an increased use of data in policy evaluation, implying that the policy cycle might become shorter.

Looking ahead, interviewees point out several challenges and opportunities. Data literacy, especially for policy-makers (at all levels of public organisations), is seen as an important issue in the near future. A number of interviewees see a potential skill gap between the public and private sector in this context. Skills are not only relevant for using more data within government but also to negotiate and collaborate more effectively with private actors (e.g. about access to private and public sectors and procurement of tools).

4. Data sources

4.1 Introduction

More types of data sources are becoming relevant for policymaking. This leads to a larger number of types of stakeholders active in data markets.

In these markets, government bodies are buyers and sellers/suppliers of data, next to their adoption of an open data strategy.

Compared to the (mature) initiatives included in the inventory, research projects in the European Commission's Seventh Framework Programme put more emphasis on 'new' data sources such as micro sensors, owned by citizens.

It has also become common practice to combine different types of data sources in one initiative. Still, there are opportunities in many policy areas to use more types of data sources, move to open data formats and improve data interoperability. Interviewees stressed in this context the importance of analytical frameworks and sound indicators ('so that the data makes sense') and the opportunities (and challenges) for statistical offices.

Privately held data makes up the largest source of digital data since it includes not just social media and search engine data, but also economic transactions of all kinds and records of services or, for examples, goods shipped around the world. Some of these data are partly restricted for legal reasons and because they provide competitive advantage. However, it has often been pointed that these data lose their benefits in terms of competitive advantage very quickly, such that it would be possible to give these data to researchers or others (for example, airline travel data in the service of examining disease spread) without loss. However, in this case, the issue might be the effort that it takes to anonymize the data or put it into suitable formats.

Some privately held data is being shared with researchers, as with Twitter data being provided to UN's Global Pulse. But these types of agreements, providing access to some researchers and not others, raise the problem of unequal access, perhaps even a 'digital data access divide'. Second, there is the issue of reproducibility and being able to 'open the black box' of how the data was collected. On the other hand, there are many types of privately held data which have been made available, such as Google Trends or samples of Twitter data and many other types of data which can be scraped from the Web (as with the Billion Prices Project).[33]

4.2 Stakeholders

The number of types of stakeholders involved in data collection is increasing in parallel to the increase in the variety of data sources.

Statistical offices and the policy analysis/evaluation community continue to invest in national and international statistics and bespoke surveys (large scale and small scale). At the same time, they explore data collection based on social media and sensors. For instance, statistical offices in Canada, Finland, Italy, the Netherlands and Slovenia experiment with social media data and/or sensor data (to measure consumer confidence and monitor transport). Another example includes work done by Eurostat. Data about prices available online, obtained by means of web scraping, is tested as an additional data source for tracking the Consumer Price Index.

The examples above illustrate that social media firms and organisations with a network of sensors (cf. the Internet of Things) have joined the group of relevant stakeholders. The same holds for mobile phone operators whose data is relevant for policy areas such

[33] http://bpp.mit.edu/

as transport (e.g. congestion), health (e.g. epidemics), safety and security (e.g. earth quakes and intelligence).

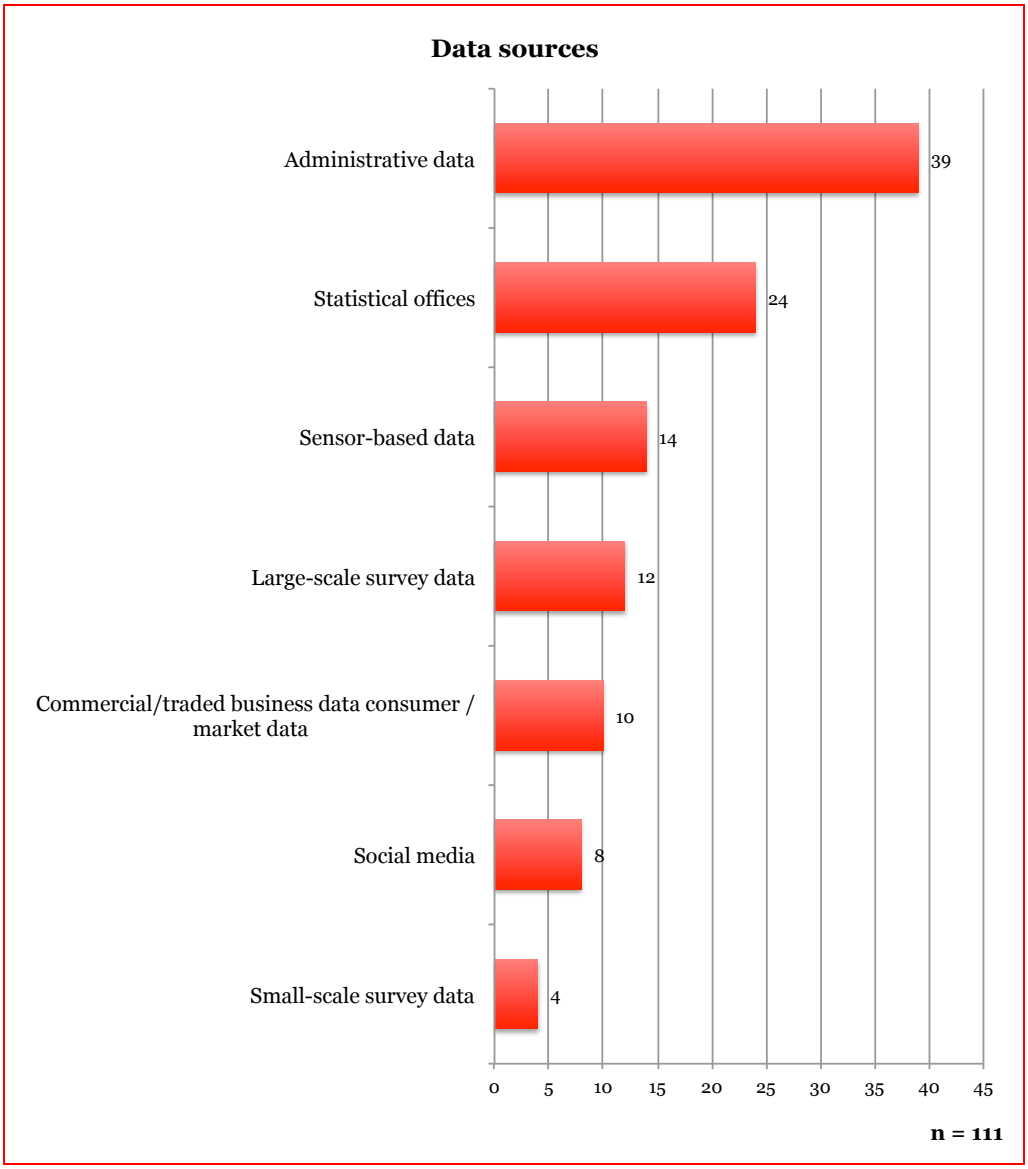
The open data phenomenon implies that public administration becomes more important as providers of (open) data for policymaking but also for commercial practices and for initiatives and campaigns by NGOs. Another opportunity for government bodies, as a buyer and seller of data, is to benefit from data brokers. One example is firms that specialise in securing access to a set of datasets (e.g. geodata, financial data and consumer data), link datasets and provide a suit of data sources and data analytics to industry and government clients.

4.3 Inventory analysis

Data sources and data velocity

The inventory of relevant initiatives (Appendix C) indicates that a variety of data sources are used to inform policymaking (Figure 6). Administrative data, either as open data or restricted data, is used in more than half the of the initiatives. Official statistics and (other) large-scale surveys continue to play an important role.

Figure 6 Data sources



Technopolis, OII, CEPS (2015)

Sensor data, commercially traded data (obtained from data brokers such as Datalogix and Epsilon), social media data and other data types (such as clickstream data and geodata) are used less frequently. Examples are the use of sensor data for transport and health; the use of commercially traded data for economic and financial policy and international development policy; and the use of social media data for transport and environmental policy.

Other data sources used include online papers, articles and blogs about research (web scraping and text mining to inform research policy); and policy documents available online (web scraping and text mining by a national policy maker to assess whether regional policy makers are preparing for taking over responsibilities that will be handed over for national to regional governments).

Data velocity to a large extent depends on the data source. Still, it is possible to observe differences in levels of velocity within one type of data source. For example, administrative data can be collected and made available real-time, on a monthly basis or once per year (e.g. using open data to continuously update court rulings, to provide monthly updates about donations to politicians or annual overviews of the number of inhabitants). Social media data is high velocity. With mobile phone data, the velocity may depend on the intended use, e.g. monitoring traffic during rush hour or intensifying data collection in case of natural disasters. Sensor-based data is collected real-time or in some cases on a daily basis. The data analysis may take place at regular intervals of days, weeks or months.

Research projects in Europe's Seventh Framework Programme (Appendix E) focus on data sources that are relatively new in the context of policy making. Examples are micro sensors, owned by citizens, to monitor air and water quality (cf. crowd sourcing), sensors along routes or in urban areas (e.g. highways and smart cities), documents available on the Internet (web crawling and text mining), social media data (e.g. sentiment analysis) and mobile phone data (e.g. combined with transport data based on sensors).

Data formats, data interoperability and data linking

Across the 58 initiatives, three data formats were mentioned more than 10 times:

- Microsoft Excel,
- CSV (Comma Separated Values standard file format), and
- SQL data format and programming language (Structured Query Language).

Other data formats mentioned are:

- GIS database format (Geographic Information Systems),
- RDF (Resource Description Framework), and other structured web-based formats including XML (Extensible Markup Language) and JSON (JavaScript Object Notation),
- HTML (Hypertext Markup Language),
- PDF (Portable Document Format), and
- Microsoft Word but also plain text (that still allows for web crawling and text mining).

In terms of the different levels of data openness, as developed by Tim Berners-Lee[34] and applied in the context of open data, the inventory provides a nuanced and positive picture. Still, there are opportunities to move from data formats such as HTML, PDF, Excel and Word (with limitations in terms of data processing and linking) to formats such as CSV and RDF that can be processed with different types of software and databases. CSV is generally the lowest-common denominator for file sharing of data, since the non-proprietary text-based files containing nothing but data values with commas separating them can be read by virtually any computer system. Other formats, on the other hand, require specialist software (some of which is free and/or open-source and others proprietary).

The inventory indicates that combining and linking different data sources has become one of the main characteristics of innovative data-driven approaches for policymaking (Figure 7).

Figure 7 Initiatives with multiple data sources

Number of data sources	Number of initiatives	Policy areas
1	14	In many policy areas, particularly multiple initiatives in economic and financial affairs and research
2	28	In many policy areas, with multiple initiatives in environment, justice and fundamental rights, economic and financial affairs and transport. Moreover many initiatives using two data sources aim at increasing the transparency of government spending, procurement and decision making,
3	11	In several policy areas, particularly multiple initiatives in the area of digital agenda/information society.
4/5	5	Few policy areas, one initiative each in the following policy areas: transport, regional policy, education, digital agenda/information society and agriculture.

Technopolis, OII, CEPS (2015)

Challenges in data interoperability and data linking were only touched upon in the inventory (e.g. initiatives that use unique identifiers of persons, cars or firms and that take a step-by-step approach in using open data formats, adding data sources and linking data). On this specific topic, the websites of the relevant initiatives provide little information. Moreover, coordinators that shared information with us did not elaborate on this topic. To some extent, data interoperability and data linking is addressed in the interviews.

4.4 Insights from the interviews

When asked for a reflection on trends in data sources for data-driven approaches for evidence-informed policymaking, interviewees addressed the following aspects: indicators, open data, transaction data, data at aggregate or individual level, data interoperability and data linking, data markets and data exchanges, and the challenges for statistical offices,

The terminology and focus of the big data trend should not imply that analytical frameworks and well-designed indicators are no longer the basis for data collection.

[34] T. Berners-Lee (2006). Linked data: Design issues. http://www.w3.org/DesignIssues/LinkedData.html and Janowicz, K., Hitzler, P., Adams, B., Kolas, D., & Vardeman II, C. (2014). Five stars of linked data vocabulary use. *Semantic Web*, *5*(3), 173-176.

They are a prerequisite for data analytics and sense making. Even when explorative approaches are taken, such as data mining, one still needs to reflect on the relation between the data itself and the variables or phenomena to which the data refers (cf. internal validity).

The importance of open data was emphasized. Political will is crucial for enabling open data, also to have constructive debates about safeguards for making data publicly available (without taking an absolutist approach towards privacy risks). Still, one should acknowledge that policymakers - and governments in general - may be more risk-averse than industry. Politicians are accountable to parliament, citizens and press. Hierarchical structures may not stimulate open data or being innovative in collaborating with researchers and firms.

Transaction data (such as credit card data) are expected to become more important, as transactions by citizens contain indications about mobility, health and social inclusion. As such, transaction data need not only serve commercial purposes.

One of the most important ways of distinguishing between data sources is between data at aggregated level (about a neighborhood, city, country, etc.) and data about individuals (human centred). In the policy area of safety and security, this difference should be stressed. In general, safety is mostly about the aggregated level (e.g. crime or fires in specific neighbourhoods) and security is mostly about the individual level (e.g. profiling, tracking, tracing potential terrorists). When new data sources are used, or when existing data sources are combined, the analysis can move from the aggregated to the individual level. If so, the ethical issues should be discussed.

The importance of data interoperability and data linking and the challenges in this sphere should not be underestimated. For instance, open data often derives from a range of departments and public agencies. They often have different identifiers of citizens or firms, different meta-data systems, different data formats, data storage solutions, etc. The challenges are considered to be more substantial at national and international level compared to the local and regional level (with many data owned by local governments and possibilities for small-scale experiments).

Data markets are becoming more important because a) more public and private organisations collect data as a by-product of their activities such as firms that collect data as a by-product of their primary process (e.g. retailers, car leasing and energy suppliers) and b) more organisations need data to improve their processes and products. Specialised data brokers or even data exchanges provide the intermediary function between supply and demand of data.

The importance of statistical offices and other well-established organizations was stressed (e.g. financial and economic departments, planning agencies, environmental and meteorological institutes, social science databases[35]). If they continue increasing their expertise in data collection, data quality and data processing, they are well positioned to combine traditional and new data sources. One of the preconditions is to do research and experiments on how to use of these new data sources.

[35] For example, http://www.cessda.net/about/members.html

5. Data analytics and data visualisation

5.1 Introduction

Analysing data and data visualisation are well-established fields, but they have taken off in new directions in recent years with new tools like Hadoop for distributed storage and processing of data and the widespread use of tools for the visual display of online information. For data analytics, it is necessary to distinguish between tools that can be used on the desktop and those done on powerful high-performance machines that few have access to – with a range in between. The expertise in being able to use these tools similarly ranges from those which students and others can learn quickly to analyse the content of Twitter for example to tools which are at the cutting edge being developed and used only by those with advanced computing degrees and/or years of professional use. The same applies to visualisation tools except that this area, due to big increases in the capabilities of desktop systems over the last 15 years, is currently not as demanding in terms of techniques and expertise. For policy, the issues are mainly to do with the quality and usability of analyses. The analysis of relevant initiatives as well as the interviews with thought leaders indicate substantial opportunities in moving to more advanced data analytics and visualisation.

5.2 Stakeholders

A general point that can be made here is that most stakeholders, including the private sector, make use of public large-scale datasets that have been collected for some time. In this respect, according to our interviewees, the main innovation comes not from new data analytic techniques, but from combining and linking datasets to produce new insights. Among Science and R&D actors (researchers but also policy makers) there have been initiatives to measure the inputs and outputs of science. One example is patent analysis in projects such as Star Metrics[36] (now hosted by the National Institute of Health) which take the approach of combining various traditional and new data sources. What is particularly important here is to enhance the granularity of the data in order to be able to arrive at more targeted interventions. The SciMaps[37] project, led by the School of Informatics and Computing at the University of Indiana, has been at the forefront of innovative visualisations relevant to science policy, and has created an evolving exhibit of science maps that have been displayed at events and institutions around the world.

Data brokers are an emerging type of stakeholders (as discussed in Chapter 4). They can integrate different economic activities, such as data collection, data cleaning (that can be an expensive step), data linking, data trading, data analytics and data visualisation. There is also sector specific expertise in data analytical tools, for example tools relating to supply of the public sector, where for instance, CKAN[38] is a tool for managing open source government data. Organizations like the Open Knowledge Foundation[39] are often contracted by governments to implement tools such as these. Private providers of data analytics and visualisation tools are in the lead in providing new analytic tools: these include companies such as Datasift[40], GNIP[41], Splunk[42], and Google[43] with tools for

[36] https://www.starmetrics.nih.gov/

[37] http://scimaps.org/

[38] http://ckan.org/

[39] https://okfn.org/

[40] http://datasift.com/

[41] https://gnip.com

[42] http://www.splunk.com

[43] www.google.com

analysing trends. Others use public data for analytics: The GDELT project[44] ingests public news data from 1979-present and repackages it for analysis and visualisation. Splunk generally works with individual firms to provide tools for analysing internal data, and visualizing the data for improved decision making. An example of a smaller start-up is Canopy Labs[45] which specializes in providing high-end data analytics and visualisation capacity to SMEs. An academic spin-off is Crimson Hexagon[46] which specializes in analysing social data related to organizations.

Among civil society actors, there are often online tools and openly available geographic tools for mapping. Among global and European policy makers, data analytics tools come via major European Commission research initiatives, such as in health and transport and the environment, or for example in the case the United Nations global pulse from sentiment mining tools in the analysis of Twitter. For National policymakers, it can be noted that the most distinctive analytics techniques here are in the area of crime and terrorism prevention, which are by their nature kept secret.

The main concern of statistical offices is to use data analytic techniques that are accurate, and to ensure that this also applies when new sources of data and techniques are used such as social media.

5.3 Inventory analysis

5.3.1 Data analytics used depends on data sources and data formats

As discussed in Chapter 4, a variety of data sources and data formats are used. From (open) administrative data, official statistics, surveys and commercial/business data to social media data, sensor data, social media data and sensor data. Data formats include long-established formats such as CSV or Excel files, but also newer web formats like JSON and XML. Some also use advanced formats for interlinked data such as RDF and SQL. There are also examples of highly specific types of data that require specialized tools to work with the data, for example extracting number plate data for comparison to police and security databases. Some data sources provide API access, although a few still only provide the data via HTML or text or PDF, which limits its reusability.

At least half of the initiatives in the inventory appear to have given thought to the question of whether people outside the organization can obtain some sort of access to and use of the data, but these still exists in a wide enough variety of formats that anyone wishing to pull data from many sources would have to deal with each one individually. As mentioned in Chapter 3, the use of sensor-based data, social media data and text mining from the open web are becoming more important.

At the intersection of data sources, data formats and data analytics is the issue whether national and international initiatives (such as those included in our inventory) are facilitated by common, international standards for data (content and technical) that would allow aggregation and comparison. For instance, the Groningen Declaration for international portability (and student ownership) of personal student data, is addressing the issue of various national and international standards related to skills, qualifications and the data formats used.

5.3.2 Data analysis tools

In this category, we find perhaps the widest variety of approaches, all the way from simple desktop based analysis such as Excel or statistics programmes through more centralized data services such as SQL databases or cloud-based solutions and on to a significant number of initiatives using high-performance computing to analyse data.

[44] http://gdeltproject.org

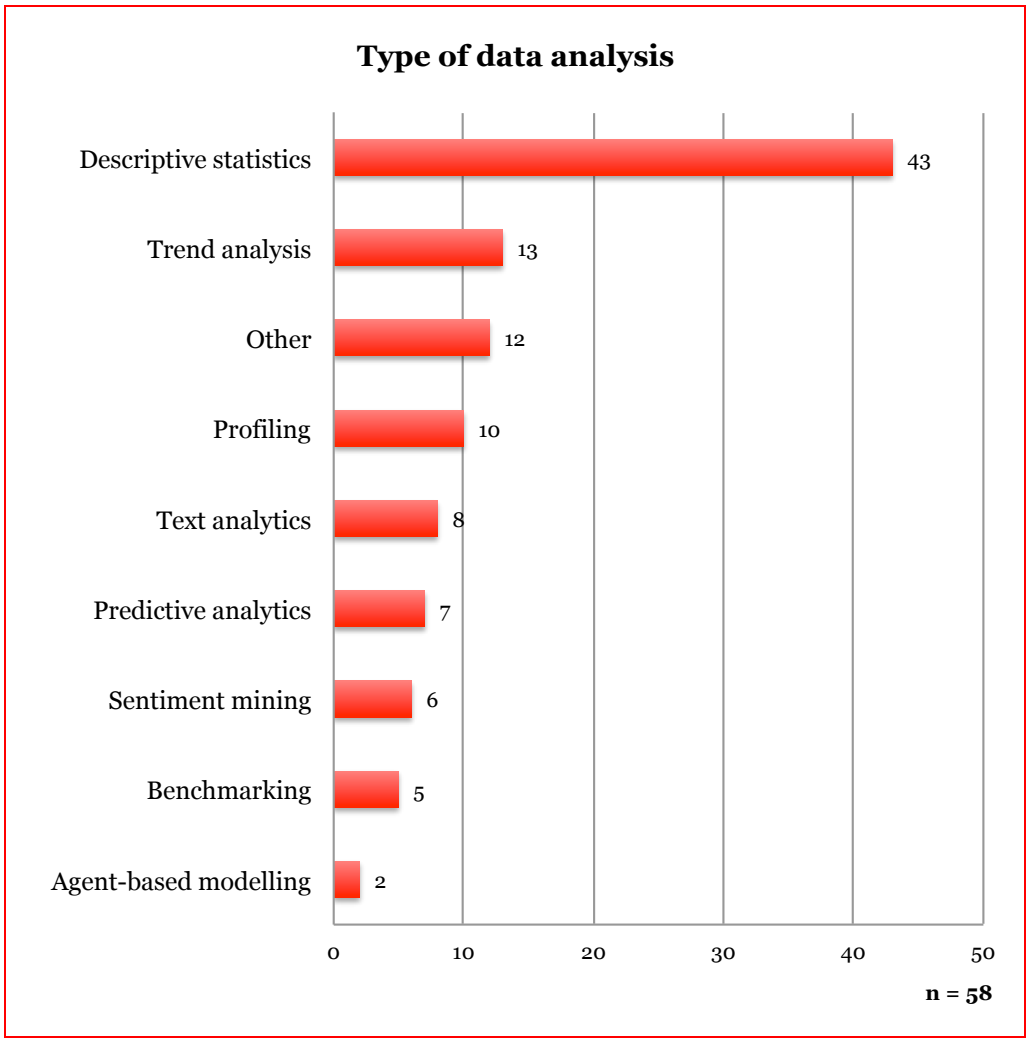
[45] https://canopylabs.com/

[46] http://www.crimsonhexagon.com/

Likewise, in terms of the data analytic approaches, these run the gamut from simple descriptive statistics and benchmarks through trend analysis and profiling all the way to advanced techniques such as sentiment mining and advance statistical analysis and predictive modelling. Only one project (Using Big Data for the Evaluation of R&D grants in the ICT sector in Spain) used Natural Language processing (or machine learning) to evaluate grants at the submission stage. However, a handful of projects used text analysis and sentiment mining, which could potentially involve machine learning.

Much of the data is sourced from administrative data and data from statistical offices (Chapter 4). These are updated anything from real-time to annual or less. Overall, we can see that the large bulk of data analytics (43) are descriptive statistics, most of which will pertain to the data collected by public bodies (Figure 8).

Figure 8 Data analysis



Technopolis, OII, CEPS (2015)

As to the remainder, it is a mixture of new techniques such text analytics and sentiment analytics or intensified existing techniques such as profiling and predictive analytics. [47]

[47] See, for examples, http://www.datasciencetoolkit.org/

5.3.3 Visualisation tools

Some of the more common visualisation tools are for displaying information geographically, for instance: specialist GIS tools including Arc-GIS or more general-purpose tools such as Google Maps. Since many of these initiatives are bounded by political geography, the choice to include mapping information spatially makes sense: for example, locating where money is being spent or how natural resources are distributed or where traffic is most congested. There are also visualisations that have to do with visualizing the patterns in a wide variety of data. These range from simple embedded Excel sheets to Word Clouds and network graphs and all the fully designer created infographics. Finally, an emerging area is visual analytics, which, as the name implies, is aimed at supporting analysis with visualisations.[48]

Interestingly, the majority of initiatives in the inventory either do not currently have any visualisation on their public-facing websites, or have some visual information where it cannot be determined how it was made. At this point, relatively few of the surveyed initiatives make data visualisations openly available, and even fewer make interactive visualisations openly available. It is not clear from this inventory what is the extent to which agencies are using interactive visualisations internally, which might be more frequent. It is likely that they all have data in bar and graph charts. The skills needed to operate Excel and Powerpoint are relatively low, but higher level skills are in short supply and medium skills will require some training or self-taught skills (Google Maps).

It seems that visualisations are often used by initiatives that need to explain complex numerical data such as budgets. Visualisations are then built often in order to help citizens to follow up and monitor the budget in specific areas. This is the case, for example, for the Italia Aid and Open Coesione projects, which provide a tool to help citizens to understand the budget distribution. While this is not an innovative use, it is a tool that is used on the government side for the citizens, but not for the government's own uses.

A scan of relevant research projects in Europe's Seventh Framework Programme (FP7) also indicates that data analytics is getting more emphasis than data visualisation. Examples in the Global Systems Science cluster of FP7 include the use of curated timelines (e.g. the EU Community project) and dynamic visualisation of group-based modelling (e.g. the GRACeFUL project). A number of relevant FP7 projects are still running, with visualisation often being the last work package.

It can be noted that although many examples of innovative visualisations exist in the academic literature, such as the visualisation of very large-scale of large-scale social networks the maps of science produced by Katy Boerner[49], in policymaking there is perhaps more of a need to make visualisations accessible on the web or usable, which may produce innovation more of in terms of best practice rather than advanced techniques.

5.4 Insights from the interviews

Interviewees discussed data analytics and visualisation in generic terms since they tended not to be involved in the details of using the tools themselves but were rather involved in data-driven approaches at a higher level (policymakers and academics using, rather being part of the data analysis process). The interviews revealed two rather separate strands of innovation. On the one hand, new techniques such as sentiment analysis, location data, trends in search engine keywords and the like, which apply novel techniques to social media and online phenomena among the public. The second strand were those who were interested in public datasets and official records, where analytics

[48] http://www.visual-analytics.eu/

[49] http://ella.slis.indiana.edu/~katy/

techniques were not central since the analysis here involves rather conventional techniques, except perhaps in aggregating and linking data sources.

In terms of visualisation, the main issue mentioned by interviewees repeatedly related not to new tools and techniques, but the need for policymakers and a wider public to be able to read and understand the new complex visualisations properly, an area where it was felt that there was currently a lack of knowledge. Put the other way around, there is a perceived danger when data analytics and visualisations are used that these could be used, without proper understanding, to draw the wrong conclusions or arrive at misleading pictures.

6. Privacy and inclusion

6.1 Introduction

Privacy has become a major concern, but it is necessary to distinguish between what is happening with the widespread media coverage of the issue in view of the Snowden revelations, Wikileaks and phone hacking scandals, as against the privacy issues to do with the security of data in data-driven analysis for policy. Further, an issue which had not been traditionally a matter of concerns about privacy – which is how data-driven analysis can pinpoint individuals and predict behaviour – is coming to the forefront even if it has not been the subject of privacy legislation and regulation as such in in the past.

In terms of inclusion, the issue of who gets left out in data-driven analyses has often been a matter of concern. This question of what or who gets left out by digital data was not noticeable as a major concern in policy-related analysis except in the traditional sense of the digital divide; i.e. people left out by not having internet or mobile data about them. Capturing too much information about individuals and blind spots in not capturing enough can be seen as mirror images, though it should be noted that 'what' gets left out can also include non-humans; i.e. areas for which no digital data exists in terms of the physical environment that is left out of sensor measurement and the like.

6.2 Stakeholders

Among science/R&D organisations, privacy is a research topic (e.g. in social sciences) as well as a consideration in designing and developing new tools (e.g. in computer sciences and in engineering).

In industries that facilitate data for policy (including mobile telecoms), an abundance of data is available, with privacy concerns leading to an under-utilization of data. Collaboration with governments and other stakeholders, including data sharing, might be hindered by unclear or out-dated legal frameworks. Privacy is an important issue and concern for private providers of data analytics and visualisation tools. However, their data protection strategy is also driven by economic incentives. The data sets in this area are often made purposefully opaque to protect the providers from losing their source of revenue. Data brokers are of course highly concerned with privacy, and this area is currently subject to legal and regulatory developments ('right to be forgotten', regulations concerns data that may be used for health and credit purposes).[50]

Civil society actors are trying to open up institutions, and so have few of the difficulties (or far less so) that beset other big data applications such as privacy. Here instead the question concerns the quality of the data and the way it is organized for making use of it.

The main privacy concerns are among global and European policymakers, where there may be different laws or unclear laws applying to different jurisdictions. For national policymakers, the key privacy issue is the collection of excessive amounts of data. Regional policymakers instead face the issue that, in trying to achieve transparency, privacy may be violated (e.g. from data about cities and neighbourhoods to data about streets and households). Statistical offices are concerned mainly with safeguarding privacy when datasets can become de-anonymized through linking data or highly granular and complete data (where privacy can be violated because people can be identified even if they are anonymous through this granularity; i.e. a person who is known to have visited a hospital on a particular day and hence a unique health event taking place in that hospital on that day). The parallel with data about industry is disclosing data about individual firms (micro data).

[50] Pasquale, F. (2015). The Black Box Society: The Secret Algorithms that Control Money and Information. Cambridge MA: Harvard University Press.

For the policy analysis and evaluation community, but also for governments, a key concern may be rather with inclusion. This mainly refers to the use of social media, mobile phone records or sensor data that are unrepresentative. An inclusion issue for the research community itself is 'gaming' the data collection procedures to assess science (e.g. bibliometrics) with perverse effects of rewarding areas of research that are better at promoting themselves than at doing research.

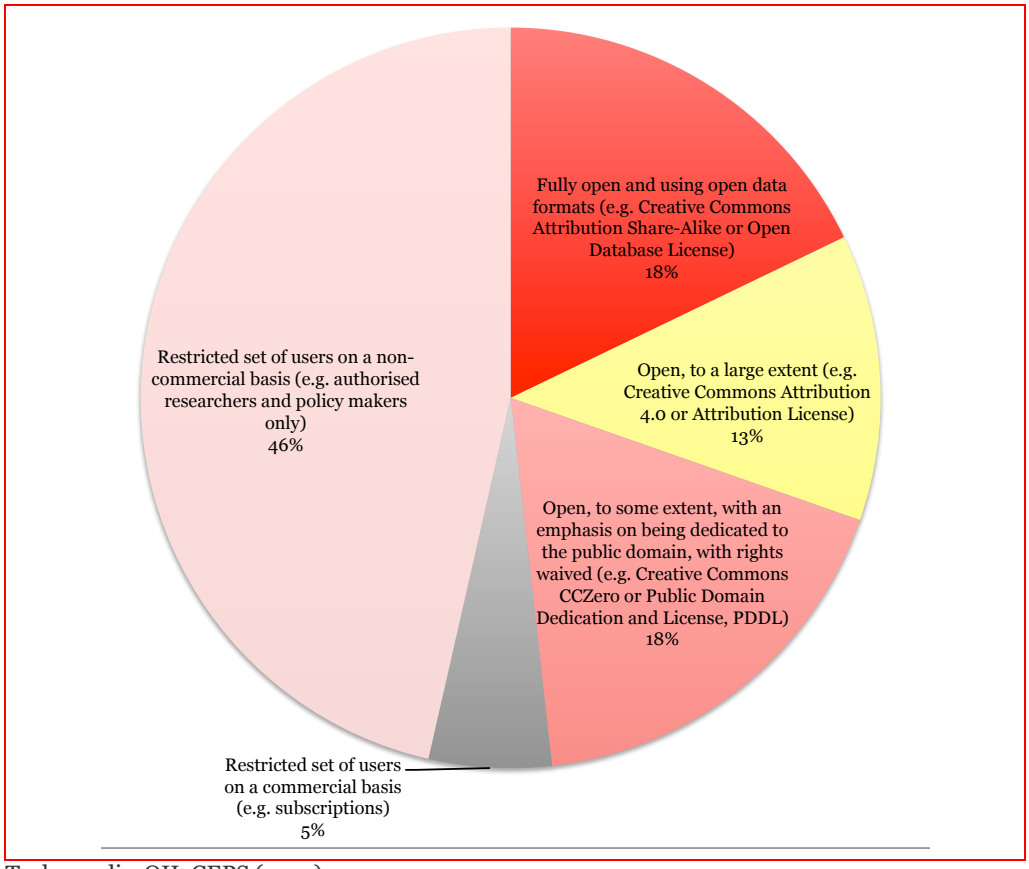
6.3 Inventory analysis

In terms of privacy issues, initiatives run the gamut from extensive and detailed privacy policies all the way to having no privacy policy listed at all. The most common way of dealing with privacy is that personal data has been removed, anonymized, or data have been aggregated so that individual responses cannot be connected to persons. Some of the initiatives, especially those gathering sensor data not linked to persons, mean that privacy issues are considered irrelevant (although there are cases between persons and a more aggregate level, for example households, which do not easily fall in this category). Unsurprisingly, the data that is most closely connected to individuals is considered most sensitive and is most likely to have policies in place for protecting privacy.

For many initiatives where data is provided on websites, there is a link to a privacy policy relating to how the data were collected or how they may be used. However, this relates mainly to sites where the data are relatively innocuous in the first place and mainly serves as a means of reassuring the public that nothing untoward has been done in the data collection process or that steps have been taken to prevent misuse.

The insights from the inventory, regarding privacy, partly depend on the extent to which the data can be accessed by large or restricted sets of users (Figure 9). If data is available only to a small set of registered users, privacy concerns tend to be smaller (although there are exceptions to this rule of thumb).

Figure 9 Openness of data



Technopolis, OII, CEPS (2015)

Relatively few of the initiatives have a clear consideration of inclusion issues. For many of the inventoried initiatives, inclusion issues were either not relevant or not mentioned. Inclusion issues do come up with regard to access (to the internet, to social media, to mobile phones, and the like) and thus to being represented. So, for instance, missing datasets, non-responding individuals or organizations, or sampling issues can result in unintentional or systematic bias in the data. For a number of national initiatives where, for example, data are being collected from different organizations (such as regional or local government, or where crowdsourcing or volunteering is used), the main inclusion issue is that it is not clear how the completeness or the updating of the data is taking place, so that uneven or out-of-date coverage may emerge as the main problem over the longer term in making data sets less useful or usable.

6.4 Insights from the interviews

A first set of concerns about privacy came from those involved in the public sector. These experts stressed that while the protection of privacy was being taken into consideration at the aggregate, there could be risks when moving from the aggregate to the personal level. Another area mentioned was that while the aggregate level often did not pose a risk directly, there could be unforeseen consequences such as when environmental or transport data analysis could affect property values for example.

The second main privacy risk was in the area of security or crime and terrorism prevention. Obviously the use of data-driven analysis, sifting through trends and looking to prevent crimes and fraud is thought to be highly useful (and, it is worth adding here that this was seen by several interviewees as the single most advanced area of actually using data-driven analysis). At the same time, what if a worrying pattern concerning an individual or a group is spotted, what should be done then? Can action be taken concerning the individual or group to prevent perceived wrongdoing, without evidence? Will mistakes be made, possibly leading to harm?

Third, it can be mentioned that apart from the question of informed consent in sensitive areas such as health (where there may also be a need for new consent models, as when data can be linked), new questions are arising as whether there need to be new more explicit models of consent when, for example, mobile phone companies or banks collect data.

As for inclusion, a point made by policy researchers especially is the extent to which, for example in the monitoring of student performance, the singling out of particular groups of students as performing poorly and in need of attention could lead to self-fulfilling prophecies (that policies are then targeted at this group which creates such a group and thus prevents a more inclusive or universal approach). Also relevant here could be seen the need to make much more use of commercial data. Working with such data, however, is bound to lead to targeting certain populations rather than others. An example of a pro-active approach towards inclusiveness can be given from the New Zealand national data strategy, where Maori values are deliberately taken into consideration in the design of the data policy, including in the form of representation from the Maori community.

7. Conclusion

This report covers much ground in a short space, and we hope it provides inspiration for policymakers and other stakeholders in data-driven approaches for evidence-informed policymaking.

By way of conclusion, we summarise the main points and provide an outlook.

- An emerging body of literature is addressing the **possibilities and limitations of data-driven approaches** in the context of policymaking and, more in general, societal challenges such as health, security and environmental sustainability. Among the key topics are validity, transparency and replication (for example, related to the use of social media data and mobile phone data), privacy, inclusion and the influence of stakeholders on data collection and data analytics ('data is not neutral'). Moreover, leading authors acknowledge the challenges in data standardisation/harmonisation, data linking and the costs of data collection, cleaning, storage, etc.
- The inventory of 58 relevant initiatives, combined with the interviews, indicate(s) that innovative data-driven approaches are implemented in a **broad range of policy areas**. Policy areas addressed by at least four initiatives include: budget, transport (that is closely linked to mobility), environment, information society and economic and financial policy. No less than 11 initiatives seek to increase government **transparency**, government accountability and/or citizen engagement (e.g. data about procurement, policy procedures and funding of political parties). Literature and experts point to innovative uses of data in areas of crime, fraud and terrorism. However, there are also concerns that it may not be in the DNA of hierarchical, centralised organisations to explore decentralised and innovative data-driven approaches.
- In terms of the policy cycle, nearly all initiatives focus on either the **first part of the policy cycle** (foresight and agenda setting, problem analysis and identification and design of policy options) or **policy implementation and monitoring**. Few initiatives address ex ante or ex post evaluation and impact assessment of policy.
- The main type of data sources, at least according to the inventory, is **public datasets** such as administrative (open) data and statistics about populations, economic indicators, education, etc. **Open data** is becoming mature as it is deployed widely by public sector organisations as well as NGOs. This applies to administrative data and, increasingly, to survey data and other types of data. From the set of 58 initiatives in the inventory, half of them the datasets among the initiatives examined were open or semi-open, with some kind of registration needed.
- The second main type of data is from **social media, sensors and mobile phones**. These data sources are relatively new, at least from the perspective of policymaking, and they have stimulated development of methods such as sentiment analysis, location mapping or advanced social network analysis. Data linking is applied in over two thirds of the initiatives in the inventory (including examples of linking four or five datasets).
- **Privately held data**, provided by market researchers, data brokers, data analytics, etc., is also used albeit to a lesser extent. More in general, the study presents a nuanced picture with respect to 'data as the new oil.' In many cases, data will be very relevant for a company itself, while additional steps are needed before the data are of added value in the context of public policy.
- The main observation with respect to **data analytics and visualisation** is that, although new analytics tools in the area of social media analysis and network analysis are being implemented, the emphasis still is on descriptive statistics rather than predictive analytics, profiling, sentiment mining, dynamic visualization, etc.

- A quick scan of relevant **European research projects** illustrate the opportunities for using a broader range of data sources and data analytics, across all stages of the policy cycle. As such, there is a normal 'gap' or time lag between advancements in research and implementation in real-life settings.
- **Privacy and inclusion** emerged as key topics during the literature review and the interviews, but not in the inventory. To some extent, this can be explained by the types of data used, e.g. statistical data instead of social media data or webscraping. Changing in European Union data protection and privacy legislations are closely monitored by experts and practitioners.

Based on valuable input of experts, while writing the report and while discussing the draft report, the following outlook can be presented:

- Currently we are moving from the height of the hype cycle in terms of big data generally and for policy making specifically, to the stage that poses **a risk of underestimating its potential**. An array of data-driven approaches is being developed and piloted. Leading uses of data-driven approaches have been developed in academic research and in the private sector, with the public sector so far mainly focusing on descriptive statistics. NGOs have been at the forefront particularly when it comes to transparency and pulling together openly available datasets.
- In terms of the sophistication of the analysis and the pace of implementation, **policymakers and public agencies have a number of particular challenges**. It may be easier to predict, for example, sales of books rather than policy interventions since the former have far fewer variables. Or again, it may be possible to use a crowdsourcing approach to filling potholes, but more difficult to use data to improve policy for a global problem like climate change. In policymaking, moreover, it is essential to ensure transparency, accuracy, inclusion and accountability. Moreover, prices of books can be changed on a daily basis, while public policy often benefits from continuity and predictability, even when real-time data may trigger 'short-termism' in policy design and implementation. These issues may apply to a lesser extent, or differently, for the private sector, academic research or NGOs.
- **A broad range of skills is needed** and there are concerns about the availability of these skills in public sector organisations. This refers to, for example, skills related to indicator design and data collection; technical and programing skills; and interpretation of (visual) data. One of the questions is whether all types of skills, to some extent, have to be available within public sector organisations, to be able to collaborate with outside experts, procure the right tools, etc.
- Data-driven approaches (e.g. using open data and involving citizen science and citizen engagement) can increase **transparency** of the policy process but at the same time, there is a risk that data-driven approaches reduce transparency. Data collection and data analytics (e.g. algorithms, machine learning) may not be fully understood and explained by policymakers and other stakeholders. To some extent, this is linked to the issue of skills.
- As the variety of data sources is increasing, policymakers should protect the balance between **collecting the data that is most crucial**, given a policy issue, relevant factors and indications, and using data that is readily available or than can be collected at low costs.
- This specific balance and, more in general, the selection of data sources and tools is influenced by stakeholders. Big data doesn't reduce the incentives for stakeholders to present **policy-based evidence** rather than contribute to evidence-based policy. In short: data is not neutral.
- The increased possibilities to work with large datasets (and to link datasets) underlines the importance of **international collaboration on data**

standardisation/harmonisation related to skills, transport, import/export, company data, R&D, air quality, etc.

- There could be a 'dual learning opportunity' in combining **policy experiments** (e.g. in specific countries or for specific target groups) and data-driven approaches for impact assessment. Does the policy intervention work? Does the data-driven approach provide timely and granular data that allows for understanding the role of various variables including the policy intervention?
- **The time lag between advances in research and implementation by policy makers and public agencies**, in a dynamic and broad field such as data-driven approaches for policymaking, calls for attention. Among the approaches to be intensified are 'match making' between developers and users of new tools; collaboration between researchers and policymakers in the development and piloting of tools (while also involving companies, NGOs, statistical offices, etc., as appropriate); and knowledge exchange among policymakers and public organisations that are working in different policy areas.

Appendix A List of interviewees

In identifying thought leaders, we took into account scientific excellence (e.g. key publications), geographic coverage (interviewees from Asia, Canada, EU Member States, Korea New Zealand and US) and a diversity of roles (interviewees that are leading academics, advisors of governments and/or practitioners).

The list of interviewed thought leaders is as follows:

Thought leader	Affiliation	Website
Prof. Viktor Mayer-Schönberger	Oxford Internet Institute, UK	http://www.oii.ox.ac.uk/people/?id=174
Prof. Miriam Lips	Victoria University of Wellington, New Zealand	http://www.victoria.ac.nz/sog/about/staff/miriam-lips
Emma Prest	DataKind UK	http://www.datakind.org/howitworks/datachapters/datakind-uk/
Dr. Andrew Hudson-Smith	University College London, UK	http://www.ucl.ac.uk/bartlett/people/?school=casa&upi=APSMI18
Prof. Dirk Helbing	ETH Zürich, Switzerland	https://www1.ethz.ch/soms/people/dhelbing
Prof. Alex 'Sandy' Pentland	MIT, US and World Economic Forum	https://www.media.mit.edu/people/sandy
Prof. Han Woo Park	YeungNam University, Korea	http://www.hanpark.net/index2.asp
Dr. Amanda Clarke	Carleton University, Canada	https://carleton.ca/sppa/people/clarke-amanda/
Prof. Philippe Laredo	University Paris-Est, France and University of Manchester, UK	http://latts.fr/philippe-laredo
Prof. Dennis Broeders	Erasmus University Rotterdam and Netherlands Scientific Council for Government Policy	http://www.eur.nl/fsw/english/sociology/profiles/profiel_mis/10532/

Appendix B Stakeholder analysis

In this appendix the relevant types of stakeholders in the field of (big) data-driven policy are introduced. The emphasis is on the *types* of stakeholders. To some extent, examples of individual organisations are provided.

The following types of stakeholders are introduced:

- Global and European policy makers
- National policy makers
- Regional policy makers
- Statistical offices
- Science and R&D organisations
- Data brokers
- Private providers of data analytics and visualisation tools
- Civil society
- Policy analysis/evaluation community

7.1 Global and European policy makers

7.1.1 Examples of individual actors

- The **European Commission** is the executive body of the European Union. It has increased in recent years its attention to better use of open data and Big Data in its policies.
- The **United Nations** (UN) is an international organisation of 51 countries committed to "maintaining international peace and security, developing friendly relaxations among nations and promoting social progress, better living standards and human rights".[51] The UN is involved via several departments and committees in (Big) Data for policy, mainly for statistics (see section 7.4) and climate change.
- The **OECD** is an international economic organisation of 24 countries with the aim to stimulate economic progress and world trade. The OECD views data-driven innovation as a key pillar for (economic) growth. It is undertaking extensive analysis on the role of data in promoting innovation, growth and well-being.[55]
- **Europol**[52] (EU) and **Interpol**[53] are international law enforcement agencies. They both have expertise in cyber related crime and are concerned with big data.

7.1.2 Associations, communities and platforms

- The European Commission has set up a partnership with other public and private partners to strengthen the data value chain (**Data PPP**) via research, development

[51] United Nations. (n.d.). *UN at a glance*. United Nations: http://www.un.org/en/aboutun/index.shtml

[52] See: https://www.europol.europa.eu

[53] See: http://www.interpol.int

and innovation.[54] This should result in a strong position of Europe in Big Data on the global market.

- Most countries selected in this study are members of the Organisation for Economic Co-operation and Development (**OECD**). In recent years the OECD has performed several studies on Big data and open data. For example, the OECD has undertaken the project "New Sources of Growth: Knowledge-Based Capital". In this project they have researched ways to improve the evidence base on the role of data for promoting growth and well-being, and have provided policy guidance on how to maximize the benefits of a data-driven economy.[55]
- The United Nations have created an Independent Expert Advisory Group on Data Revolution for Sustainable Development (**Data Revolution Group**).[56] The group consists of 24 members from public and private organisations from all over the globe. The group is concerned with the assessment of "new opportunities linked to innovation, technical progress and the surge of new public and private data providers to support and complement conventional statistical systems and strengthen accountability at the global, regional and national levels".[57] Big data is part of their scope.

7.1.3 Pilots and real-life examples of data analytics and visualisation

The European Commission outlined in 2014 its new strategy on big data, which focuses on research on innovation on big data. It has led to several project themes connected to EU policy objectives, among which are:[58]

- Healthcare: saving lives with better diagnostics. A theme in which, for example, an ICT tool for brain trauma patients was developed (TBICARE) and a platform to promote a healthier daily life (DAPHNE).
- Transport: fewer accidents and traffic jams. A theme in which a project, for example, focussed on fighting road congestion (VIAJEAO).
- Environment: reducing energy consumption. A theme in which, for example, a project focussed on the best place for wind farms (Sopcawind).

The World Bank Group launched a new internal programme in 2014 to spread the use of big data and analytics in development. The goal of this 'Innovations in Big Data & Analytics for Development' programme is to identify, incentivize and support emerging work on big data. A first operationalization of this programme was the WBG Big Data Innovation Challenge. From 130 proposals, 14 winners received 1.5 million dollars in seed and growth funding in total.[59] A lot of initiatives focus on the innovative use of call data records (CDRs) or geo-information (e.g. from satellites) in various policy areas such as agricultural policies (climate change data), labour shocks due to migration, transport policy, energy policy (e.g. monitoring rural electrification from space), education policy (predicting student's achievements), tax policy etc. The World Bank recently issued a

---

[54] European Commission. (2014, December 9). *Data Public-Private Partnership* . Digital Agends for Europe: http://ec.europa.eu/digital-agenda/en/data-public-private-partnership

[55] OECD. (n.d.). *Data-driven innovation for growth and well-being*. OECD - Better policies for better lives: http://www.oecd.org/sti/ieconomy/data-driven-innovation.htm

[56] http://www.undatarevolution.org/

[57] UN Data Revolution Group. (n.d.). *About the Independent Expert Advisory Group*. Data Revolution Group: http://www.undatarevolution.org/about-ieag/

[58] European Commission. (2014, October 29). *What can big data do for you?* Digital Agenda for Europe: http://ec.europa.eu/digital-agenda/en/what-big-data-can-do-you

[59] For a full list of winners: http://blogs.worldbank.org/voices/meet-winners-and-finalists-first-wbg-big-data-innovation-challenge

special report[60] on the opportunities of big data in various areas. Big data is considered especially useful for early warning systems (e.g. predicting riots using food price data), real-time information provision (e.g. commodity prices), and contributing to understanding complex social systems (e.g. favela growth using satellite images).

The United Nations has set up several studies and projects on big data for the public sector. They organised congresses and studied especially the possibilities of big data for statistics, the latter will be discussed in section 7.4. A selection of the other initiatives are:[61]

- The Data-Pop Alliance, a research centred initiative on big data for sustainable development, focusing on improving decisions and empowering people to prevent a new digital divide, dehumanisation and de-democratisation.
- The United Nations Global Pulse initiative is initiated to realise innovation in data use for better-informed policymaking and planning of development programmes. The focus is on public-private R&D to accelerate the development, discovery and use of big data for sustainable development and humanitarian action.

Furthermore, the European Commission and the United Nations have set up several projects on open data. Collecting and making available large sets of data, could facilitate big data analysis.

7.1.4 Societal challenges and/or policy areas addressed

Big Data is stimulated in Europe to promote jobs and economic growth, to promote industrial leadership and an open society (open data). It is connected to the many societal challenges that the European Commission has defined, among which are 'health, demographic change and wellbeing', 'smart, green and integrated transport' and 'climate action, environment, resource efficiency and raw materials'. However, no projects could be found in which the European Commission uses big data itself for direct use in its own policy cycle (apart from the activities of EUROSTAT, see section 7.4).

7.1.5 Relevant resources

Global and European policy makers potentially possess a lot of **data** that could be used for evidence-informed policymaking. Opening up these data could be a first step (open data). Furthermore, the organisations of these policy makers have significant **financial means** to set up projects and improve big data for policy. The global and European policy makers are able to design **legislation** for big data that could clarify for which purposes it may and may not be used. This would transfer some ethical aspects, such as privacy, from the ethical domain to the legislative domain.

7.1.6 Main benefits

Big data can help global and European policy makers to improve (speed and accuracy) their decision-making process, increase the relevance, effectiveness and efficiency of their interventions, and, in the end, better serve their citizens.

7.1.7 Main concerns

Internationally, there are different or unclear laws for privacy and the use of big data. There are also differences in ethics for the openness, gathering, use and purpose of big data. Furthermore, there is a difference in the geographical distribution of available data: more developed countries have generally much more data available than less

[60] World Bank. Big Data: In Action for Development. From: http://blogs.worldbank.org/voices/meet-winners-and-finalists-first-wbg-big-data-innovation-challenge

[61] UN Data Revolution Group. (n.d.). *Initiatives on the data revoluitin for post-2015.* Data Revolution Group: http://www.undatarevolution.org/catalog/2/

developed countries. Ownership and cross-border gathering of data (e.g. for intelligence) might also be an issue.

7.2 National policy makers

7.2.1 Examples of individual actors

National policy makers are part of **ministries or government departments**, **government agencies**, **policy related institutes** and other **official public organisations**. They either advise on policymaking or are directly responsible for policymaking and policy implementation. Several ministries and government agencies, such as intelligence agencies, are looking into the possibilities of big data and innovative data-driven approaches. Policy related institutes, such as advising bodies, are investigating the possibilities of big data for policy. Official organisations, such as statistical offices, are sometimes already using big data (see section 7.4). Also water boards, climate and highway agencies produce, collect and analyse a large amount of data. However, some of these operate on a regional level (see also section 7.3).

7.2.2 Associations, communities and platforms

The national organisations are working together in several international organisations. Some of these organisations are discussed in section 7.1.

- Within a European context the ministers cooperate in the **Council of the European Union**, they formally took the initiative to set up the Data PPP[62] (see paragraph 7.1.2).
- In the US the **Association of Public Data Users (APDU)** is a national network that links users, producers, and disseminators of government statistical data. APDU connects stakeholders with an interest and a concern about the collection, dissemination, preservation, and interpretation of public data.[63]

7.2.3 Pilots and real-life examples of data analytics and visualisation

- Intelligence agencies gather and analyse large amounts of communication and location data to have timely and accurate intelligence in order to protect national security. Recently, these kinds of big data projects have been revealed, most notably the PRISM project in the USA and the Tempora program in the UK. [64]
- The US Department of Agriculture has used big data analytics to reduce the rate of food stamp fraud by 60%. Using big data, it was able to quickly identify illegal traffickers of food stamps.
- An agency of the Dutch Ministry of Infrastructure and the environment, Rijkswaterstaat, together with a few other partners, have started a project (Digitale Delta) to use different data sources for preventing floods and improving water quality.[65]

---

[62] European Commission. (2014, December 9). *Data Public-Private Partnership* . Digital Agends for Europe: http://ec.europa.eu/digital-agenda/en/data-public-private-partnership

[63] Association of Public Data Users. (n.d.). Retrievied on April 29 2015 from APDU: http://apdu.org/about-apdu/

[64] Hobbs, A. (2014, July). Big Data, Crime and Security. *POSTnote* (470), pp. 1-5

[65] Rooney, B. (2013, June 25). *The Netherlands Looks to Big Data to Tackle Flooding*. The Wall Steet Journal: http://blogs.wsj.com/tech-europe/2013/06/25/the-netherlands-looks-to-big-data-to-tackle-flooding/; http://www.digitaledelta.nl/en/

7.2.4 Societal challenges and/or policy areas addressed

Big data can address many different national policy areas. Currently big data is often used in the areas of transport, where innovative sensor-data provides relevant information. Moreover, it is useful in detecting fraud, reducing crime and improving national security, both via defence and intelligence. Furthermore, big data (and especially open data) can increase transparency of government spending and procurement.

7.2.5 Relevant resources

National policy makers potentially possess a lot of **data** that could be used for informed policymaking using big data analyses. Opening up these data could be a first step (open data). Furthermore, the organisations of these policy makers have significant **financial means** to set up projects and improve big data for policy. The national policy makers are able to design **legislation** for big data that could clarify for which purposes it may and may not be used. This would transfer some ethical aspects, such as privacy, from the ethical domain to the legislative domain. In addition, legislation can also facilitate the use of data in various domains.

7.2.6 Main benefits

Big data can help national policy makers to improve (speed and accuracy) their decision-making, increase their overall efficiency and, in the end, better serve their citizens. Furthermore it can help them to reduce all sorts of crime and to protect national security.

7.2.7 Main concerns

Privacy and laws to regulate the use of big data is one of the main concerns, especially in the use of massively collecting internet and location data by intelligence agencies. Furthermore, although big data offers many new opportunities, it is often more unreliable and ambiguous as well. National policy makers are thus cautious in using it.

7.3 Regional policy makers

7.3.1 Examples of individual actors

Regional policy makers are first of all policy makers at **municipalities** (towns and cities), **provinces** and **states** (e.g. in Germany and the USA). However, regional policy makers also include policy makers at **water boards** (Waterschappen, Wasserverbände, Fresh water supply district, Agence de l'eau etc.) and **government agencies**, such as the **highways agencies** (Rijkswaterstaat, Highway Agency, Federal Highway Administration etc.) and the **local police departments**. Although the last two might be national agencies, they operate within the local policy domain, i.e. provinces or municipalities are responsible for the local infrastructure and the police.

7.3.2 Associations, communities and platforms

- Europe Regions Research and Innovation Network (**ERRIN**)[66] is a network of more than 90 European regions and their policy makers concerning topics about regional research and innovation. The network has some attention for big data, especially within the ICT working group.

[66] http://www.errin.eu

- **EUROCITIES**[67] is the network of more than 130 major European cities. The network has had some attention for big data in several events. Big data is addressed in the context of Smart Cities.
- **Associations of municipalities**, such as the Dutch Vereniging voor Nederlandse Gemeenten (VNG).[68] The VNG, for example, has a committee on information policy, which is also concerned with big data.

7.3.3 Pilots and real-life examples of data analytics and visualisation

- A first illustration of the many examples at regional and local is the US city of Dubuque (Iowa). Big data from mobile phones is used to track commuters and to improve transit services.[69]
- Miami-Dade County (USA, Florida) uses big data for water management (smart metering), intelligent policing, transparency and transportation efficiency.[69]
- The city of Louisville (USA, Kentucky) uses big data for public health of its citizens. The city has many inhabitants with respiratory disorders, which frequently results in peak episodes in the visit of emergency rooms. Inhalers with GPS transponders provide big data (place and time) that is analysed in order to predict when and where such episodes occur. This is used to improve the care for these patients.[69]The Kent Police developed a predictive policing software package, PredPol, which is able to produce a map of hot spots with the highest probability of crime in the next 12 hours. The software analyses twice a day real-time recorded crime location data about burglary, street violence, theft from vehicles and anti-social behaviour in combination with historic location data from the past five years. In addition, there are several similar big data projects at other police departments.[64]

7.3.4 Societal challenges and/or policy areas addressed

At the regional level big data could address policy issues concerning traffic, road safety, critical infrastructure, waste management, safety and security and public health. In contrast to national policymaking, data for regional policy focuses more on the policy implementation instead of agenda setting.

7.3.5 Relevant resources

Regional and local governments and agencies have a lot of **data** collected on a variety of aspects. Combining these data could result in new and more complete insights for informed policymaking. Furthermore, they have **financial means** as well as flexibility to set up projects and improve big data for policy.

7.3.6 Main benefits

Here, the same applies as for national policy makers. Big data can help to improve the decision-making process and the effectiveness and efficient of policy interventions.[70] As mentioned in 1.3.4 data for regional policy focuses on policy implementation and the improvement of efficiency.

---

[67] http://www.eurocities.eu

[68] See: http://www.vng.nl

[69] IBM. (2014, February 18). *Big Data is a Big Deal for Local Government*. ICMA: http://icma.org/en/icma/knowledge_network/blogs/blogpost/2162/Big_Data_Is_a_Big_Deal_for_Local_Government

[70] Leahy, E. (2013, April 29). *State and Local Agencies expect data to double in next four years; show little adoption of Big Data*. MeriTalk - The Government IT Network: http://www.meritalk.com/pdfs/state-and-local-big-data/State_and_Local_Big_Data_Gap_Release.pdf

7.3.7 Main concerns

At the first place local policy makers are confronted with their concern that the outcomes and the use of big data projects for policy are uncertain and unpredictable. Furthermore there are issues with ethics, transparency, data ownership and data responsibility that policy makers need to address.[71] Privacy is one of the ethical issues, but also a legal issue: privacy laws need to be respected and could pose restrictions on the use of big data for policy means.

7.4 Statistical offices

7.4.1 Examples of individual actors

The **statistical offices** are **national** or **international** organisations that collect and statistically analyse data within a specific geographical region for policy (related) means. The countries within the scope of this report have one or more statistical offices. The European Union, the United Nations, the World Health Organisation and the European Free Trade Association have their own statistical offices, which concern international statistics. A non-exhaustive list of examples of statistical offices is given below.

National Statistical Offices:[72]

- Statistics Netherlands (CBS)
- Korea National Statistical Office
- Statistics Canada
- Central Statistics Office Ireland
- Office for National Statistics (ONS) (United Kingdom)
- National Statistical Institute (Belgium)
- Statistics Denmark
- Federal Statistics Office (Germany)
- FEDSTATS and STAT- USA (USA)[73]
- Statistik Austria
- National Institute of Statistics Italy (ISTAT)
- National Institute of Statistics and Economic Studies (INSEE) (France)

International statistical offices:

- United Nations Statistics Division
- EUROSTAT (EU)
- OECD statistics
- WHO Statistical Information System
- EFTA Statistical Office

7.4.2 Associations, communities and platforms

The individual statistical offices are collaborating or they are affiliated with several statistical associations or platforms. A few are listed below.

[71] Lewis, C. (2014, July 2). *Using 'Big Data'*. Manchester Policy Blogs: Science and Technology: http://blog.policy.manchester.ac.uk/sci-tech/2014/07/using-big-data/

[72] Fore a longer list, see http://unstats.un.org/unsd/methods/inter-natlinks/sd_natstat.asp

[73] The USA have also statistical offices for certain specific policy domains: Bureau of Labor Statistics, National Center for Health Statistics, Bureau of Transportation Statistics and National Agricultural Statistics Service.

- The High-Level Group (**HGL**) **for the Modernisation of Statistical Production and Services**[74] is part of the Statistics Division of the United Nations Economic Commission for Europe. The HGL promotes international standards-based modernisation of statistical production and services. Big data has their attention.
- International Statistical Institute (**ISI**)[75] is a scientific association that links national statistical societies and statistical offices. It promotes and sponsors several conferences in which big data is addressed.
- International Association for Official Statistics (**IAOS**)[76] is an international platform for users and producers of official statistics. Its goal is to promote the understanding and advancement of official statistics and to stimulate statistical services globally. Big data is also a topic within this organisation.

7.4.3 Pilots and real-life examples of data analytics and visualisation

The statistical offices and the related associations have performed several projects to explore the possibilities of big data. Learning is often a central goal in these projects.

- The UNECE Statistics High-Level Group for the Modernisation of Statistical Production and Services sponsors a series of international collaboration projects related to big data and government statistics. These projects have the goal to identify strategic and methodological issues, demonstrate feasibility for novel and current official statistics and to facilitate the sharing of knowledge, tools and methods for big data official statistics.[77]
- Canada Statistics performed a pilot on Smart Meter Data. The data is available from electricity consumption sensors and analysed for official statistics. Canada Statistics concluded that the use of big data can offer more accurate, relevant and coherent data than surveys.[78]
- Eurostat has experimented with determining the Consumer Price Index (CPI) from internet price data. They developed a tool to scrape prices from the internet to improve the CPI.[79]
- Eurostat has performed a project on generating tourism statistics from mobile phone time and positioning data in several European countries.[80]
- The National Institute of Statistics Italy (ISTAT) has analysed mobile phone times and positions to estimate the flow of people in a certain administrative district. In their analysis they identified the flow of residents, embedded users, daily users and

---

74 Vale, S. (2015, January 13). *High-Level Group for the Modernisation of Statistical Production and Services*. UNECE Statistics: http://www1.unece.org/stat/platform/display/hlgbas/High-Level+Group+for+the+Modernisation+of+Statistical+Production+and+Services

75 ISI. (n.d.). *About ISI*. The International Statistical Institute: http://www.isi-web.org/

76 IAOS. (n.d.). *About us*. International Ossociation for Official Statistics: http://isi.cbs.nl/iaos/

77 Willis-Núñez, F. (2015, January 13). *Big Data Projects*. UNECE Statistics: http://www1.unece.org/stat/platform/display/bigdata/Big+Data+Projects

78 Trépanier, J. (2014, September 6). *Canada (Statistics Canada) - Pilot on Smart Meter Data (electricity consumption)*. UNECE Statistics: http://www1.unece.org/stat/platform/pages/viewpage.action?pageId=102039600

79 Willis-Núñez, F. (2013, October 2). *Eurostat - Consumer Price Index from internet price data*. UNECE Statistics: http://www1.unece.org/stat/platform/display/bigdata/Eurostat+-+Consumer+Price+Index+from+internet+price+data

80 Karlberg, M. (2013, October 3). *Eurostat - Mobile positioning data for tourism statistics*. UNECE Statistics: http://www1.unece.org/stat/platform/display/bigdata/Eurostat+-+Mobile+positioning+data+for+tourism+statistics

visitors. For that they combined big data with administrative data and used a machine learning method.[81]

- The Statistical Office of the Republic of Slovenia has tried to enhance the statistics about the Slovenian population distribution regarding time and location and the mobility (migration) of its population by using mobile positioning data.[82]
- Statistics Finland used data from geographically distributed traffic sensors to enrich the modelling of commuting time to better estimate the driving speed during rush hours. This project was in collaboration with the Finnish Transport Agency.[83]
- Statistics Netherlands has conducted multiple experiments with big data collection and analysis. It has researched the possibilities of vehicle detection loops for transport data, social media sentiment analysis of consumer confidence data and mobile phone sensor-data for data about people's location.[84]

7.4.4 Societal challenges and/or policy areas addressed

The statistical offices use big data to acquire better official statistics for policy means. These may concern all sorts of policy areas. Societal challenges that could be addressed are, for example, energy efficiency, infrastructure, smart transport and demographic change. The pilots mentioned in the previous paragraph concern these challenges. Other challenges could be fraud detection and control of epidemics.

7.4.5 Relevant resources

The most relevant resources that the statistical offices have are **knowledge** and **skills** related to statistically analysing large sets of data. They may also have the needed **technological infrastructure** to store and process big data. They have **financial means** to acquire and analyse data for official policy. Still, they may have to expand the experience and IT knowledge and equipment needed for big data. Most pilots are performed in cooperation with external institutes.

7.4.6 Main benefits

The main benefits of big data for the statistical offices is improving the accuracy, timeliness and relevance of their statistics and reducing costs.[85] For example, using social media data and having access to data about offline and offline retail revenues is less expensive than large-scale surveys (re-using and matching data versus collecting data).

---

[81] Barcaroli, G. (2013, October 4). *taly (Istat) - Specific purpose geographic basins and population statistics using mobile phone tracking data*. UNECE Statistics: http://www1.unece.org/stat/platform/display/bigdata/Italy+%28Istat%29+-+Specific+purpose+geographic+basins+and+population+statistics+using+mobile+phone+tracking+data

[82] Kuzma, I. (2013, September 9). *Slovenia - Population statistics using mobile positioning data*. UNECE Statistics: http://www1.unece.org/stat/platform/display/bigdata/Slovenia+-+Population+statistics+using+mobile+positioning+data

[83] Pasi, P. (2015, January 19). *Statistics Finland - Traffic sensor data for commuting statistics*. UNECE Statistics: http://www1.unece.org/stat/platform/display/bigdata/Statistics+Finland+-+Traffic+sensor+data+for+commuting+statistics

[84] Daas. P, Van der Loo, M. (2013). Big Data (and official statistics). UNECE Statistics: http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.50/2013/Topic_4_Daas.pdf

[85] Landefeld, S. (2014). Uses of Big Data for Official Statistics: Privacy, Incentives, Statistical Challenges, and Other Issues. *International Conference on Big Data for Official Statistics* (pp. 1-20). Beijing: United Nations Statistics Division. http://unstats.un.org/unsd/trade/events/2014/beijing/Steve%20Landefeld%20-%20Uses%20of%20Big%20Data%20for%20official%20statistics.pdf

7.4.7 Main concerns

The main concerns of the statistical offices as a stakeholder in big data for policy means is that data is reliable, accurate and representative.[85] Statistical offices have more control of this when collecting the data using conventional methods such as large-scale surveys. In addition, policy makers and researchers expect to receive precise statistics with a non-changing and clear process for analysis and interpretation. Statistical offices not only have less control due to the characteristics of big data, they also often have less control over the data collection itself. Statistical offices have to rely on third (commercial) parties, such as supermarkets providing scanner-data or telecom services providing mobile location data. Furthermore privacy issues remain a concern in the collection of large data sets and the connection of several data sets. The data needs to be anonymised.

The role of the statistician might change with the introduction of big data. Some have suggested that statistical offices should play a role in auditing and certifying the accuracy of big data.[85]

7.5 Science and R&D organisations

7.5.1 Examples of individual actors

There are multiple ways in which research and innovation communities are a stakeholder. First of all the science community performs research concerning data-driven approaches in policymaking. Secondly, they increasingly use big data provided by the government (some through open data platforms) in their research. Thirdly, the science community educates students and policy makers on data science and data-driven policymaking. Examples of these actors are:

- Universities have faculties and research groups specialized in data science (for policy)
- Several data centres have emerged. These interdisciplinary data centres connect scientists from different faculties and universities. One of the activities of these data centres is the research concerning data-driven policymaking.
- Big Data for Regional Policy – ESRC Business and Local Government Data Research Centre (research institute).[86]
- Scientists and researchers at international organisation such as the EU, UN, OECD, World Bank, WHO, WTO etc.

There is also an increasing interest from policy makers in using data for science policy. These policy makers use data-driven approaches to monitor and assess government spending for science policy and the realised output.

- National governments typically have departments to promote research and innovation within and beyond universities.
- International organisations regarding science policy include UNESCO, the OECD Directorate for Science, Technology and Industry, national funding bodies, and of course the EU's various science and R&D related directorates (e.g. DG Connect).

7.5.2 Associations, communities and platforms

- The Business and Local Government Data Research Centre[86] is leading the way in advanced analytics to help companies, local authorities and academics explore data, enhance knowledge and empower our society. The centre is funded by the Economic

[86] See: http://www.blgdataresearchcentre.com

and Social Research Council and includes a consortium of the University of Essex, University of East Anglia and London School of Economics.

- The Urban Big Data Centre[87], funded by the UK Economic and Social Research Council, the UBDC is a facility that brings together the expertise of urban social scientists, data scientists and statisticians from the University of Glasgow and six partner universities. They support policy makers, businesses, third sector organisations and citizens to harness the potential of big data to develop solutions for environmentally sustainable, economically resilient and socially just cities.
- Platforms for measuring science output using data include various scientometrics sources (e.g. Alt-metrics, Web of Science, Scopus, Google Scholar) and organizations (e.g. Leiden's Centre for Science and Technology Studies[88] or the Shanghai Jiao Tong University rankings[89])
- The Global Systems Science (GSS)[90] is a cluster of European research projects. Global Systems Science develops know-how about global systems. Examples include the internet and the worldwide system of cities.

7.5.3 Pilots and real-life examples of data analytics and visualisation

The science community has published a vast amount of papers regarding data-driven approaches in policymaking. Moreover, in the last couple of years, it has established educational courses and study programs for data-driven policymaking. Some examples:

- There are several FP7 research projects concerning big data. For example, the ePolicy project of the Italian government. In this project a proof-of-concept for an open source decision support system on renewable energy was developed.
- The Leiden Centre of Data Science offers a master class for policy makers on big data and the Leiden Peace Informatics Lab offers a summer school in big data for peace making.[91]

There are a couple of notable initiatives regarding data-driven science policymaking:

- There have been initiatives to measure the inputs and outputs of science including patents by projects such as Star Metrics[92], which take the approach of combining various traditional and new data sources.
- The SciMaps project[93], led by Katy Borner, has been at the forefront of innovative visualisations relevant to science policy.

7.5.4 Societal challenges and/or policy areas addressed

The science community supports policy makers in all policy areas, on all governmental levels and in all steps of the policy cycles.

---

[87] See: http://ubdc.ac.uk/about/overview/vision-and-objectives/

[88] CWTS (Centre for Science and Technology Studies, http://www.cwts.nl/Home) is the home of the Leiden Rankings (http://www.leidenranking.com) which measure scientific impact and collaboration.

[89] Academic Ranking of World Universities (Shanghai Jiao Tong University, http://www.shanghairanking.com/).

[90] See: http://global-systems-science.eu

[91] Leiden Centre of Data Science: http://campusdenhaag.leiden.edu/centre4innovation/events/big-data-policymakers.html and Peace Informatics Lab : http://www.peaceinformaticslab.org

[92] See: https://www.starmetrics.nih.gov

[93] See: http://scimaps.org

Concerning science policy, the main policy questions have included how to promote an environment which protects intellectual property and supports the most effective organization of disciplines and teams and resources. Steering the large resources devoted to research into the most useful and beneficial channels can be of great benefit to society, and this area has been one where there is great sophistication in the analysis and much data available.

7.5.5 Relevant resources

The most relevant resources that the science community has are **knowledge** and **skills** related to statistically analysing large sets of data and using an evidence-based approach when researching the data-driven approaches. They often also have the needed **technological infrastructure** to store and process big data. They have **financial means** to conduct research. Moreover they have the possibility to connect multiple disciplines in their research (as the data centres demonstrate). Lastly, they possess or have access to a vast amount of large data sets (e.g. climate data, civil engineering data, social and behavioural data) and can thus more easily connect different data sets.

7.5.6 Main benefits

The main benefit of big data for the research community is the new possibilities it opens for research in all areas.

The main benefits of big data for science policy makers are the improved accuracy, timeliness and relevance of the data. Policy makers can more accurately assess the efficiency and effectiveness of various scientific organisations and disciplines.

7.5.7 Main concerns

Concerning science policy, a number of critics have argued that 'gaming' the ways that science is analysed can have perverse effects of rewarding areas of research that are better at promoting themselves than at doing research. Likewise, 'picking winners' is regarded by some as counterproductive. Finally new methods (alt-metrics) are still regarded as too unknown to be of value for some.

7.6 Data brokers

7.6.1 Examples of individual actors

Data brokers are companies that collect, combine and repackage all kinds of data into commercial databases that are rent or sold to customers.[94] Examples of data that are sold by data brokers are names, addresses, phone numbers, e-mail addresses, hobbies, credit card information, buying behaviour, clickstream data etc. Virtually all data that can be captured from the web can be combined, analysed and repackaged for customers. Moreover some data brokers collect sensor-data, such as retail scanner data of mobile phone locations.

There are three main groups of actors in this field: commercial actors who collect and sell this data as their main activities; commercial actors who produce the data as a by-product or mainly for internal use; and public actors. Well-known commercial data brokers are for example **Datalogix**, **Epsilon**, **Acxiom**, **Experian**, **eBureau** and **ID Analytics**.[95] Companies that produce data as a by-product or for internal use can include retailers, all kind of companies with a website/app and companies using sensors. Many **retailers** collect information on their customers via store loyalty programs and

[94] Kitchin, R. (2014). *The Data Revolution*. London: SAGE.

[95] See: http://www.datalogix.com; http://www.epsilon.com; http://www.acxiom.com; http://www.experian.com; http://www.ebureau.com; http://www.idanalytics.com

some of this information is sold. But also companies behind **websites, apps and social media**, such as Facebook, Google and Twitter, collect and sell information about their users to customers, mainly for marketing purposes. There are also companies creating and selling very specific information, such as spatial or geographical information (**GIS-data**). Public actors include traditional collect data as part of their public mission, but also sell it to third parties. For example the **Chambers of Commerce** of many countries sell information on companies (e.g. the Dutch KvK). The alternative strategy, of course, is open data.

7.6.2 Associations, communities and platforms

- The **Association of Independent Information Professionals (AIIP)**[96] is an association with members from many countries that are active in providing and analysing information to clients.
- There are associations for people working at data brokers, for example the Digital Analytics Association. This association connects these data scientists, offers access to research and professional development.[97]
- NEMODE[98] (New Economic Models in the Digital Economy) has analysed various industry efforts to use data analytics and the impact of these efforts on the economy

7.6.3 Pilots and real-life examples of data analytics and visualisation

- Google has collected and analysed information on the flu activity throughout the world, this information could be used by government agencies to combat flu epidemics. The information from **Google Flu Trends**[99] turned out to be faster and more accurate[100] than the traditional monitoring of flue activity (through GP visits). The main limitation of the Google Flu Trends project is that the research cannot be replicated ('the black box of Google Trends').[101]
- The official statistics office of the Netherlands, Statistics Netherlands, has multiple contracts and partnerships with data brokers. The agency receives scanner data from supermarkets, social media sentiment analysis from the data broker Coosto and mobile phone location information from a telecom provider. The statistical office uses this information to produce statistics on inflation, consumer confidence and people's whereabouts.[102]

7.6.4 Societal challenges and/or policy areas addressed

Data brokers could provide their data for all kinds of societal challenges and/or policy areas. An example is healthcare, in which Google Flue Trends is active. Data brokers often do not analyse or actually use the data; they often only provide it for the other actors.

---

[96] See: http://www.aiip.org

[97] See: http://www.digitalanalyticsassociation.org

[98] See: http://www.nemode.ac.uk

[99] Google. (n.d.). *How does this work?* Google.org Flue Trends: https://www.google.org/flutrends/intl/en_us/about/how.html

[100] Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature , 457*, 1012-1014.

[101] Lazer, D., Kennedy, R., King, G., and Vespignani, V. (2014). The Parable of Google Flu: Traps in Big Data Analysis. Science. 343, no. 14 March: 1203-1205.

[102] Daas. P. & Van der Loo, M. (2013). Big Data and official statistics http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.50/2013/Topic_4_Daas.pdf

7.6.5 Relevant resources

Data brokers have as their main resource **data sets** on specific groups or on societies as large. Furthermore, they have **knowledge** of and **skills** in data collection and analysis, for which they have dedicated **tools**. As most of the data is commercially traded, they have the **financial** means and incentives to invest in the improvement of data collection, storage and analysis.

7.6.6 Main benefits

The main benefit of big data for data brokers is the fact that it created a real market for all kinds of information that previously did not seem interesting. All information can be valuable, because by combining and analysing several data sources new valuable information can be extracted and sold. Ever more tools are available to handle large data sets and to find correlations within these data sets.

7.6.7 Main concerns

The main concerns of data brokers are privacy, transparency and security issues. Data brokers collect lots of information on citizens, who are not always aware of this (transparency). There is lots of debate on the control citizens have over their personal information (privacy). In Europe this has led for example to the "right to be forgotten". Stricter regulations may influence the business of data brokers. Furthermore data brokers need to store their information securely to protect their business.

Some data brokers have responded to the concerns about transparency and privacy. For example, Acxiom, one of the world's largest data brokers, launched a portal (aboutthedata.com) that shows consumers the data that the company has on them and lets tem make changes or opt out.[103]

Moreover, data analytics are typically expensive because the extent of data cleaning required is typically underestimated. Data may be kept proprietary, so validation can be an issue.

7.7 Private providers of data analytics and visualisation tools

7.7.1 Examples of individual actors

Some of the data brokers discussed above also provide data analytics and visualisations tools (e.g. companies such as Datasift, GNIP, Splunk and Google).

7.7.2 Associations, communities and platforms

- There are conferences devoted to data analytics and visualisation in both the private sector and in the academic world. These are often done in conjunction with larger events that attract related professionals (such as computer scientists, applications developers, etc.).

7.7.3 Pilots and real-life examples of data analytics and visualisation

- GDELT[104] ingests public news data from 1979-present and repackages it for analysis and visualisation.
- Splunk[105] generally works with individual firms to provide tools for analysing internal data, and visualizing the data for improved decision making.

---

[103] Singer, N. (September, 2014). Acxiom lets consumers see data it collects. http://www.nytimes.com/2013/09/05/technology/acxiom-lets-consumers-see-data-it-collects.html?_r=0

[104] See: http://gdeltproject.org

[105] See: http://www.splunk.com

- Canopy Labs[106] specializes in providing high-end data analytics and visualisation capacity to SMEs.
- Crimson Hexagon[107] specialize in analysing social data related to organizations.

7.7.4 Societal challenges and/or policy areas addressed

These tend to be concentrated in areas like risk management, credit, terrorism, and marketing. They also have to do with operational efficiency of organizations.

7.7.5 Relevant resources

The actors in the field have the skills and means to clean and analyse open data, social media and public datasets.

As many of these companies are relatively focused and agile, tool development tends to occur more quickly in terms of the ability to work with newly emerging sources of data.

The data cleaning and consolidation skills represented in these companies far exceeds those in most public bodies or SMEs.

7.7.6 Main benefits

Analyses based on novel data sources that are less costly than surveys and other means of gathering data, while also being larger and more granular.

By letting data analysts outsource the expertise needed to clean and combine data to these firms, the analysts and policy makers can devote their time to more high-level issues rather than the (often tedious and time-consuming) details of working with a diverse set of data sources

7.7.7 Main concerns

The data sets in this area are often made purposefully opaque to protect the providers from losing their source of revenue.

Many organizations, both public and private, are reported to be struggling to implement big data solutions that provide actual new business intelligence

7.8 Civil society

7.8.1 Examples of individual actors

Civil society actors are operating data in the field of transparency, accountability and anti - corruption, and promote the publishing of data in order to re-use it in different civil applications.

- Fair play in Slovakia[108] is an NGO which promotes accountability through the use of Open Data.
- Open Cohesion in Italy is a civil association which monitor EU funded policies in Italy by submitting reports and DataKind UK is a chapter of the global DataKind which uses volunteers to analyse problems and solve them by using data.[109]
- Others, like the Sunlight Foundation[110] in the US are promoting policies of open data and use big data in order to promote these policies.

---

[106] See: https://canopylabs.com

[107] See: http://www.crimsonhexagon.com

[108] See: http://www.fair-play.sk

[109] See: http://www.opencoesione.gov.it and http://www.datakind.org

- 'Civic tech' is a movement which creates applications in order to improve or solve social problems by creating tailored open source web or mobile applications. One of the leading organizations in the field is Code for America, that sends data analysts, graphic designers and coders on fellowships into government to improve public service. It has been replicated in Code for Germany, Code for Australia, and other regions including Africa.[111]

7.8.2 Associations, communities and platforms

The Open Government Partnership (OGP)[112] is a international governmental body which has a civil society forum which allows to promote and transfer knowledge regarding Open Data and data analysis. However, there are only 40 countries which are associated in the program. Open Knowledge[113], a NGO which promotes open data and data literacy and use in decision making has a global network of civil society organisations who consult one another on these topics.

There are regional associations and partnerships between civil society organisations in Africa (Code4Africa), Latin America (DAL) and the US, but there is no such partnership for European civil societies organisations.

7.8.3 Pilots and real-life examples of data analytics and visualisation

- Open Courts by Transparency international Slovakia is a real life example of use of big and open data to policies in judiciary.[114]
- DataKind UK is doing data analytics in many ways in order to promote social causes.[115]
- I Pay Bribe in India tries to show where bribes are paid in India in order to abolish the phenomenon.[116]

7.8.4 Societal challenges and/or policy areas addressed

Civil society will mostly look at Anti corruption and accountability, which include public spending, budgeting and judicial sector. Sometime it will include other focuses such as home affairs or crime.

In addition, NGOs and other associations try to promote the public good, and are mainly interested in influencing the media or researchers or government policymakers by bringing pressure to bear on them.

7.8.5 Relevant resources

The Sunlight Foundation is funded by charities, which gives the Foundation the necessary staffing to influence government in a way that they would not otherwise have.

Civil society has knowledge as a resource that government does not have, for example volunteers who are data scientists as in DataKind, training resources like schoolofdata.org where there are courses available (can train civil society where there may be a digital divide gap).

---

[110] See: http://sunlightfoundation.com

[111] See: http://codeforamerica.org/api/

[112] See: http://www.opengovpartnership.org

[113] See: www.okfn.org

[114] See: http://www.otvorenesudy.sk

[115] See: http://www.datakind.org/howitworks/datachapters/datakind-uk/

[116] See: http://www.ipaidabribe.com/

The new.opendatahandbook.org gives resources to inspire and teach other organizations about other open data implementation which can be used by civil society.

There are also various websites listing, for example, open government datasets and the like, such as the open knowledge foundation open government data index[117].

Data often comes from public sources, and freedom of information laws ensure access in many countries, but by no means all countries.

Datakind is also an example of where the organization has data where data professionals in government are too silo'd to understand what the problems in certain sectors outside their own. Put differently, government is not problem-solving driven enough as civil society in this area.

7.8.6 Main benefits

Holding the powerful to account plays an important role for promoting a better society, and brining data analysis to bear on opening institutions is regarded as useful and entailing few of the difficulties (or far less so) that beset other big data applications such as privacy.

7.8.7 Main concerns

The quality of the data and the way it is organized for making use of it will vary, and so there is a question mark over how reliably and accessibly it can be used. Perhaps the main question mark is over how widely it will be used, apart from committed activists and those in the media and academia, even if the purpose of these data is open up data to as a large a population as possible.

7.9 Policy analysis/evaluation community

7.9.1 Examples of individual actors

The policy analysis and evaluation community consists of actors in both the private and the public sector. **Policy analysers and evaluators** are active **at government departments and semi-government organisations** in several stages of the policy cycle. In the private sector **consultancy firms focus on independent policy analysis and evaluation**.

7.9.2 Associations, communities and platforms

- **BetterEvaluation** is an international collaboration to improve evaluation practice and theory. The members share and generate information about methods, processes and approaches. Within BetterEvaluations there are several discussions and articles about big data.[118]
- The goal of the European Evaluation Society (**EES**) is "to stimulate, guide and promote the theory, practice and utilization of evaluation in Europe and beyond". Within the EES there is some attention for the use of big data in evaluation.[119]
- The American Evaluation Association (**AEA**) is an international association of professionals devoted to the exploration and application of evaluation.[120]

---

[117] See: http://index.okfn.org

[118] BetterEvaluation. (n.d.). BetterEvaluation: http://betterevaluation.org

[119] EES. (n.d.). *About EES*. European Evaluation Society: http://www.europeanevaluation.org/about-ees

[120] AEA. (n.d.). *About AEA*. American Evaluation Association: http://www.eval.org/p/cm/ld/fid=4

- **TAFTIE**, The European Network of Innovation Agencies, is a group of 29 European organisations that closely collaborate in the field of implementation of national technology programmes.[121]

7.9.3 Pilots and real-life examples of data analytics and visualisation

- In the **Nairobi Project** data from mobile communications devices, such as cell phones, are used to identify the needs of the poor in order to develop and monitor programs and policies that address these needs. The evaluation of policies is the main objective of this project.[122]

7.9.4 Societal challenges and/or policy areas addressed

Big data used by the policy analysis community could cover all kinds of policy areas. Big data fulfils the political need to improve the analysis and the evaluation of policies. This is ultimately connected to the need for good policies and the wish for accountability and openness of the public sector.

7.9.5 Relevant resources

The policy analysis and evaluation community has **knowledge** of and **experience** in selecting indicators, collecting or processing data, analysing data and communicating the results to policymakers.

7.9.6 Main benefits

The main benefits of big data for the policy analysis and evaluation community is the fact that big data allows to reach respondent groups that normally are hard to reach. Also aspects that are generally hard to measure by means of surveys could be measured using big data.[123] Moreover, using techniques such as data and text mining could reduce the cost of projects and the burden on interviewees (cf. survey fatigue). It could thus be an additional asset to policy analysis and evaluation.

7.9.7 Main concerns

Big data allows to acquire and to analyse large amounts of data in a fairly short time, which could be a threat to the services of the analysis and evaluation community. Big data may capture the characteristics of a whole group and algorithms can do the analysis quickly. Traditional evaluation, with sample groups and with consultants who collect and analyse the data, could become more or less superfluous.[124] Of course, this community could also use big data themselves, but then new skills might be needed. Concerns to be dealt with when using big data for policy analysis of evaluation are privacy and confidentiality, inclusion (e.g. when using social media data), unknown population representation, data quality, diversity/representativeness, accuracy and uncertainty.[125]

---

[121] Taftie. (n.d.). *About*. Taftie: http://www.taftie.org/about

[122] Schintler, L. A., & Kulkarni, R. (2014). Big Data for Policy Analysis: The Good, The Bad and The Ugly. *Review of Policy Research , 31* (4), 343-348

[123] Friedman, A. (2013, November 6). *Increasing efficiency in government programming with Big Data*. Nonlinear Digital: http://www.nonlinearcreations.com.br/Digital/how-we-think/articles/2013/11/3-considerations-for-increasing-efficiency-in-government.aspx

[124] Lu, S. (2013, April 30). *Can Big Data be Friends with Evaluation?* FSG: http://www.fsg.org/KnowledgeExchange/Blogs/StrategicEvaluation/PostID/440.aspx

[125] Schenker, N. (2013, September 19). Big Data: Uses and Limitations. http://www.cdc.gov/nchs/data/bsc/bscpres_schenker_091913.pdf

Appendix C List and summaries of inventory initiatives

Inventory initiatives

Name of initiative & website	Organisation	Summary[126]
Riigiraha http://riigiraha.fin.ee	Ministry of Economic Affairs and Communications (MKM)	The web portal provides citizens and policy makers with Estonian local governments accounting data and analytics tools. The purpose of the portal is to make the government financial activities more transparent and understandable. Users have the opportunity to create their own statistical overviews and to analyse the data on a detailed level. The Ministry of Finance is able to offer such a portal, as it collects all local governments accounting data without any interference. The submitted data covers the period from 2004 to 2014. Furthermore, the Ministry added further statistics for each local government that it collects autonomously. The data comes from statistical offices and administrative data. The main data analysis methodologies used are descriptive statistics and benchmarking. This initiative is unique in its kind as it combines financial accounting data from municipalities with traditional statistics. The website allows users to combine different data with visualisation tools. This gives users a detailed insight in the financial activities of Estonia.
DataKindUK http://www.datakind.org/howitworks/datachapters/datakind-uk/	DataKind UK	The UK data-for-good community is poised to make a big impact, with enthusiastic data scientists ready to give, a government deeply invested in using data to drive innovation and a vibrant charitable sector ready to embrace data to advance its work. DataKind™ harnesses the power of data science in the service of humanity. They engage data science and social sector experts on projects addressing critical humanitarian problems and lead the conversation about how data science can be applied to solve the world's biggest challenges. The data sources used are statistical offices, small-scale survey data, administrative data and social media. The main data analysis methodologies include agent-based modelling, descriptive statistics, predictive analytics and trend analysis.
Open Coesione http://www.opencoesione.gov.it	Department for Development and Economic Cohesion (DPS)	OpenCoesione is Italy's first national web portal on the implementation of investments programmed in the 2007-2013 programming cycle by Regions and State Central Administrations via cohesion policy resources. Publication of data from now on allows Italian citizens to evaluate if and how both implementation projects meet their needs and whether financial resources are allocated effectively. The cohesion policy promotes the economic, territorial and social cohesion, reduces the disparities in terms of development among regions and makes the opportunities of citizens even. In Italy, the Cohesion Policy is financed by additional resources from the EU (Structural Funds with the principle of additionality) and the national budget (cofunding to the EU Structural Funds and national funds). The web portal contains information about any single project carried out to implement cohesion policy, and more specifically: funds used, places and categories, subjects involved and implementation timeframes. Users can either download raw data or surf through interactive diagrams

[126] Summaries based on website of initiatives or interviews with contact persons

Name of initiative & website	Organisation	Summary[126]
		itemised by expenditure categories, places and type of intervention, as well as have access to files on single projects and subjects involved. Data on the local economy and social context are provided as well. The lack of transparency on how public money is spent is one of the main reasons for the slow pace in implementing development policies and in understanding whether investment projects actually respond to local demand. This is a particularly hot topic in Italy at present given the low absorption rate of EU Structural Funds and the debate all over Europe continues on the extent to which, after decades of subsidies, whether the European Regional Policy is effective or not. The OpenCoesione web portal provides analysis and monitoring on the use of cohesion policy resources, offering information, accessible to anyone, on what is funded, who is involved and where. The main data source is administrative data and the main analysis methodology is descriptive statistics. The analysis is done using tools such as excel and statistical programs. This initiative uses innovative visualization techniques. The website does not merely provide data, but visualises spending in different sectors in different areas (e.g. by using maps).
SUPERVIZOR http://supervizor.kpk-rs.si	Commission for the Prevention of Corruption	The present Web services to the general public, media, art and public authorities provides insight into the transaction of public institutions and state-owned companies and municipalities, relating to goods and services, wages, social benefits, pensions, grants, scholarships, etc. Public flow of money between the public and private enhances accountability of public office holders for the efficient and effective use of public funds, provides a reasoned discussion of the measures taken and planned investment and reduce the risk of mismanagement, abuse of authority, especially limiting systemic corruption, unfair competition and clientelism. Commission for Prevention of Corruption with this project opens up a new paradigm of transparency of government operations and limiting corruption risks. The fundamental mission of the KPK is: strengthening the rule of law, integrity and transparency, elimination of corruption risks and conflicts of interest. The main data source is administrative data and the main analysis methodology is descriptive statistics. The analysis is done using tools such as excel and statistical programs. The main data source is administrative data and the main analysis methodology is descriptive statistics. The analysis is done using tools such as excel and statistical programs.
REGIA http://www.regia.lt	Centre of Registers	Three years ago the Centre of Registers has started to develop a regional geo-information service called REGIA, which is specifically developed for municipalities, i.e. their citizens, officials and therein-operating businesses. REGIA facilitates the decision making process based on geographic location as well as information exchange. REGIA is based on the Lithuanian cadastral map maintained by the Centre of Registers including the data of three basic state registers integrated therein: The Address Register, the Register of Legal Entities and the Real Property Register. Users of REGIA have a possibility to create and manage their own data layers within the service environment. In their own data layers users can collect and manage their data, information or documents, as well as to create and provide services based on geo-referential data. A user decides individually whether their data layers are displayed publicly and whether or not other potential users can use it. The data mainly comes from administrative data. The main data analysis methodologies used include descriptive

Name of initiative & website	Organisation	Summary[126]
		statistics and benchmarking.
Using big data for the evaluation of R&D grants in the ICT sector	Secretary of State for Telecommunications and for the Information Society, Spain	Started in January 2013, the project has two main objectives: (1) Improve knowledge of the ICT sector and (2) Design and implement more effective public policies, especially in the promotion of R&D&I. The main reasons that require an improvement of the knowledge of the ICT sector are the following ones: • The ICT sector is a very dynamic and highly specialized sector • It is also very horizontal, as it reflects all productive and social areas • There are important gaps in current knowledge of the ICT sector Furthermore, knowledge about the structure and size of the ICT sector are important for the procedure of the awarding of ICT R&D grants. The grant award procedure necessitates the evaluation of a large volume and variety of technologies and products. In addition, the evaluation of projects admissible for grants takes place in a short period of time. Therefore, the Big Data Project shall support the understanding of structure and size of the ICT sector for statistical purposes, but also in order to improve efficiency and effectiveness of the evaluation process for ICT R&D grants. As this Project entered into the design phase, it became clear the variety of possible sources of information that could be useful. At this moment, these are the different sources planned to be used at the project: Summary of the R&D&I projects submitted to calls, public procurement, patents and trademarks, papers & academic publications, websites of ICT companies, blogs/news sites in the ICT sector, job portals. In this project a wide range of new technologies and techniques have been tested and developed: Natural Language Processing (NLP), Machine Learning, Latent Semantic Ontology of ICT Sector, Topic Analysis, etc. The main objective of the Big Data project is to develop a system capable of performing the following functions: 1. Duplicate Detection (intracorpus): Scans all the elements of a corpus in order to detect similarities between various elements. 2. Identification of related information: Given an element (Grant Application), searches for other elements of the corpus or other sources related with (e.g. patents with similar subject, other grants, etc.). 3. Textual search: Allows complex text searches on all the corpus and sort the results according to different parameters or document metadata. 4. Conceptual Tagging: Scans an item and classifies it in one or more ICT categories, assigning labels and grouping those elements that deal with the same knowledge area. 5. Global Analysis and Technology surveillance: Analyses corpus and provides information on the concepts and ICT topics. Shows the hierarchical relationships between those concepts or topics and their evolution.

Name of initiative & website	Organisation	Summary[126]
simTD - Secure intelligent mobility http://www.simtd.de	• Federal Ministry for Economic Affairs and Energy (BMWi) • Federal Ministry of Education and Research (BMBF) • German Association of the Automotive Industry (VDA)	simTD is pursuing the following principle objectives: • Increased road safety and improved efficiency of the existing traffic system through the use of car-to-x communication • Definition and validation of a roll-out scenario for the identified functions and applications for scientific questions through practice-oriented experiments and field operational tests • Consolidation of car-to-x functions from the categories of traffic efficiency, driving and safety as well as value-added services • Definition, analysis, specification and documentation of those functions that are to be developed and tested, as well as of the resulting requirements for the overall system for selected functions and tests within sub-project 1 • Development of test and validation metrics and methods in each phase of the overall system development in order to allow measurement and evaluation of the results • Consolidation and harmonisation of requirements from the perspective of feasibility and performance as well as their compatibility of requirements within the sub-projects • Verification of functions and requirements within the context of individual milestones The main data sources used include sensor-based data and administrative data. The data is analysed using descriptive statistics
Price efficient, participatory measurement of quality attributes in bicycle traffic through smart phone applications	Technical University Dresden	Currently, the measurement of quality attributes in bicycle traffic, such as the quality of traffic flows, is very difficult and time consuming. The aim of this project therefore is the development of a cost-efficient approach to analyse the quality of large bicycle traffic networks through data generated by smart phone apps. The approach is innovative, as it relies for the first time on objective sensor data and involves a significant number of bicycle traffic participants. The pilot will be conducted in the city of Dresden. However, the project is funded by the German Federal Ministry of Transport and Digital Infrastructure with the aim of applying the research results in other cities in Germany. The planning and maintaining of communal bicycle traffic networks is one of the main targets of the German National Cycling Plan 2020 (NRVP). The systematic measurement of bicycle traffic demand is of key importance to that. However, measuring cycling is in general very costly and can not be undertaken by smaller, less affluent communities and cities. Therefore, the main goal of this project is to provide cities and communities with a tool to measure cycling traffic on the basis of big data analytics. The project uses sensor-based commercial data, generated by fitness apps and wristbands. The data is analysed using descriptive statistics.
Smart Data for Mobility (SD4M)	Research consortium	The mobility sector is characterised by a high volatility of demand, and scarce resources on the supply side. Transport capacity is a non-reusable good: If a seat on a plane or on a subway remains empty, it is not re-sellable on the same journey. Therefore, the precise forecast of demand is important to optimally plan for capacities. Another factor of the mobility sector is its high degree of interconnection (Complex Data Value Chains). Large urban areas are highly complex super-systems where single traffic systems (rail, road, air) are deeply interconnected and subject to external effects, such as large events. SD4m proposes an innovative approach to forecasting on the basis of interconnected data. Its overall goal is the development of an ecosystem for multi-modal smart mobility services on the basis of a big data analytics platform for intelligent data management in order to optimise capacity, planning and forecasts of mobility service

Name of initiative & website	Organisation	Summary[126]
		providers, allowing for individual and flexible price design and amelioration of services for travellers. The main data source is social media. Other data sources used are: administrative data, commercial/traded business/consumer data and sensor-based data. Data is analysed using different techniques: text analytics, sentiment mining, trend analysis and predictive analytics.
Tender Tracking (Közpénzkeresö) http://www.tendertracking.eu	Corruption Research Center Budapest	Transparency is the most powerful insurance that a modern state spends taxpayer money for the common good. To strengthen transparency in Hungary, the Corruption Research Center Budapest has built the Tender Tracking Beta, a database of public procurement contracts in Hungary graded by a corruption risk scorecard. Our goal is to make every contract, issuer and winner and their assigned CRI (Corruption Risk Index) scores available to the public with better accessibility than from official sources. The CRI is a composite index of elementary corruption risk indicators checking for the presence of red flags in the life of the contract. In total there are currently 72,535 public procurement contracts in the Tender Tracking Beta from 2005 to 2012. Future work will expand the database in both directions, including next-day inclusion for newly published contracts. The Corruption Research Center Budapest was created in November 2013 in response to the growing need for independent research on corruption and quality of government in Hungary. The central aim of the Center is to systematically explore the causes, characteristics, and consequences of low quality of government, corruption, and regulatory failure using an inter-disciplinary approach. In addition, the Center also aims to help citizens to hold governments accountable through the use of robust evidence. Our unique research approach combines qualitative and quantitative methods to analyse micro-level actor behaviour and generates novel hard data on the phenomena under scrutiny. The Tender Tracking Beta is built using administrative data and data collected from the Hungarian Public Procurement Office's notices of public contracts from 2005 to 2012, and, from 2009 to 2012, each contract's associated Corruption Risk Index score. The data are cleaned, formatted and structured by Corruption Research Center Budapest. Subsequently they are analysed using descriptive statistics.
Market Model	Hungarian Central Bank	The purpose of creating the banking industry simulation model is to support decision makers at the Hungarian Central Bank to plan and calculate their decisions based on precise data, statistics and simulation and to allow them to test various different scenarios. The simulation model is based on statistical data and survey data from various sources. It uses data from the Hungarian Central Statistical Office and market surveys as well as expert opinions. The simulation model is validated by real-time measurements. Validation and calibration are necessary to ensure authenticity and reliability. The simulation shows a probability space with potential options which plot all the possible outcomes of a decision. Different data analysis methodologies are used: descriptive statistics, sentiment mining, predictive analytics, profiling and agent-based modelling
Webcrawl	The Dutch Ministry of Interior and Kingdom Relations	The Dutch central government is transferring some of its tasks to municipalities, provinces and water authorities. This will give municipalities and provinces more say in issues that directly affect their residents, like youth care and public

Name of initiative & website	Organisation	Summary[126]
		transport, and permit a smaller central government. The formal decentralization was to be implemented starting 1 January 2015. In 2014 municipalities had time to adjust in order to take on these new tasks. A transition commission 'Social Domain' was established that would monitor the progress around the decentralization. In the summer of 2014 they asked the ministry of Interior and Kingdom Relations to do a quick scan: to what extend are municipalities ready for the decentralization? The ministry wanted to conduct this research with minimal administrative impact on the municipalities, who were already busy with the transition. Consequently, they decided to use open data and a webcrawl of policy documents to get an idea of the progress. About 25 indicators were established. Part of these indicators were based on statistics and administrative data such as financials and information on collaborations in the region. The other part of the indicators was based on a webcrawl of the official council documents of the municipalities. These documents reflect what councils in municipalities discuss in their meetings. The ministry established relevant combinations of words and the webcrawl searched through all the documents for the combinations of words. Municipalities was assumed to be further in the progress of decentralisation if more words combinations were found in the council documents. There main underlying assumption is that these municipalities have more attention and time for decentralisation, and thus are further along. The ministry recognized that the adverse relationship is also possible: councils that discuss decentralisation extensively have more trouble with the transition. However, the ministry sees this as decentralisation being a priority on the agenda and thus as a step in the right direction. After the data collection, the data was corrected for the difference in the extend that municipalities publish council documents. Some only publish the agenda while others publish detailed minutes of the meetings. The 25 indicators (including the results from the webcrawl) were analysed in Excel using descriptive statistics One of the main disadvantages was the delay of the council documents compared to the actual policy implementation. First some policy steps are informally taken before the council discusses them formally. Secondly, the council documents are published with a delay.
Austrian Register Census http://www.statistik.at/web_de/statistiken/bevoelkerung/volkszaehlungen_registerzaehlungen/	Statistik Austria	In June 2000 the Austrian federal government has decided that the census of 2001 was the last conventional census in Austria. From 2011 onwards a register census is used. This means that existing registers are cleverly combined in order to determine the measures that in a traditional census are surveyed. The data from 15 different registers (statistical and administrative data) are anonymously encrypted and statistically combined and analysed by Statistik Austria. Descriptive statistics is the goal and the result of the analysis. In 2006 the method was tested and proved to be reliable; the resulting data were of good quality. The method has reduced the costs for obtaining census data, is much faster than a traditional census, and has removed the burden for citizens to fill in lengthy surveys. The data are well protected and privacy is assured.

Name of initiative & website	Organisation	Summary[126]
		The method used can be classified as an early form of big data in the sense that several different sources are analysed and correlated, large data sets are used and N=all citizens is used.
Digital Delta http://www.digitaledelta.nl/en	• Ministry of Infrastructure & Environment • Local Water Authority Delfland • University of Delft • Applied Science Institute Deltares • IBM	The Digital Delta is aimed at the unlimited and readily accessible provision of water and climate services. Various actors in the water management field provide their data in a standardized form. The Digital Delta coordinates this and if possible, combines these data sources. It provides a catalogues where this data can be found. Not all data is and can be provided in a standardized form. The work is being carried out by the Digital Delta consortium, which is drawn from the private sector, public bodies and research institutions. The Digital Delta is intended to enable the various water managers to work more efficiently in carrying out their primary water management tasks. The Digital Delta can also contain and provide access to information for professional and non-professional users of water-related data. Every year the Netherlands spends some 7 billion euro on managing the water systems: a complex network of tidal inlets and dykes, canals and sluices, ditches and pumping stations, and sewerage and drinking water systems. The costs of water management will rise by around 1 to 2 billion euro by 2020. At the same time, data is collected, processed and analysed on a large scale in order to manage water effectively and efficiently and in order to be able to cope with the increasing fluctuations in the climate. This is leading to a massive growth in the amount of available data and required management applications, but also to fragmentation, incompatibility and a lack of transparency of solutions which restrict the effective use by the user for integrated water management. With these growing challenges in prospect, doing nothing is not an option. Within the Digital Delta research programme, Dutch parties are exploring how the integration and analysis of water data and a wide range of existing and new data sources can help to reduce the cost of future water projects by 20 to 30%. Different data source are used: administrative data, sensor-based data (water level) and static geo-information. No analysis is conduced by Digital Data; they only supply the data (in a uniform format if possible)
Consumer Price Index http://www.cbs.nl/nl-NL/menu/themas/prijzen/methoden/dataverzameling/korte-onderzoeksbeschrijvingen/2006-cpi-art.htm	Statistics Netherlands	Statistics Netherlands, the Dutch statistical agency (CBS), publishes the consumer price index statistic every month. This index indicates how prices change over time of products that consumers buy. For this statistic, the CBS collects data on prices of many different products. Part of the data collection is conducted using conventional data collection methods (surveys, visits to store) and part of the data is collected using scanner data from supermarkets. The main advantage of using scanner data over the conventional data collection methods is that it is faster, has more volume, is cheaper compared to on-site data collection and that it is less of a burden for supermarkets. The data is analysed with descriptive statistics; the CBS does not analyse the data further.
Mobile data for mobility http://www.cbs.nl/NR/rdonlyres/2A8D34FF-75D9-46BC-B2BB-53428F1699E3/0/IMnr0	Statistics Netherlands (CBS)	The Dutch Statistical Office (CBS) has agreed with the mobile provider Vodafone that it will receive data on the whereabouts of people that Vodafone gathers through their mobile services. The main goal is to gather information on where people are during the day, how they move about and how this varies during different days (holidays, weather etc.). The data

Name of initiative & website	Organisation	Summary[126]
9Projectmobieletelefonie.pdf		source is sensor-based data and CBS will publish descriptive statistics. This data is useful for policy makers in various domains. For example, it is useful for designing disaster contingency plans or for policy regarding commuter traffic. There is interest from various policy makers. However, the results of this project are not yet published. Thus, no policy makers use it yet. Once the results are published, the CBS will contact various policy makers and departments to discuss how this new data can be used.
Analysis of traffic sensor data for better commuting statistics http://www1.unece.org/stat/platform/display/bigdata/Statistics+Finland+-+Traffic+sensor+data+for+commuting+statistics	Statistics Finland	Statistics Finland uses currently data from geographically distributed traffic sensors to enrich the modelling of commuting times to better estimate the driving speed during rush hours. This project was in collaboration with the Finnish Transport Agency. The project is an experiment on the use of big data for official statistics in Finland. The raw data are accessed and processed in the "cloud". Besides the sensor-based data, statistical data from their Social Statistics Data Warehouse is used: workplace coordinates, demographics, dwelling coordinates, employment information etc. The data is analysed using statistical data.
Social media consumer sentiment analysis http://www.ecb.europa.eu/pub/pdf/scpsps/ecbsp5.pdf	Statistics Netherlands	Statistics Netherlands, the Dutch statistical agency (CBS), reports a monthly consumer confidence indicator: a statistic that indicates the confidence of consumers in the Dutch economy. This statistic is based on monthly survey data. The CBS has investigated whether this statistic can be constructed with sentiment mining of social media data. CBS asked Coosto, a company specialized in sentiment mining, to analyse social media and construct sentiment scores for consumer confidence. The CBS analysed the data and concluded that the sentiment scores correlated with the official consumer confidence statistic very well. Especially with social media messages from Facebook. The sentiment mining method has some advantages over the official statistic. It can be gathered more frequently and data can be analysed quicker. In contrast, the analysis of the survey data takes over a month. However, the interpretation of the sentiment data is more difficult. First of all because it is influenced by other events, such as Christmas or large sport events (national football etc.). Secondly, because not all Dutch people are active on social media. So the data does not represent the Dutch population. As the consumer confidence statistics is very important and cited often in the media, CBS has decided not to publish the sentiment data results as official statistics. However, they might publish it on an experimental website. Furthermore there is interest for this sort of data from other policy makers (than those concerned with consumer confidence). For example, policy makers are interested to know how Dutch people feel about safety in the Netherlands.
Vehicle detection loops	Statistics Netherlands	Statistics Netherlands, the Dutch statistics agency, reports a traffic intensity statistic every 3 months. This statistic is based upon a sample of points where the traffic is counted. Statistics Netherlands wants to update this statistic more frequently and report more detailed statistics using vehicle detection loops (sensor-based data). It has already done the research for this and will publish results in 2015 (descriptive statistics). The new statistics will be published monthly and will have

Name of initiative & website	Organisation	Summary[126]
		more details, e.g. the type of cars. This has implication for different policy makers. Policy concerning mobility/traffic will have better information to design policy. Moreover, it can be useful for (regional) economic monitoring, e.g. an increase in the number of trucks can indicate increased economic activity.
Automatic Number Plate Recognition (ANPR)	National Police of Denmark	After a pilot ended in 2013 the Danish National police has decided to implement an ANPR-system near border crossings and on (some) police vehicles. It is part of the implementation of the proposed new foreigners law in Denmark. It will be used to find criminals who cross the Danish border, stolen cars or cars registered for other reasons in the police's hot list and some other databases (e.g. Europol). The ANPR-sensors continuously scan the number plates of cars passing the sensor and check them with a central database. Thus sensor-based data is combined with administrative data. A lot of data is continuously recorded, analysed (descriptive statistics and profiling) and (partly) stored (for a limited time). The implementation of the ANPR system will be enrolled in may 2015 (planning).
Flanders Image Processing Chain https://www.agiv.be/producten/beeldverwerkingsketen	AGIV (Belgium agency for geo-information)	The Belgium agency for Geo-innovation collects a lot of raw remote sensing data, e.g. from satellites and aerial photography. The agency is working on products that automatically analyse this (big) sensor-based data, combine it with administrative sources (e.g. data on housing) and convert the data into maps/visualised products (descriptive statistics). Two main examples are: 1) Buildings mutation: AGIV uses aerial photography to automatically assess what building have disappeared, are new or have changed in an area. This data is combined with GRD-data (Large Scale Reference Database). This is interesting for policy regarding urban planning. 2) Vegetation map: AGIV uses remote sensor data (infrared sensors) to map vegetation in areas. This is interesting for policy regarding nature reservation. AGIV develops these products either on demand of clients (often the government) or in order to demonstrate their usefulness to policy makers.
OpenFisca http://www.openfisca.fr/en/	Team consisting of France Stratégie, IDEP and Etalab.	Free innovative software for open-source micro-simulation of the tax-benefit system in France. The government observed that far too many people in France are not requesting social benefits whilst being entitled. Either they do not know, or the system is too complicated (old system which grew more complex over the years). The initial objective of OpenFisca was to increase the awareness of citizens for social benefits and to let them calculate whether they are entitled and what effect this would have on their income. The data citizens fill in to calculate this is combined with statistical data. However, it enables the user also to simulate the effects of tax reforms. Users can calculate tax optimization for specific social groups (non-agent based modelling).
NOMAD http://nomad-project.eu/	European FP7-funded consortium coordinated by the University of the Aegean (Greece)	Collaboration and crowdsourcing are the realities of today's public internet. The so-called "Web 2.0" represents a precious repository of thematic information, thanks to the heterogeneous content that is inserted daily and spontaneously updated by its users. It is possible to use this insight to information at multiple stages of the policy-life cycle to support the definition of the political agenda, the creation, the implementation and the monitoring of policy proposals. In this context, modern politicians can test, detect and understand how citizens perceive their own political

Name of initiative & website	Organisation	Summary[126]
		agendas, and also stimulate the emergence of discussions and contributions on the informal web (e.g. forums, social networks, blogs, newsgroups and wikis), so as to gather useful feedback for immediate (re)action. In this way, politicians can create a stable feedback loop between information gathered on the Web and the definition of their political agendas based on this contribution. The ability to leverage the vast amount of user-generated content for supporting governments in their political decisions requires new ICT tools that are able to analyze and classify the opinions expressed on the informal Web, or stimulate responses, as well as to put data from sources as diverse as blogs, online opinion polls and government reports to an effective use. Different analysis methodologies are used: mainly sentiment mining and text analysis, but also descriptive statistics. To this end, NOMAD introduced these different new dimensions into the experience of policymaking by providing decision-makers with fully automated solutions for content search, selection, acquisition, categorisation and visualisation that work in a collaborative form in the policy making arena.
OpenExpo http://dati.openexpo2015.it	Expo 2015 SpA	The Universal Expo is a global expo and a national project of the government of Italy. Open Expo is an initiative of Expo 2015 which aims to ensure total transparency at the Universal Exhibition of 2015 by publishing in open formats of all information concerning the management, planning, organisation and the conduct of the event. OpenExpo2015.it intends to enrich the information already released by Expo 2015 SpA in the administration by making information available concerning the economic management of the event (income and expenditure, purchases, payments and their beneficiaries), the works (construction sites, description of the works, the amounts required for their implementation) and any variations in the performance of the event (number of visitors, mobility and public transport). The data is analysed using descriptive statistics. OpenExpo2015.it data is also available on the portal that has innovative dynamic info graphics, which are constantly updated. The info graphics are very detailed use advanced visualisation techniques. This allow anyone to have a clear idea, at a glance, of the state of progress of work and, more generally, the trend of the entire event." This platform was created by the management if the event after a big corruption story was exposed by civil society.
Moni-thon http://www.monithon.it	Monitoring Marathons	Moni-thon" comes from "monitor" and "marathon", and this is precisely what this platform seeks to help with: an intensive activity of observing and reporting of public policies in Italy. What's there to monitor? Monithon was born as an independently developed initiative to promote the citizen monitoring of development projects funded both by the Italian government and the EU through the Cohesion (aka. Regional) Policy. Projects include a wide range of interventions such as large transport, digital, research or environmental infrastructures (railroads, highways, broadband networks, waste management systems…), aids to enterprises to support innovation and competitiveness, and other funding for energy efficiency, social inclusion, education and training, occupation and workers mobility, tourism, etc. Citizen monitoring of these projects is possible thanks to a

Name of initiative & website	Organisation	Summary[126]
		combination of open government data and citizens' collaboration, joined by the goal of controlling how the projects are progressing, and whether they deliver actual results. Monithon means active involvement of communities and a shared methodology. Citizens, journalist, experts, researchers, students – or all combined – collect information on a specific project chosen from the OpenCoesione database. Then this information can be uploaded on the Monithon platform (based on Ushahidi) by selecting the projects from a list and it can be geo-referenced and enriched with interviews, quantitative data, pictures, videos. The result is a form of civic, bottom-down, collective data storytelling. All the "wannabe monithoners" can download this simple toolkit, a 10-page document that describes the initiative and explains how to pick a project to monitor and get things started. Users can do this individually, but the whole activity increases in its civic importance when done collectively, as it happens during the "Monithon Days" such as the National Monitoring Marathon of February 22, 2014 organized as part of the International Open Data Day. During these events groups of citizens - sometimes under the guide of local organizations – set out on real explorations around their area, to gather information on specific projects of local interest. In doing so, not only do the participants collect useful material to evaluate the effectiveness of the funding and generate awareness around these finance plans, but they also experiment with new forms of control over public policies. The data is not further analysed by Monithon as is only provides the qualitative projects reports.
The Department of Public Expenditure and Reform http://www.per.gov.ie	The Department of Public Expenditure and Reform (Ireland)	The Department of Public Expenditure and Reform's mission statement is: To serve the country, its people and the Government by delivering well-managed and well-targeted public spending, through modernised, effective and accountable public services. This Department was established in 2011 with twin goals: to reduce public spending to more sustainable levels and to reform and improve our public services The website of the Department aims to keep people fully up to date with all of its major initiatives. The design will help people explore their main policy areas, view their publications, view expenditure details, videos and photos, and press releases. Data from statistical offices and administrative data is used. (Descriptive) statistical models are used to analyze the data and find ways to improve the budget.
Risk Assessment and Horizon Scanning (RAHS) http://www.rahs.gov.sg	The RAHS Program Office	The Risk Assessment and Horizon Scanning (RAHS) programme was launched in 2004, as part of the National Security Coordination Secretariat (NSCS). The RAHS program explores methods and tools that complement scenario planning in anticipating strategic issues with significant possible impact on Singapore. With the help of international partners, as well as through experimentation, the RAHS Programme has today developed an extensive range of processes that enabled agencies to collect, analyse, inform, model and monitor emerging strategic issues. Armed with the processes and products, the RAHS team has been actively engaging government agencies, academic institutions and international partners through training, consultancy and joint projects. These projects generally

Name of initiative & website	Organisation	Summary[126]
		explored emerging issues with the aim of enhancing strategic anticipation capabilities for the agencies. Data sources used are statistical offices, large-scale survey data and sensor-based data. Sentiment mining, text analysis, narrative scanning and descriptive statistics are used to analyse the data.
Real Estate Market Analysis http://www.trgnepremicnin.si/en/vsebine-portala/stanje-trga	Geodetic institute of Slovenia	Based on our good knowledge about the contents and principles governing the functioning of the existing real estate and other spatial records in Slovenia, the Geodetic institute of Slovenia develops, in cooperation with external partners, own applications and information solutions that offer expert support in real estate valuation. The institute regularly monitors the situation at the Slovene real estate market. The price trends and real estate transaction movements in the area of Slovenia are regularly updated and comprehensively presented at the portal Trg nepremičnin (Real Estate Market), offering different analyses for the general and expert public. We offer our customers target-oriented Slovene real estate market analyses. At the portal Trg nepremičnin, a service is being prepared that will present beside the data about realised prices also the information on advertised prices, adequately analysed to establish gaps between them. The application TRGOSKOP presents the easiest and fastest way of monitoring events on the Slovene real estate market, based on data about concluded real estate sales/purchase agreements. The application uses data from Real estate market register, which the real estate agents, public notaries, administrative units and the Tax Administration of the Republic of Slovenia are obliged to provide. The data is analysed using descriptive statistics and predictive statistics to predict property value.
Public Spending http://publicspending.net	Independent initiative	Governments around the globe are opening up public spending data in order to promote transparency and citizen awareness. However, data openness by itself is not enough to guarantee that the data is consumed efficiently and in meaningful ways. In the publicspending.net project administrative data and commercial/traded data from seven governments, both local and national, with total value of almost 1,5 trillion euro, are processed, cleansed and converted to Linked Open Data, following best practices. In cases where Open Data exist, the basic obstacle is the fact that there are not even standards for representing the main actors (i.e. payers, payees) and the type of payments. Therefore, it is impossible to interlink the available data in meaningful ways and support evidenced-based policies. Publicspending.net addresses the aforementioned obstacles by employing efficient transformations and standards. The results of the project have been used in analysing the market and competition conditions and in forming evidence-based public policies especially in Greece. The data can be analysed with network analysis (spending networks) or profiling at the business level.
SlowWolf http://www.volkovi.si/	Volkovi	The LIFE+ SloWolf project, which started with planned activities on 1 January 2010 and will continue through the end of 2013, represents the first large-scale project on wolves in Slovenia. The project is carried out collaboratively by partners from the University of Ljubljana, the Slovenian Forest Service, and the Society for the Conservation, Research and Sustainable Development of the Dinaric Alps – Dinaricum. The applicant and leader of the project is the Animal Ecology Group at the Department of Biology of the

Name of initiative & website	Organisation	Summary[126]
		Biotechnical Faculty of the University of Ljubljana. The goal of the project is the long-term conservation of the wolf population, its main prey and habitats in Slovenia, and improvement of their coexistence with humans. Wolves are natural predators of free-ranging ungulates living in Slovenian forests. They live in packs and require a large territory to survive. In Slovenia, many important details about how they live, for example the number of animals in the average pack, the number of packs, and the size of their territories, are as yet unknown to us. The attitude of people towards wolves also varies greatly. Individual groups of people, for example hunters and livestock keepers, may come into conflict with wolves because of their predation on wildlife and sheep. At the same time wolves increasingly symbolize wild and unspoiled nature. The continued existence of wolves in Slovenia depends on a number of factors: these will be the subject of study by the project, and the findings and knowledge thereby obtained will form the basis for the successful conservation and management of this large carnivore species. The project area covers some of the best preserved forests in the country and in Europe. The Dinaric region is considered one of Europe's biodiversity hotspots. Together with the neighbouring region of Gorski Kotar in Croatia, this is the largest non-fragmented forest complex in central Europe." The main data sources are large-scale survey data and sensor-based data. The data is analysed using descriptive statistic and geo-location analysis.
Canadian international development portal (CIDP) http://cidpnsi.ca	The North - South Institute think tank at the Norman Paterson School of International Affairs, at Carleton University (Ottawa).	The goal of the Canadian International Development Platform is to leverage open data and big data, to enhance policy relevant analysis of international development issues, and to engage Canadians on development issues by grounding discussions in the best available evidence. The CIDP began as an initiative of The North-South Institute (NSI), Canada's only independent international development think-tank, and is now housed at the Norman Paterson School of International Affairs, at Carleton University (Ottawa). The CIDP builds on NSI's nearly four-decade track record of research on international development, and over 15 years experience compiling detailed data on Canada's engagement with developing countries. Canada's engagement with developing countries is multifaceted and complex. Being an important foreign aid donor and key member of the OECD Development Assistance Committee, aid forms an important component of what is covered in the CIDP. CIDP brings together a wide range of data and analysis on foreign aid, including highly detailed disaggregated data compiled from multiple sources and visualized for ease of use and rapid analysis. The main data sources are statistical offices, administrative data and commercial/traded business/consumer data. One of our guiding convictions in building the platform is that development is much bigger than foreign aid. To provide a more comprehensive picture of Canada's engagement with developing countries, CIDP covers a range of other development data including trade, investment and migration and remittance flows. Data is analysed using descriptive statistics and trend analysis (e.g. trend in finance)
Znasichdani.sk http://znasichdani.sk	Fair-Play Alliance and Company register	Znasichdani.sk aims to show the public which entrepreneurs and companies are successful in securing contracts with the state in public procurements. The government is bound to inform about handling the official commissions through the

Name of initiative & website	Organisation	Summary[126]
		Public Procurement Bulletin. According to this, information about the flow of taxpayers' money can be traced from state to the companies. But it is rarely possible to track the flow of money to specific individuals involved in these companies, since the direct link between procurements and specific persons has never been available. Such information is not provided in the Bulletin. Znasichdani.sk is based on the assumption that if people get access to this kind of detailed information, management of public money in Slovakia will become more transparent. Procurement of overpriced goods and services will get a new dimension if citizens are able to connect the benefits from these procurements with specific names and faces. Information about contracts with the state are obtained from the Public Procurement Bulletin, information about connections between entrepreneurs and companies come from the Business Register of the Slovak Republic. The service can find owners and other stakeholders of companies successful in public procurements, as well as the volume of money they gain through commissions from the state and details about these procurements. The project is a collaboration of the Fair-Play Alliance and the Company Register project (Firemny-register.sk). The project is funded by Open Society Institute Data from statistical offices and administrative data is used. Data is used using descriptive statistics.
Otvorene Sudy http://otvorenesudy.sk	Transparency International Slovakia	The judicial system has a major impact on the quality of life and democracy. Functioning and reliable judiciary is a necessity for good business but also for the ordinary citizens. The state of the judiciary system is not satisfying. According to the Eurobarometer of autumn 2009 only 29% of Slovaks trust the court. At the same time, each twelfth household has had contact with the judiciary paid bribe, according to a survey of the Global Corruption Barometer 2013. The activities of judges and the courts are not known yet, much because of limited access to information. At the same time the debate on the functioning and quality of justice suffered from two problems. Firstly, they are often paušalizujúce ("Judges are corrupt") without clearly naming the culprits or problems. Secondly - a number of discussions are kept within unnecessarily small group of experts from the legal profession. Project Open Courts is a contribution to the debate on improving Slovak judiciary in three ways: 1. Improving the availability of data on the activities and performance of courts and judges, 2. The data interconnects and allows comparison. 3. Selectively - the level of individual judges and the courts. The data used, regarding judiciary in Slovokia, stems from administrative sources. The analysis is conducted using benchmarking and descriptive statistics. For this they use a self-developed central database on Hithub. Project Open Courts was created with the support of the Secretariat of Transparency Internationalin Berlin and Project Restart organized by the Centre for Philanthropy . Updates website were sponsored by the Fund for Transparent Slovakia in the Pontis Foundation
Pol-On https://polon.nauka.gov.pl	OPI Informacji-National Research Institute	POL-on is an integrated system of information on higher education, which supports the work of the Ministry of Science and Higher Education, and the Central Statistical Office and the Central Commission for Academic Degrees and Titles. Its essential task is to create a global database of scientific institutions, universities and learning Polish. Using this

Name of initiative & website	Organisation	Summary[126]
		information, collected support the decision-making processes of the Ministry of Science and Higher Education with regard to Polish universities and research units. POL-on makes it easy to make decisions about targeting spending on education and material assistance to universities. A range of information collected in the system is made available to the public. The data collected stems from administrative data and is analysed using descriptive statistics (using self-developed tools). The project is being conducted as a partnership project system within the Human Capital Programme (priority IV, Action 4, sub-measure 4.1.3).
myGov.in https://mygov.in	National Informatics Centre, Department of Electronics & Information Technology, Ministry of Communications and IT, Government of India.	MyGov is an innovative platform to build a partnership between Citizens and Government with the help of technology for growth and development of India. Through this platform, the Government aims to encourage Citizen Participation towards Good Governance by seeking their ideas, suggestions and grass roots level contribution. Citizens can participate in this unique initiative of nation building. For the very first time in the history of this country, citizens from across India will come together to share their expert thoughts, ideas and suggestions with the Government in areas related to various policies, programs, schemes etc. MyGov aims to empower citizens to work hand in hand with the Government." Data sources used are statistical offices, large-scale surveys and social media. The data is analysed using, text, sentiment and trend analysis.
GeoConnections http://geodiscover.cgdi.ca/web/guest/home	Natural Resources Canada	GeoConnections is a national partnership initiative led by Natural Resources Canada designed to facilitate access to and use of geospatial information in Canada through the development, integration and use of the Canadian Geospatial Data Infrastructure (CGDI). The CGDI is an on-line resource that improves the sharing, access and use of geospatial information – information tied to geographic locations in Canada. It helps decision-makers from all levels of government, the private sector, non–government organizations and academia make better decisions on social, economic and environmental priorities. The infrastructure itself consists of data, standards, policies, technologies and partnerships that are in place to allow the sharing and visualisation of information on the internet. The main data sources stem from sensor-based data and large-scale survey data. The data is analysed using geo analytical techniques.
Groningen Declaration Network www.groningendeclaration.org	Dienst Uitvoering Onderwijs (Education Executive Agency)	The management of student credentials today is like a cottage industry: it is inconsistent, localized, and inefficient. A new approach is needed that makes learners the owners of their data and enables them to control access. To redefine the digital educational data ecosystem, barriers related to privacy, security, and other issues must be resolved. This new paradigm also must be flexible enough to accommodate differences across countries. The Groningen Declaration signatories realize that it is not practical to dictate how countries exchange digital student data. Data standards can make systems more scalable; however, multiple models will be used. For example, some countries may embrace a distributed approach where each institution maintains student information, while other countries may take a more centralized approach. Each has

Name of initiative & website	Organisation	Summary[126]
		advantages and disadvantages, but neither will be right for every situation. The Groningen Declaration acknowledges this and encourages stakeholders in the Digital Student Data Ecosystem to seek convergence and work on the realization of a truly citizen centred approach to global human capital cross border mobility. By bringing this vision to life, big data on student and skilled worker mobility will easily become available for large-scale data analysis, providing new inferences and insights about education worldwide. Tools and techniques exist to analyse large pools of data and generate useful insights about education. This sort of analysis can only succeed, however, with access to digital data. As more and more data is digitized, it will be possible to evaluate the success of different education programs and also predict what contributes to student success. Data is used from statistical offices and administrative data sources. Often the student data repositories do not analyse the data themselves, although there are exceptions (e.g. the US based national Student Clearinghouse).
I Paid Bribe http://www.ipaidabribe.com	Janaagraha	Ipaidabribe.com is Janaagraha's unique initiative to tackle corruption by harnessing the collective energy of citizens. You can report on the nature, number, pattern, types, location, frequency and values of actual corrupt acts on this website. Your reports will, perhaps for the first time, provide a snapshot of bribes occurring across your city. Janaagraha will use them to argue for improving governance systems and procedures, tightening law enforcement and regulation and thereby reduce the scope for corruption in obtaining services from the government. Janaagraha invites people to register any recent or old bribes you have paid. IPaidABribe.com is an initiative to tackle corruption by tapping into the collective energy of citizens, almost every one of whom is either a victim, perpetrator or beneficiary, or a combination of these. Janaagraha has a very rigorous system of aligning its institutional objectives with the performance of its programmes. Each institutional objective is further aligned with several strategic themes. Ipab aligns with Janaagraha's goal of making a measurable difference in the quality of Citizenship in cities. It does this by adopting the strategy of empowering citizens using a net-plus- roots approach – we use a net-based platform for recording the grassroots experiences of citizens. Through this approach, ipaidabribe.com aligns completely with Janaagraha's key thrust of increasing the number of citizen change agents in the city. We believe that every citizen who reports a story on our website about paying a bribe, is angry enough to begin to resist it. Of course, it goes without saying that somebody who has resisted paying a bribe and reported the story is already very much on the path of being a citizen change agent. Data used is a large-scale crow sourced survey. Janaagraha generates simple descriptive statistics of the data.
CIHI http://www.cihi.ca	Canadian Institute for Health Information	CIHI is an independent, not-for-profit organization that provides essential information on Canada's health system and the health of Canadians. Its stakeholders use their broad range of health databases, measurements and standards, together with the evidence-based reports and analyses, in their day-to-day decision-making. CIHI protects the privacy of Canadians by ensuring the confidentiality, integrity and availability of our health care information. CIHI's vision: Better data. Better decisions. Healthier Canadians. 28 pan-Canadian databases across health sectors are included in this project. The data stems from administrative data and statistical offices. The data is analysed using descriptive statistics and profiling.

Name of initiative & website	Organisation	Summary[126]
Public Policy.ie http://www.publicpolicy.ie	Irish Fiscal Policy Research Centre Limited	The goal of this project is to promote better economic, social and environmental policies. The Irish Fiscal Policy Research Centre does this by carrying out independent research and communicating the results to make it as easy as possible for interested citizens to understand the choices involved in addressing public policy issues and their implications. They believe in the value of evidence in helping to make the right policy choices. The Research Centre is independent and has no party political view and is not aligned with any sectional interest. They paid particular attention to three areas where the Government was committed to raising significant amounts of revenue to close the gap between revenue and expenditure, Local property tax domestic water charges, Reform and reduction of tax reliefs to promote pension saving They also cover the issue of public sector reform. If productivity in the public sector is increased, reductions in front line services will be less. Particular attention is devoted to the reforms in the health service. It is essential that the adjustments in the public finances are not only fair but seen to be fair. As a result they present evidence on the distribution of incomes in Ireland and the equity of individual policy choices. How the Reseach Centre operates: They present the evidence as they find it, to help the readers come to an informed view on the issues. Where they express an opinion, they set out the evidence to allow people make up their own minds as to whether or not they agree with the conclusion. Data collected stems from administrative data sources and small-scale surveys. The data is analysed with descriptive statistics and trend modelling.
Ireland National Transport Authority data analysis https://www.nationaltransport.ie/planning-policy/data-analysis/	Ireland National Transport Authority - Udaras Nasiunata Lompair	To support its transport-planning role, the authority undertakes analysis of available land use and transport data for the country on an on-going basis. The development of the GDA transport strategy and the provision of advice and support on transport and planning issues nationally is backed-up by analysis of population and travel behaviour from data made available primarily through the Census. The Authority maintains a transport model to support its transport planning remit within the GDA and is in the process of developing a National Transport Model. The Authority also periodically undertakes its own transport surveys to supplement the Census data. Full reports on these surveys – including analysis of the survey data may be found here. The demographic and transport data collected in Census 2011 was released to the Authority in full in late 2012. This data set provides detailed information on the patterns of travel to education – in addition to the travel to work patterns that have been recorded since the 2002 Census. The Integrated Planning Team has been using this data extensively across the state in order to provide evidence and support for the following key functions of the Authority: - Land use and transport planning; - Advice and support to other authorities; - Transport capital investment; - Public transport network reviews; - Bus licensing; - Smarter Travel workplaces; - Green Schools. Data collected stems from administrative data sources, large-scale surveys and statistical offices. The data is

Name of initiative & website	Organisation	Summary[126]
		analysed with descriptive statistics and GDA model network analysis, using a self-developed platform.
e-petitions http://epetitions.direct.gov.uk	HM Government	A website which allows citizens to petition the government and parliament. Petitions with other 10000 signatures and with over 100000 signatures are considered for parliamentary debate. Approximately 20000 petitions have been made, but only a small fraction has met these thresholds. The main data source is thus the petition/forms filled in on the website. However, log data and data about social media where petitions are disseminated have been collected. Government has used only descriptive statistics, researchers have done more (trend analysis mainly).
Data.gov.uk http://data.gov.uk	UK government (cabinet office, government digital service)	The Government is releasing public data to help people understand how government works and how policies are made. Some of this data is already available, but data.gov.uk brings it together in one searchable website. Making this data easily available means it will be easier for people to make decisions and suggestions about government policies based on detailed information. Hear more about the Government's Transparency agenda from the Prime Minister. There are datasets available from all central government departments and a number of other public sector bodies and local authorities. From data.gov.uk, people can access the raw data driving government forward. This can then be used by people to build useful applications that help society, or investigate how effective policy changes have been over time. There are different data sources used: small and large scale survey data, administrative data, data from statistical offices and commercial/traded business/consumer data. It's an open data platform, so only some descriptive/trend statistics are given. Users are supposed to use ow tools for further analysis.
AADHAAR project https://uidai.gov.in	Unique Identification Authority of India	The UID sets out to provide 'residents' of India (not just its citizens) with a unique 12 digit identification number linked to certain biometric and personal demographic data about them. The project has been promoted as a way to provide legal identity to residents of India and ensure that genuine beneficiaries are able to access welfare services. A central feature of this scheme is that the UIDAI through private agencies are collecting personal biometric information including fingerprints and a scan of both the iris in the individuals' eyes. The Unique Identification Authority of India (UIDAI) was set up within the Planning Commission in January 2009 and made responsible for development and implementation of the technical, legal and institutional infrastructure necessary for this system of identification numbers. The UIDAI also "owns" the data collected, according to an official notification. Through a somewhat unusual appointment procedure, the Prime Minister 'invited' Nandan Nilekani, a successful entrepreneur in the information technology sector, to be the Chairman of the project with the rank of a Cabinet Minister. The UID's stated purpose is a way for people who before had participated in only informal markets to be included into the formal economic and legal system, as well as creating a system with the possibility of enabling efficient delivery of services. One of the main policy applications for the UID is to act as a bridge for financial inclusion, driven by transfer of benefits in cash on a banking platform electronically, making the UID platform available for beneficiary identification and

Name of initiative & website	Organisation	Summary[126]
		verification at various stages. The UIDAI has announced plans to roll out UID-enabled payment system using a concept called the micro ATM. The UID draft law also allows disclosure of the personally identifiable information to government agencies and the court in the interest of national security. Provisions allow for disclosure of information to government agencies and the courts as well as the creation of a transaction record of the UID number holder. This suggests that the UID is intended to be a key part of the law enforcement and intelligence apparatus. The Additional Director-General of Police stated publicly that the UIDAI would eventually 'converge' with the Police department over time Although the UID Project was launched in September, 2010, the 'National Identification Authority of India Bill, 2010' was introduced in the Parliament on 3.12.2010. This Bill was referred to the Parliamentary Standing Committee on Finance which tabled its report in Parliament in 13.12.2011. The Committee found the draft law wanting in several respects, it also expressed its disapproval on how the Union Government had proceeded with the project even when it awaited Parliamentary sanction
The Riksdag's open data, data.riksdagen.se http://data.riksdagen.se	The Riksdag Administration	In 2010, the Riksdag opened up its databases, making them freely available to anyone who wishes to use them. The open data system is an important tool for providing access to the work of the Riksdag and its decisions. During 2013-2014, the open data has been further developed. According to the Swedish principle of public access to official documents, all citizens have the right to access official documents. The principle of public access to official documents has been constitutionally regulated since the 18th century. In the PSI Directive, the EU has decided that it should be easier for both private people and companies to use public information for both commercial and non-commercial purposes. PSI stands for "public sector information". In Sweden, the PSI Directive is implemented through the Act (2010:566) on the Re-use of Public Administration Documents, which the Riksdag decided on in 2010. The Act limits such things as the conditions an authority or agency may impose on people and companies who wish to re-use information provided by the authority or agency. The data includes administrative data and 300 000 documents (committee reports, private members' motions) going back to 1971, laws from 1809, calendars/planning information, debates, information about members, results of votes, member's speeches in the Chamber and so on. With the help of this data, developers, journalist, researchers and other interested parties can build their own applications and analyse the data. Analytics such a text analysis, profiling and descriptive statistics are often used.
PSI-datakollen (Public sector information data protocols) http://www.psidatakollen.se	VINNOVA, plus the e-delegation	Sweden has taken the initiative, joint between VINNOVA, a public sector agency which is promoting innovation, and the E-delegation, a department within the ministry for trade and industry, to open data for re-use in accordance with the EU's PSI directive. The main data sources are administrative data and commercial/traded business/consumer data.
Sunlight Foundation http://sunlightfoundation.com	Free standing foundation	The Sunlight Foundation is a nonpartisan non-profit that advocates for open government globally and uses technology to make government more accountable to all. The Sunlight Foundation does so by creating tools, open data, policy

Name of initiative & website	Organisation	Summary[126]
		recommendations, journalism and grant opportunities to dramatically expand access to vital government information to create accountability of our public officials. Their vision is to use technology to enable more complete, equitable and effective democratic participation. The overarching goal is to achieve changes in the law to require real-time, online transparency for all government information, with a special focus on the political money flow and who tries to influence government and how government responds. And, while our scope began with only a focus on the U.S. Congress, we now are defining open government on the local, state, federal and international level. The Sunlight Foundation believes that information is power, or, to put it more finely, disproportionate access to information is power. They are committed to improving access to government information by making it available online, indeed redefining "public" information as meaning "online." The Foundation approaches its work in a number of ways. The main data sources are all kind of administrative data and also data on election campaigns. They work with thousands of software developers, local transparency activists, bloggers, on and off-line active citizens and journalists, involving them in distributed research projects, hackathons, targeted lobbying and training. Sunlight's Policy team pushes for improved transparency policy through NGO efforts like OpeningParliament.org, and through traditional lobbying of government. Their reporters cover political influence stories both through reporting and through close collaboration with technical staff, leveraging computer-assisted reporting and data visualisation techniques. And in Sunlight Labs, teams of technologists and designers create apps and websites to bring information directly to citizens, as well as building and maintaining APIs (that support js on querying datasets) that power the applications of others.
Global Terrorism Dataset http://www.start.umd.edu/gtd/	National Consortium for the Study of Terrorism and Responses to Terrorism	The Global Terrorism Database (GTD) is an open-source database including information on terrorist events around the world from 1970 through 2013 (with additional annual updates planned for the future). Unlike many other event databases, the GTD includes systematic data on domestic as well as transnational and international terrorist incidents that have occurred during this time period and now includes more than 125,000 cases. For each GTD incident, information is available on the date and location of the incident, the weapons used and nature of the target, the number of casualties, and--when identifiable--the group or individual responsible. Statistical information contained in the Global Terrorism Database is based on reports from a variety of open media sources. Information is not added to the GTD unless and until we have determined the sources are credible. Users should not infer any additional actions or results beyond what is presented in a GTD entry and specifically, users should not infer an individual associated with a particular incident was tried and convicted of terrorism or any other criminal offense. If new documentation about an event becomes available, an entry may be modified, as necessary and appropriate. Thus the data is crowd-sourced. The National Consortium for the Study of Terrorism and Responses to Terrorism (START) makes the GTD available via this online interface in an effort to increase understanding of terrorist violence so that it can be more readily studied and defeated. They provide descriptive and trend statistics.
Star Metrics https://www.starmetrics.nih.gov/Star/About	NIH (hosts the system), plus various US federal agencies including	STAR METRICS® is a federal and research institution collaboration to create a repository of data and tools that will be useful to assess the impact of federal R&D investments. The National Institutes of Health (NIH) and the National

Name of initiative & website	Organisation	Summary[126]
	Energy, Agriculture.	Science Foundation (NSF), under the auspices of Office of Science and Technology Policy (OSTP), are leading this project. This project has been developed after a successful pilot project was conducted with several research institutions in the Federal Demonstration Partnership (FDP). The STAR METRICS® project consists of two implementation levels: Level I Developing uniform, auditable and standardized measures of the impact of science spending (ARRA and non-ARRA) on job creation, using data from research institutions' existing database records. No personally identifiable information (PII) is collected in Level I. Level II Through the Federal RePORTER tool, developing an open and automated data infrastructure that will enable the documentation and analysis of a subset of the inputs, outputs, and outcomes resulting from federal investments in science.
UN Global Pulse http://www.unglobalpulse.org	United Nations	Global Pulse is a flagship innovation initiative of the United Nations Secretary-General on big data. Its vision is a future in which big data is harnessed safely and responsibly as a public good. Its mission is to accelerate discovery, development and scaled adoption of big data innovation for sustainable development and humanitarian action. The initiative was established based on the recognition that digital data offers the opportunity to gain a better understanding of changes in human well-being, and to get real-time feedback on how well policy responses are working. To this end, Global Pulse is working to promote awareness of the opportunities big data presents for relief and development, forge public-private data sharing partnerships, generate high-impact analytical tools and approaches through its network of Pulse Labs, and drive broad adoption of useful innovations across the UN System. Global Pulse functions as a network of innovation labs where research on big data for Development is conceived and coordinated. The program has partnerships with experts from UN agencies, governments, academia, and the private sector to research, develop, and mainstream approaches for applying real-time digital data to 21st century development challenges. An example of a project is the analysis of Twitter data to map climate change discourse around the world. Using text analysis the discourse is mapped and subsequently analysed for trends.
Research and Innovation Observatory (RIO)	European Commission: DG JRC (J6)	The Research and Innovation Observatory was established in 2012 in response to the request from Commissioner Geoghegan-Quinn for high quality timely data on research and innovation. The role of the Observatory is to provide European policy makers and stakeholders, and above all DG RTD as RIO's main client, with data, policy information and analyses on R&I in the EU Member States (and to a degree in the Associated Countries). For this purpose a database has been set up which is gradually populated with R&D&I data, indicators (including their visualisations), policy documents and research studies from different accredited sources. Key indicators are produced and stored in the database through filtering, merging and aggregating quantitative data. All Eurostat, OECD and World Bank data, and RIO key indicators based

Name of initiative & website	Organisation	Summary[126]
		on them are automatically updated on a regularly base. At the same time, a website which will provide access to this repository is under development jointly with DG RTD. All data comes from statistical offices of the member states. Data is collected and stored using centralised databases (Oracle SQL) and analysed using desktop tools (Excel, R, SPSS, SAS).
Geography of Digital Innovation & Technologies (GeoDIT)	European Commission, DG JRC.J03 Information Society Unit	The operational pilot GeoDIT started last year building on the big amount of data already available at unit level and on the previous experience in different projects, aiming at deploying higher detail of information in a flexible and effective manner, so to better serve different policy needs by allowing for monitoring, benchmarking, profiling and analysis different aspects of ICT industry segments and providing insights at the level of actors, technologies and locations. The project is powered by the exploitation of different data analysis approaches, so to explore newest capabilities offered by data science on one hand to provide new perspectives and on the other to serve as a basis for further implementations. The project aims at providing insights suitable to support specific policy questions, coming from presently running projects for different client DGs, involving the merging and simultaneous use of many data sources (presently more than 10, encompassing both private and public sources), thus allowing for dynamic analysis both at micro and at aggregate level, with wide geographical scope (EU at city or functional urban unit, NUTS3, NUTS2, Country level; US and other major economies). The methodologies applied are those typical of big data analysis, as well as the tools used, even if the datasets on which they are applied are not sharing all the characteristics generally considered to refer to big data as such. A strong emphasis is put on the exploration of new techniques and methodologies such as knowledge discovery, pattern recognition, text and semantic analysis, as well as on the visualisation techniques.
European ICT Poles of Excellence (EIPE) http://is.jrc.ec.europa.eu/pages/ISG/EIPE.html	European Commission, DG JRC.J03 Information Society Unit	The European ICT Poles of Excellence (EIPE) Project (cofounded by DG CNECT, JRC-IPTS / DG CONNECT Joint Project nr. 31786-2010-06) is to set the general conceptual and methodological conditions for defining, identifying, analysing and monitoring the existence and progress of current and future EIPE, as to develop a clear capacity in distinguishing those among the many European ICT clusters, benchmarking them with non-European poles, observing their dynamics and offer a thorough analysis of their characteristics. The EIPE research project investigated the issues of growth, jobs and innovation, which have become the main priorities of the European Union's growth strategy programme 'Europe 2020'. The overall objectives of the EIPE project are to set the general conceptual and methodological conditions for defining, identifying, analysing and monitoring the existence and progress of current and future EIPE, in order to develop a clear capacity to distinguish these among the many European ICT clusters, observe their dynamics and offer an analysis of their characteristics. The EIPE project spanned the period between 2010 and 2013. Over this time, it developed a tool based on a database of original ICT activity indicators, which was enriched with geographical information to allow localisation and aggregation at NUTS 3. The tool helps to answer such questions as:

Name of initiative & website	Organisation	Summary[126]
		• How is ICT R&D, innovation and economic activity distributed in Europe? • Which locations are attracting new investments in the ICT sector? • What is the position of individual European locations in the global network of ICT activity? The project makes use of several sources of data, ranging from official statistics provided by Eurostat and OECD (and other Statistics offices, and including data resulting from large scale survey like CIS) to commercial business data and data about investments, venture capital, etc., to data regarding EC funded programmes (FP7, H2020, etc.). Data produces in the framework of other JRC J03 projects (such as the PREDICT 2 project) are used as well. Patent data provided by EPO (addressing patent submitted to most of Patent Offices worldwide) are included too.
Prospective Insights on R&D in ICT (PREDICT, 2nd phase) http://is.jrc.ec.europa.eu/pages/ISG/PREDICT/PREDICT2014/home.html	European Commission, DG JRC.J03 Information Society Unit	The main objective of the PREDICT project (cofounded by DG CNECT) is to provide an analysis of private and public R&D investments in the EU Information and Communication Technology (ICT) sector. It also analyses ICT R&D performance through a worldwide analysis of patenting in ICT, and it developed in its first phase an analysis of the increasing internationalisation of ICT R&D. In its second phase that started in 2011, PREDICT will also analyse macroeconomic impacts of ICT R&D through a modelling approach. PREDICT has become a unique source of information on ICT R&D investments in the EU, also benchmarking them against those of the EU's main global competitors. PREDICT combines in a unique way three complementary perspectives: national statistics (covering both private and public R&D expenditures), company data, and technology-based indicators. It relies on the latest available official statistics delivered by Member States, Eurostat and the OECD. This data still contains gaps and where this is the case, rigorous cross-checking and estimating methods have been applied by JRC-IPTS to provide the study with the necessary set of data. The data are analysed using descriptive statistics and trend analysis.
S3 Platform: assisting EU countries and regions to develop, implement and review their Research and Innovation Strategies for Smart Specialisation (RIS3) http://s3platform.jrc.ec.europa.eu/s3-tools	*JRC-IPTS*	The S3 Platform assists EU countries and regions to develop, implement and review their Research and Innovation Strategies for Smart Specialisation (RIS3). The platform has complementary three tools that use (big) data: - **Eye@RIS3**: an online database of RIS3 priorities. It is intended as a tool to help strategy development rather than a source of statistical data. Regions are requested to introduce/update input in the database, which will produce a realistic map of the process of RIS3 development (descriptive statistics). This tool collects and visualises the innovation priorities of almost all states and regions in the European Union and in neighbouring countries. It shows market trends and facilitates cooperation. Data are validated by policy makers who form part of our platform/network. - **Regional benchmarking tool**: A key to build sound innovation strategies for smart specialisation at the regional level is to identify opportunities for learning policy lessons and transferring practices from other regions. For this, it is important to start by identifying regions that share similar structural conditions that are relevant for innovation-driven development (social, economic, technological, institutional and geographical characteristics). That is, characteristics that cannot be easily changed in the short term and that are demonstrated to affect the way innovation and economic

Name of initiative & website	Organisation	Summary[126]
		evolution take place in a region. We have developed an interactive tool that allows the identification of reference regions across Europe based on a methodology jointly developed by Orkestra – Basque Institute of Competitiveness and JRC-IPTS. Data from statistical offices and large-scale survey data is used. The data is analysed using descriptive statistics and benchmarking. - **S3 Inter-regional Trade and Competition Tool:** This is a web tool by the Joint Research Centre, the European Commission's in-house science service and the PBL Netherlands Environmental Assessment Agency. It aims to provide evidence-based scientific support to the European policy making process. It is based on inter-regional trade data developed by the PBL Netherlands Environmental Assessment Agency. It is composed of four integrated applications: Regional import/export map, revelled competitors and regional competitiveness scoreboard map, competitiveness scores and trade network scores/map. The tool uses data from statistical offices and commercial/traded business data. Different analysis methodologies are used: Descriptive statistics, benchmarking, profiling and network analysis
Big-data based Youngju Apple Harvest Project http://bigapple.yeongju.go.kr	Gyeongsangbuk-Do - The Ministry of the Ministry of Science, ICT and Future Planning - The National Information Society Agency	Gyeongsangbuk-Do (a local government), the Ministry of Science, ICT and Future Planning, and the National Information Society Agency (a statutory agency), developed a big data model to detect damage by blithe and insects, and to predict supply, demand, and price of apples. The initiative is based on the Youngju region, which locates within Gyeongsangbuk-Do, since Youngju produces the largest amount of apples in South Korea, and about 60% of the national output of apples is dependent on the Kyungsangbook-Do. They collected and analyzed various forms of data, such as daily climate statistics, real-time u-IT sensor data from farms, daily/monthly/annual market data, and social media keywords. Based on the model, they have been providing relevant information such as the trends of apple prices and outputs to farmers and citizens, via web, IPTVs, and smartphones. For this information service, they also established wifi in 25 apple farms, and provided IPTVs to 7 community halls. They used descriptive statistics, text analytics, trend analysis and predictive analytics. The initiative has implications as it provides useful information for an apple output increase, marketing strategies, and farm-products price policies.
European Consumer Complaints Registration System (ECCRS)	European Commission, DG JUST	In 2010, more than one in five Europeans encountered a problem with a product or service for which they thought they had cause for complaint. It was estimated that the total detriment European consumers incurred from problems amounted to about 0.4% of EU GDP. Consumer complaints collected by the relevant consumer organisations are the key source of information on the state of all sectors and markets across EU and challenges faced by consumers. Hence they are also key evidence for formulating any policy response, be it at the EU or at the national level. However the multiplicity of complaints handling bodies and different complaints classification used made very difficult and cost extensive to know in details citizens problems. Therefore In May 2010 the Commission adopted the Recommendation on the use of a harmonised methodology for classifying and reporting consumer complaints and enquiries. The Recommendation is addressed to all third-party complaint bodies (national authorities, consumer organisations, ADR bodies) and calls on them to classify

Name of initiative & website	Organisation	Summary[126]
		complaints according to common taxonomy and to report the data to the Commission. The system does not cover consumer complaint handling mechanisms operated by traders. Step by step, the data-driven approach is being improved in terms of geographic coverage (adding countries), data sources, data analytics (complaints per sector, product, selling method, etc.), openness of data (moving from a restricted set of users to sharing data with citizens, which also allows for confronting the rigorous, European dataset with other types of data, e.g. social media data about consumer complaints).
Digital Single Market: survey data and clickstream data	European Commission, DG JUST	In 2015, the European Commission (DG JUST) is conducting a consumer survey as supporting evidence base to a Commission study identifying the main cross-border obstacles to the Digital Single Market and where they matter most. Survey data will be complemented with clickstream data being collected from the survey respondents. Combining the objective clickstream data with more subjective survey data should generate a richer picture of the drivers and impediments to online consumer behavior for domestic and cross-border online services. More specifically, the consumer clickstream and survey data could be used to estimate an econometric model of online consumer expenditure. It is requested that from the clickstream data the contractor gets detailed insights into the time consumer respondents have spent in the various websites or internet platforms and the location of these websites (in order to also validate whether these relate to domestic or cross-border commerce). Given the main objective of this clickstream survey, i.e. studying online purchase journey, clickstream data appears to be necessary to assess consumer respondents' online behaviour, i.e. websites visited, time spent on each website, type of online purchases carried out (where relevant) etc. It is important to know from which websites consumers buy these online services. The website name can then lead us to the country of origin of the supplier of these services and thereby reconstruct online cross-border trade in the EU Digital Single Market. Ideally, that information could be obtained in an objectively measurable way from underlying clickstream data in online consumer panels. However, we are still asking the question if consumers buy domestically or abroad, to check consistency between subjective replies and objective measurement.

Technopolis, OII, CEPS (2015)

Appendix D Inventory template

Data for policy: big data and other innovative data-driven approaches for evidence-informed policymaking.

Topics for the inventory of relevant initiatives, v6 (final), 9 March 2015

Inventory template

Variable	Options *(and variable type)*	Results
Name of the initiative	• Name: *(Open)*	
Web address	• URL: *(Open)*	
Direct contact *(When possible)*	• Name, function, contact: *(Open)*	
Description of the initiative *(Short, stand-alone description answering to the basic who? what? where? how? when? questions, and highlighting other note-worthy elements. The description should clarify both the policy context and the data-driven approach)*	• Free text, max. 500 words *(Open)*	
Launch of the initiative	• Year: *(Single choice)*	
Status of the initiative *(Has the initiative been used for policy yet, or is it still in development?)*	• Pilot or demonstrator, full implementation to be discussed • Pilot or demonstrator, full implementation foreseen • Full implementation • Other *(Single choice. If possible, distinguish between pilot and demonstrator)*	
Responsible authority *(Which organization runs this?)*	• Name: *(Open)*	
	• Country or location of main office: *(Open)*	
Type of responsible authority *(Overall umbrella into which the organization fits)*	• International institution other than EU (WHO, UN, OECD, NAFTA, etc.) • EU institution • National government • National agency • Regional government • Regional agency • NGO • Other, please explain *(Multiple choice, in case initiatives are the result of close collaboration between two types of*	

	authorities)	
Policy area(s) addressed by the data for policy initiative *(Categories taken from EC. If initiatives do not fit, please write 'other' and explain)*	• Agriculture • Humanitarian Aid • Banking and finance • Budget • Civil Protection • Climate Action • Communications Networks • Competition • Consumer Affairs • Culture • Customs • Development Aid • Digital Agenda/Information Society • Economic and Financial Affairs • Education • Employment • Energy • Enlargement (of European Union) • Enterprise and Industry • Environment • Equal opportunities • External relations (EC terminology) • International relations • Fisheries • Food Safety • Home Affairs • Internal/single market • Justice and Fundamental Rights • Mobility • Public Health • Regional Policy • Research • Social Affairs • Sport • Taxation • Trade • Trans-European networks (EC terminology) • Transport • Youth • Other, please explain *(Multiple choice: please choose one primary, max. 3 secondary)*	
Policy objective(s) addressed	• Main policy objectives or policy issues addressed:	

	(Open)	
(Expected) use of the data in the policy cycle *(In terms of policy, what does this initiative do, and is it make policy [=early in the cycle] or evaluate it [=late] or in between?)*	• Foresight and agenda setting • Problem analysis • Identification and design of policy options • Ex-ante impact assessment and evaluation • Policy implementation • Monitoring and interim evaluation • Ex-post evaluation and impact assessment • Other, please explain *(Single primary use, max. 3 secondary use)*	
Characteristics of data in terms of data sources and data variety *(Some categories overlap but this is hard to avoid, e.g. statistical offices use multiple data sources, try to identify the main ones)*	• Statistical offices (irrespective of how these offices obtain the data) • Large-scale survey data (recurring international and national surveys that cover several themes, e.g. the European Social Survey) • Small-scale survey data (e.g. recurring or ad-hoc surveys that address a specific target group and a specific theme, e.g. participants of Horizon 2020) • Administrative data (e.g. about citizens, housing prices, businesses and subsidies) • Commercial/traded business data, consumer data, market data • Social media • Sensor-based data • Other, please explain *(Multiple choice. Please explain)*	
Data velocity *(If several choices, make sure to explain which velocity applies to which choice)*	*(link these to data sources above so these are sub-types)* • Real-time data • Daily • Weekly • Monthly • Annual • Less than annual *(Multiple choice)*	
Data format, data interoperability and data linking	• Short description, including specific examples, e.g. *SQL, TSV, Excel, RDF (Open)*	
Data volume	• Deep data: a large number of observations per unit of observation • Broad data: a small number of observations per unit of observation • Both/mixture • Other *(Single choice. Please explain)*	

Data veracity *(Here, we work with three 'pairs' of two options, against the background of very generic issues such as correct/incorrect data, and truthfulness/deception (hard to assess), objectivity/subjectivity (this correlates with data source) and credibility/implausibility (hard to assess)).*	• Verified data (e.g. explicit verification by firms or citizens) • Unverified data • Near complete data (e.g. in terms of units of observation, number and timing of observations, geographic scope) • Incomplete data • Highly consistent data (e.g. in terms of definitions and indicators used, methods of data collection, data sources) • Inconsistent data *(Single choice per pair. Please explain and add other observations about data veracity. See examples mentioned on the left)*	

Geographic scope and granularity of the data	• What is the geographic scope/coverage of the data:
	  • Multiple choice
	  • Global
	  • European/regional
	  • Country
	  • Sub-country level
	  • Other, please explain
	  (Single choice)
	• To what extent is the data available at disaggregated level (e.g. per country, region or sector)? *(open)*	

Characteristics of data in terms of level of openness of data *(If there is ambiguity here, please explain)*	• Fully open and using open data formats (e.g. Creative Commons Attribution Share-Alike or Open Database License) • Open, to a large extent (e.g. Creative Commons Attribution 4.0 or Attribution License) • Open, to some extent, with an emphasis on being dedicated to the public domain, with rights waived (e.g. Creative Commons CCZero or Public Domain Dedication and License, PDDL) • Restricted set of users on a commercial basis (e.g. subscriptions) • Restricted set of users on a non-commercial basis (e.g. authorised researchers and policy makers only) *(Multiple choice, please explain. Some data (see question on data sources) may be less open than other data).*	
Data analysis methodologies	• Descriptive statistics • Benchmarking • Text analytics • Sentiment mining • Trend analysis	

	• Predictive analytics • Profiling • Agent-based modelling • Other *(Multiple choice, please explain)*	
Data analytics tools and platform *(Link data analytics tools/platforms to data analysis methods, previous question?)*	• Brief description, mentioning the specific tools and platforms and (if possible) position these under one of three categories): - High performance computing, e.g. Mapreduce, Hadoop ('high powered') - Centralized data servers, e.g. SQL databases and simple cloud based ('medium powered') - Desktop tools, e.g. Excel and Stata Databases ('low powered') • Link data analytics tools to data analysis methods (previous question). *(Open)*	
Visualisation tools	• Provide names of specific tools, e.g. Polymaps, Google Databoard *(Open)*	
Use of the data for policy initiative in policymaking? *(Are there ways to assess how data are being used, by whom, and for what?)*	• To what extent is the data for policy initiative used by policy makers, citizens, business and other stakeholders, and how is this monitored? *(Open)*	
Privacy	• How are privacy issues described and addressed? *(Open)*	
Inclusion *(If relevant)*	• How are inclusion issues addressed, e.g. being transparent about or addressing any bias in terms of age groups, income groups or urban and rural areas? For example, who does the initiative collects data from persons or with or without digital traces? *(Open)*	
Estimated added value *(What are the benefits, and are they being realized?)*	• Explain the extent to which the data for policy initiative allows for an improved assessment (more detailed, more often, etc.) of the impact of policy interventions and/or for identifying and analysing economic, social or environmental developments that require new or adapted policy interventions. *(Open)*	
Concluding remarks	• Any best practice, highly innovative or unique elements of the initiative? • Main barriers and how they are addressed? • Potential for replication? *(Open)*	

Technopolis, OII, CEPS (2015)

Appendix E Relevant research projects within Europe's 7th Framework Programme

In this appendix research projects within FP7 on innovative data-driven approaches for evidence-based policymaking are listed. The Global Systems Science platform bundles several of these projects, including some Horizon 2020 projects (such as the GRACeFUL project). These projects are presented with a short summary in section E.1 Other relevant FP7 projects are presented in a similar format in section E.2 Summaries are taken from project websites or the CORDIS project database.

E.1 Global Systems Science platform

The table below contains all the projects that are part of the Global Systems Science platform.[127]

Projects within the Global Systems Science platform

Project	Organisation (Consortium lead partner)	Summary
CIMPLEX: Bringing CItizens, Models and Data together in Participatory, Interactive SociaL EXploratories https://www.cimplex-project.eu	Deutsches Forschungszentrum für Künstliche Intelligenz	We propose visionary research to develop modelling, computational, and ICT tools needed to predict and influence disease spread and other contagion phenomena in complex social systems. To achieve non-incremental advances we will combine large scale, realistic, data-driven models with participatory data-collection and advanced methods for big data analysis. In particular we will go beyond the one-dimensional focus of current approaches tackling one aspect of the problem at a time. We will interconnect contagion progression (e.g. epidemics) with social adaptation, the economic impact and other systemic aspects that will finally allow a complete analysis of the inherent systemic risk. We will develop models dealing with multiple time and length scales simultaneously, leading to the definition of new, layered computational approaches. Towards policy impact and social response we will work to close the loop between models, data, behaviour and perception and develop new concepts for the explanation, visualisation and interaction with data and models both on individual and on collective level. We will cast the fundamental advances into an integrated system building on widely accepted open ICT technologies that will be used and useful beyond the project. As a tangible ICT outcome directed at facilitating the uptake and impact of the project, we will implement "Interactive Social Exploratories" defined as interactive environments which act as a front-end to a set of parameterisable and adjustable models, data analysis techniques, visualisation methods and data collection frameworks.
Consensus: Multi-Objective Decision Making Tools through Citizen Engagement http://www.consen	Institute of Communication and Computer Systems	Every real-world planning problem, especially in governance and policymaking, possesses several objectives that are typically subject to inherent conflicts with underlying trade-offs to be discovered. Policy makers are in need of proper tools that will utilize an overall analytical process, that is, assist in modelling the real-world planning process, automatically obtain the best attainable trade-offs, and facilitate efficient exploration in order to reach a final decision on which policy to implement.

[127] http://www.global-systems-science.eu

Project	Organisation (Consortium lead partner)	Summary
sus-project.eu		Consensus will strive to model existing real-world use-cases within the relevant policymaking context, and consequently employ measurable quantifiers in order to investigate how and whether preferable trade-offs can be identified. Those quantifiers will be sought in multiple realms – such as analytical models, numerical simulations, statistical tools and even public opinion evaluators – in order to link the domain data to the set of objectives, and by that to reflect the expected success-rate of policies and their implementation. Furthermore, Consensus intends to investigate the balance shift of the objectives, in scenarios where certain resources are being deployed to primarily address one of them through EU- or international-level policies. This investigation is meant to cover two important real-world use-cases: the one dealing with Biofuels and Climate Change (EU Renewable Energy Directive), and the other dealing with Transportation Networks (trans-European transport network guidelines). Consensus will also seek the citizens' involvement in policymaking according to this scheme, since their input can potentially become highly valuable in various stages, from gathering the necessary data, through formulating public opinion as one of the objectives in the model, to eventually playing the role of exploring the attained trade-offs and contributing to their weighing.
CRISIS: Complexity Research Initiative for Systemic Instabilities http://www.crisis-economics.eu	Universita Cattolica del Sacro Cuore	In his opening address to the ECB Central Banking Conference, Jean-Claude Trichet said that "in the face of the crisis, we felt abandoned by conventional tools", and went on to call for the development of complex systems based approaches to augment existing ways of understanding the economy. This project proposes to address this need by creating an integrated set of complex systems-based ICT tools for modelling the economy, of practical use to policy makers, to be used in both simulation and gaming modes. The core element of our project will be a pair of coupled agent-based models of the European economy, one for the financial system and one for the macro-economy. The European model will in turn be coupled to a corresponding American model developed in an independently funded sister project. These models will be carefully calibrated using a comprehensive data set. The models will be constructed around the available data, so that the resulting model can be placed in the existing state of the real economy at any given point in time, and used to simulate that economy going forward. The decision-making components of the model will also be calibrated using laboratory experiments with human subjects. The final result will include a sophisticated graphical user interface with open-source software. This model will allow us to do research to better understand the combined European and American economies, freed from the constraint of unrealistic assumptions such as perfect rationality and representative agents. The output of the research will be used to provide new insights for policy makers and evaluate quantitatively policy measures at the European level.
MD: Emergence By Design http://emergencebydesign.org	Universita Ca' Foscari Venezia	MD research aspires to satisfy two kinds of objective. The first is to develop the foundations of a theory of innovation dynamics that concatenates design and emergence. The second is to design a set of processes, enabled in part by new ICT, that can help to mobilize civil society to construct a socially sustainable future. These processes will assist the managers of online communities of networks of innovators to enhance the generative potential of relationships among members of these communities; monitor innovation cascades to detect signals pointing to the onset of endogenously generated social crises, as part of a system innovation policy that goes beyond the current strategy of priming the pump of invention; and implement a new kind of

Project	Organisation (Consortium lead partner)	Summary
		dynamic evaluation for socially-oriented innovation projects, which can provide stakeholders with a multivalent representation of the social consequences induced by the projects to help them steer the resulting cascades of change in socially positive directions. The process designs will be informed by the theory, and the theory will be informed by the experience of consortium partners immersed in the world of relevant practice.
EUNOIA: Evolutive User-centric Networks for Intraurban Accessibility www.eunoia-project.eu	Universitat de Les Illes Balears	Urban transport is essential for citizens to perform their daily activities, but it also constitutes a major source of pollution. The goal of EUNOIA is to take advantage of smart city technologies and complex systems science to develop new models and tools empowering city governments and their citizens to design sustainable mobility policies. EUNOIA will pursue advances in three complementary directions: • Use of data. The massive penetration of ICT is modifying social relationships and travel behaviour, and at the same time is providing us with a huge amount of heterogeneous data: intelligent transport systems, internet social networks, mobile phone call logs, e-transactions. EUNOIA will investigate how to exploit these data to characterise mobility and location patterns in different European cities. • Urban transportation models. EUNOIA will investigate the interactions between social networks and travel behaviour, e.g. the influence of social networks on the planning of joint trips. This will allow a more comprehensive assessment of mobility policies, particularly of new services emerging around the idea of a shared access to resources, such as car sharing. The new travel behaviour models will be integrated into state-of-the-art agent-based simulation tools. • Link between modellers, decision makers, and societal actors. The potential of urban simulation models is still little exploited in policy decision contexts. EUNOIA will develop tools, e.g. 3D visual analytics, allowing stakeholders' interaction with the simulation results, as well as a methodology for collaborative, multi-stakeholder policy assessment. In order to ensure maximum credibility and usability of the project results, the models and methodologies developed by EUNOIA will be tested and refined through several case studies conducted in close cooperation with policy makers and mobility stakeholders from the three cities participating in the project: Barcelona, London, and Zurich.
EU Community http://project.eucommunity.eu	Intrasoft International	EU Community merges ICT and social media networking with established online media and stakeholder groups to cultivate transparency, enhance efficiency and stimulate fresh ideas for EU policymaking. Three challenges are identified to reduce complexity and leverage under used technologies: • Ever more complex policy processes, thousands of contributors and documents, not easily understandable by people. REQUIRE reputation ranks and relevance filters; • eGov projects grow locally but have limited EU impact due to language, interface and trust barriers. REQUIRE co-development with stakeholders, leveraging visualisation technologies such as curated timelines; • Social media is prevalent but not helping EU legislation-making yet (except blocking ACTA). Novel approach: over 36 months, a consortium of leading research

Project	Organisation (Consortium lead partner)	Summary
		centres, ICT SME's and a large media network, will go from existing tools to further advanced prototype, pilot-testing and roll-out. They are supported by a number of high-calibre experts and a foundation serving as community guarantor. The results will be tested and deployed over an EU policy media network, with a track record of sustainability and multilingualism. Three pilots suiting the EU political mandates 2014-2019 have been selected (Future of EU, Renewable Energy and Innovation Strategy) and will be undertaken by a network of European stakeholders (policy-makers, journalists, experts, NGO's and informed citizens) in several EU countries, supported by localised policy media. Objectve: EU Community goes beyond current generation of policy modelling and argumentation tools. It provides decision makers with better policy options by combining social media interactions, qualified contributors, document curation, visual analysis plus online and offline trust-building tools. The results will be open source platforms, and the data itself will be open to re-use by other apps developers.
EveryAware: Enhance Environmental Awareness through Social Information Technologies. http://www.everyaware.eu	Fondazione Istituto per l'Interscambio Scientifico (I.S.I)	There is now overwhelming evidence that the current organisation of our economies and societies is seriously damaging biological ecosystems and human living conditions in the very short term, with potentially catastrophic effects in the long term. The enforcement of novel policies may be triggered by a grass root approach, with a key contribution from information and communication technologies (ICT). Nowadays low-cost sensing technologies allow the citizens to directly assess the state of the environment; social networking tools allow effective data and opinion collection and real-time information spreading processes. In addition, theoretical and modelling tools developed by physicists, computer scientists and sociologists have reached the maturity to analyse, interpret and visualize complex datasets. The proposed project intends to integrate all crucial phases (environmental monitoring, awareness enhancement, behavioural change) in the management of the environment in a unified framework, by creating a new technological platform combining sensing technologies, networking applications and data-processing tools; the internet and the existing mobile communication networks will provide the infrastructure hosting such a platform, allowing its replication indifferent times and places. Case studies concerning different numbers of participants will test the scalability of the platform, aiming at involving as many citizens as possible leveraging on the low cost and high usability of the sensing devices. The integration of participatory sensing with the monitoring of subjective opinions is novel and crucial, as it exposes the mechanisms by which the local perception of an environmental issue, corroborated by quantitative data, evolves into socially-shared opinions, eventually driving behavioural changes. Enabling this level of transparency critically allows an effective communication of desirable environmental strategies to the general public and to institutional agencies.
FOC-II: Forecasting Financial Crises www.focproject.eu	Scuola IMT Alti Studi di Lucca	In this project we set up an interdisciplinary consortium of computer scientists, physicists, economists and policy makers to deal with the problem of understanding and forecasting systemic risk and global financial instabilities. By leveraging on expertise in the various disciplines, we want to provide a novel integrated and network-oriented approach to the issue. On one hand, we will offer a theoretical framework to measure systemic risk in global financial market and financial networks. On the other hand, we will deliver an ICT collaborative platform for monitoring systemic fragility and the propagation of financial distress across institutions and markets around the world. Experts will be able to evaluate algorithms and models to

Project	Organisation (Consortium lead partner)	Summary
		forecast financial crises as well as visualise interactively possible future scenarios.
GLODERS: Global Dynamics of Extortion Racket Systems http://www.gloders.eu	University of Surrey	The GLODERS research project is directed towards development of an ICT model for understanding a specific aspect of the dynamics of the global financial system: Extortion Racket Systems (ERSs). ERSs, of which the Mafia is but one example, are spreading globally from a small number of seed locations, causing massive disruption to economies. Yet there is no good understanding of their dynamics and thus how they may be countered. ERSs are not only powerful criminal organizations, operating at several hierarchical levels, but also prosperous economic enterprises and highly dynamic systems, likely to reinvest in new markets. If stakeholders - legislators and law enforcers - are to be successful in attacking ERSs, they need the much better understanding of the evolution of ERSs that computational models and ICT tools can give them. GLODERS will provide a theory-driven set of computational tools, developed through a process of participatory modelling with stakeholders, to study, monitor, and possibly predict the dynamics of ERSs, as they spread from local through regional into global influence. The research will draw on expertise already developed in the small, but highly experienced multidisciplinary consortium to use: • computer-assisted qualitative text mining of documentary evidence; • guided semi-automatic semantic analysis of stakeholder narratives and other textual data; and • multi-level, stakeholder-centred agent-based modelling of the distributed negotiations between normative agents. These methods will advance the state of the art for using data to inform policy decisions. Throughout, the project will interact with a large, international group of stakeholder representatives from EU Ministries of Justice and police forces. The output will provide a set of ICT tools to facilitate strategic policies that could prevent the further penetration and extension of the global menace posed by ERSs.
GRACeFUL: Global systems Rapid Assessment tools through Constraint FUnctional Languages http://www.gracefulproject.eu	Universitat Politècnica de Catalunya	Global Systems Science is a FET Proactive initiative under Horizon 2020 that seeks to improve the way scientific knowledge can help inform and evaluate policy and societal responses to global challenges like climate change and global financial crises. The making of policies coping with such global systems is a process that necessarily involves the participation of stakeholders with very diverse backgrounds (political, legal, social, economic, technical, ecologic, etc.), each of them with their own interests, expectations, constraints, targets and objectives. People play a central role in this kind of collective decision making and generally the quest for solutions to a problem intertwines its very understanding and specification. To some extent, traditional “what-if” style computer simulations can assist in this process provided they employ the right narratives and adequate high-level, qualitative models to separate the policy question from the underlying scientific details. Domain-specific Languages (DSL) embedded in Functional Programming (FP) languages like Haskell offer a promising way to implement scalable and verifiable simulators. But the traditional use of simulators is essentially an iterative trial-and-error process, way too tedious for “on-line” execution in a moderated group session: possible policy measures have to be imagined and tried, their simulated effects

Project	Organisation (Consortium lead partner)	Summary
		contrasted with the desired objectives and targets, and this process has to be repeated many times until results are considered good enough for everyone. A paradigm shift is needed towards rapid assessment tools that are real problem solvers in which targets and stakeholders' objectives can be taken along as input from the very beginning and the required actions and measures to be implemented are the solutions produced. Constraint Programming (CP) has already demonstrated to enable a similar paradigm shift e.g. in the case of managed physical systems like water and power distribution networks. The GRACeFUL project pursues laying a base for domain-specific languages aimed at building scalable rapid assessment tools for collective policymaking in global systems. It involves several different disciplines. At the top policy-modelling level, we adopt and adapt the socially-inspired discipline of Group Model Building, well-known from system dynamics. This process is formalized through a dedicated embedded DSL and backed by visual forms of conceptual modelling and flavoured with gamification aspects and visual analytics. At the host-language level, we work on combining the declarative paradigms of CP and FP. In the latter context, specific work is being done on domain-specific constraints, constraint composition, and composable solvers and heuristics. Results are applied and validated for a specific problem case of Climate-Resilient Urban Design in the city of Dordrecht in The Netherlands, but the ambition is providing a general framework and approach applicable to several other Global Systems.
GROWTH-COM: Growth and Innovation Policy Modelling: Applying Next Generation Tools, Aata, and Economic Complexity Ideas http://www.growthcom.eu	Consiglio Nazionale delle Ricerche	In the wake of the financial crisis one of the major challenges faced by policymakers in Europe and around the world is re-kindling economic growth and innovation. Recent research portrays economic growth as a process of evolution of ecosystems of technologies and industrial capabilities. This research has also shown a strong relationship between the development of the complexity of economies in a globalized market and overall output growth. ICT based tools, in particular complex systems analysis, simulation, systems science methods, and big data capabilities offer policymakers new opportunities to empirically map technology and capability ecosystems of countries and industrial sectors, analyse their structure, understand their dynamics, measure economic complexity, and design policy interventions more likely to have positive impacts on growth over time. This project proposes to apply ICT tools in this way to yield insights on the industrial competitiveness and fragility of countries, the evolution of technologies and capabilities, the network of products, the adaptability of companies, and the ecology of E-commerce. In particular, E-commerce provides an important example of the active involvement of citizen in terms of their feedbacks on products and companies. The project will be a European and global collaboration amongst a network of researchers highly experienced in these topics. The teams involved represent an interdisciplinary consortium in which various research lines will be integrated in a network of collaborative effort to address challenging problems for a new vision of a data driven fundamental economics in a strongly connected, globalised world. The results of the project will provide a novel basis for economic forecasting and risk analysis for countries, companies and technological sectors, and they will also provide a new perspective for growth and innovation policymaking.

Project	Organisation (Consortium lead partner)	Summary
INSIGHT: Innovative Policy Modelling and Governance Tools for Sustainable Post-Crisis Urban Development http://www.insight-fp7.eu	Universidad Politecnica de Madrid	Cities embody the twofold challenge currently facing the European Union: how to improve competitiveness while achieving social cohesion and environmental sustainability. They are fertile ground for science and technology, innovation and cultural activity, but also places where problems such as environmental pollution, unemployment, segregation and poverty are concentrated. INSIGHT aims to investigate how ICT, with particular focus on data science and complexity theory, can help European cities formulate and evaluate policies to stimulate a balanced economic recovery and a sustainable urban development. The objectives of the project are the following: • to investigate how data from multiple distributed sources available in the context of the open data, the big data and the smart city movements, can be managed, analysed and visualised to understand urban development patterns; • to apply these data mining functionalities to characterise the drivers of the spatial distribution of activities in European cities, focusing on the retail, housing, and public services sectors, and paying special attention to the impact of the current economic crisis; • to develop enhanced spatial interaction and location models for retail, housing, and public services; • to integrate the new theoretical models into state-of-the-art simulation tools, in order to develop enhanced decision support systems able to provide scientific evidence in support of policy options for post-crisis urban development; • to develop innovative visualisation tools to enable stakeholder interaction with the new urban simulation and decision support tools and facilitate the analysis and interpretation of the simulation outcomes; • to develop methodological procedures for the use of the tools in policy design processes, and evaluate and demonstrate the capabilities of the tools through four case studies carried out in cooperation with the cities of Barcelona, Madrid, London, and Rotterdam.
NESS: Non-Equilibrium Social Science in ICT and Economics http://www.nessnet.eu	Universite de Fribourg	Traditional economics uses models that assume limited interaction between agents, simplistic behavioural assumptions and equilibrium analysis. Yet economies and social systems are often strongly correlated, heterogeneous and far from equilibrium hence traditional approaches do not provide the tools to understand, model or design systems in rapid change. The current financial and world order constitute such systems and heavily rely on ICT (such as trading platforms, social media and novel e-business models) that is transforming the way modern economies and societies function. For many years researchers from diverse disciplines and communities, such as Complex Systems, non-orthodox Economics and Agent-Based Modelling have proposed alternative formalised models that capture aspects of non-equilibrium social systems yet these endeavours have been fragmented and peripheral. With this co-ordination action we will bring together leading researchers within Non-Equilibrium Social Science (NESS) to build a community, address current pressing issues and engage with policy makers, business and other stakeholders. We aim to create a new ICT and NESS-aware paradigm for economics and foster its application in areas of policy and business for social and economic progress.
Policy Compass http://policycompa	Fraunhofer Gesellschaft zur Förderung der	The main goal of the project is to develop a research prototype of an easy-to-use, highly visual and intuitive tool for social networks and e-participation platforms, called the Policy Compass, enabling citizens and public officials to easily create, apply, share, embed,

Project	Organisation (Consortium lead partner)	Summary
ss.eu	Angewandten Forschung	annotate and discuss causal models, charts and graphs of historical data from trusted open data sources. The aim is to develop methods and tools that facilitate more factual, evidence-based, transparent and accountable policy evaluation and analysis. The research question we want to address in the Policy Compass project is whether and how open public data, social media, e-participation platforms, fuzzy cognitive maps and argumentation technology can be integrated and applied to provide better tools on the World-Wide Web for constructing, sharing, visualising and debating progress metrics and causal models of policies. The Policy Compass will make better use of Europe's open public data resources and empower policy-makers and citizens (especially the younger generation) to better assess government policies in the policy analysis and monitoring phases of the policy cycle.
SENSE4US: Data Insights for Policy Makers and Citizens http://www.sense4us.eu	University of Southampton	Making and implementing policy at any level of government is fraught with difficulty. The impact of decisions made are not always obvious at the time the policy is formulated or enacted, and any short-comings of the policy become known too late to change it. This is not due to a lack of information, it is due to the difficulty of finding and aggregating the right data out of the sea of information which characterises our modern world. Having once formulated a policy it is then impossible to make useful predictions around its likely impact and effectiveness. Policy specialists lack the resources and the methodology to be able to access most current data and are unable to take into account the views of citizens on policy issues expressed in real time through social network discussions. SENSE4US is creating an integrated package of utilities based on cutting-edge research that meets this need for tools and techniques to support information gathering, analysing and policy modelling in real time. Through close interaction with policy makers around Europe the project will validate results in complex policymaking settings and direct the research towards the support of more timely, more effective and better understood policy creation. The SENSE4US project will tackle these challenges of policymaking and implementation, integrating the benefits of both quantitative open data sources and qualitative social media data. We will provide tools enabling policy makers to find and select relevant information; link and homogenise the data; model policy in terms of constraints and intent; validate the policy; discover and incorporate views from NGOs and public; predict social impact of policy; provide decision support; provide understandable visualisation. The ultimate objective of the SENSE4US project is to advance policy modelling and simulation, data analytics and social network discussion dynamics, providing economic and social benefits at all governmental levels across Europe.
SIMPOL: Financial Systems Simulation and Policy Modelling http://www.simpolproject.eu	Universität Zürich	This project proceeds from the vision that fundamental advances of policy modelling in finance and climate finance can only stem from a genuinely interdisciplinary approach where network science, big data and ICT's meet economics and financial regulation. Accordingly, on the one hand we will develop new methods to assess the systemic importance of market players in (climate) financial networks and we will evaluate the effect of regulations in close collaborations with representatives of various regulatory bodies. The results will contribute to the discussion on how financial innovations could ignite a transition towards a greener economy and a more sustainable financial system. On the other hand, we will leverage on open data initiatives and semantic web to empower citizens with a more

Project	Organisation (Consortium lead partner)	Summary
		active role in relation to EU policies, by crowd-sourcing the mapping of the networks of influence involved in the policy making process.
SYMPHONY: Orchestrating Information Technologies and Global Systems Science for Policy Design and Regulation of a Resilient and Sustainable Global Economy http://www.projectsymphony.eu	Universita Degli Studi di Genova	Over the last few years, the greatest concern of the global political agenda has been to find a way to overcome the crisis and therefore to look for effective policy instruments and even for new theoretical economic frameworks. Some effort has been made in order to respond to this concern, but we still miss adequate tools for exploring and governing the complex global dynamics of our economic and social system. On the top of the urgency to find a way out of the crisis, and to prevent future crises by making the global economy more resilient, we need new tools to explore the possibility of broad societal transitions to sustainable patterns of production and consumption. In particular, it is crucial to explore financial sector regulations and macroeconomic policies that help to trigger these sustainability transitions. SYMPHONY aims at providing a set of innovative ICT tools, integrated in a platform designed to tackle these pressing issues. The work will be developed in strict cooperation with stakeholders and policy-makers, involved in the project to devise appropriate scenarios and modelling requirements. The main objective of the project is to develop a framework for designing and testing policies and regulatory measures regarding: • preventing and mitigating economic and financial crises; • fostering an economically and ecologically sustainable growth path. SYMPHONY's strategy to successfully accomplish its mission is to orchestrate a set of tools that will be able to: • collect and analyse relevant information by means of social media mining tools and web-based information markets; • simulate the complex economic dynamics by means of an agent based model of the global economy, explicitly designed for policymaking; • involve citizens in the decision making process through a serious game interface, and through a set of information markets on the artificial economy that will allow us to overcome the huge economic impasse of properly modelling expectations.

Technopolis, OII, CEPS (2015) based upon CORDIS and project websites

E.2 Other FP7 projects

A search for FP7 projects in the CORDIS project database[128] has resulted in a list of research projects in which innovative data-driven approaches for evidence-informed policymaking were used. The key words used are: big data, data policymaking and data policy. Several projects were found that are also part of the Global Systems Science platform; those projects are excluded from the table below. Summaries are taken from project websites or the CORDIS project database.

[128] http://cordis.europa.eu/search/advanced_en

Other relevant FP7 projects

Project	Organisation (Consortium lead partner)	Summary
+Spaces: Policy Simulation in Virtual Spaces http://cordis.europa.eu/project/rcn/93838_en.html	IBM Israel - Science and Technologt Ltd	Positive Spaces (+Spaces) aims to provide novel technologies and instruments that will allow government bodies to measure public opinion on a large scale and seize the outcomes of prospective policy measures by leveraging the power of the communities. From the +Spaces perspective, Virtual Spaces (VS) form a vast reserve of group knowledge that, once aggregated and properly processed, can be incorporated in the policy making process. Currently, government agencies seek to refine their policy making process by assessing the impact of prospective policies to the society prior to their implementation. Traditional tools for this include polls and forums, but these have limited capabilities and are inadequate for purpose variety of reasons. +Spaces aims to address this issue by using VSs as testbeds for policy simulation. Contemporary VSs can be viewed as micro-societies with dynamics resembling those of real world societies, the most evolved having virtual economies as well as regulations analogous to a legislative framework. Moreover, VSs are controlled environments in which all parameters of users' reactions and interactions can be tracked. +Spaces will provide governments with tools for testing their legislation in a number of virtual environments as well as data mining tools for tracking the users' reactions. We will develop technologies for aggregating, filtering and analysing this information, as well as recommender and reputation systems that will ensure that only reliable sources of data are taken into account. Finally we will represent this information in an appropriate fashion that will enable government bodies to draw conclusions on the potential outcomes of the policies they propose, taking into account the analogies between VS and real life concepts. We believe that the +Spaces platform will constitute a powerful tool for testing policies and forecasting their impact and will assist government bodies to include the public opinion in the policy making process.
Apps4EU: Turning Data into Business http://www.appsforeurope.eu	Stichting Waag Society	Apps for Europe - turning Data into Business will create a thematic network to organise competitions for using open data, stimulate the winners to start business ventures, and maximise the socio-economic impact and overall benefits of open data. It will contribute to the open data policy of the Commission and stimulate the reuse of public sector information from governmental, scientific and cultural sources. It brings together 19 partners from ten countries that have a background in hosting and facilitating open data competitions, providing data and supporting business ventures. Together, we will build on and align 20 existing local and national open data competitions held over the next two years by the partners and stimulate the participants to turn their inventions into viable businesses. The project will develop a new, transferable and tested programme that will increase the business knowledge and potential for success of participants in open data competitions: the Business Lounge. The Business Lounge educates developers in creating and commercialising apps with market potential. Participants enter their prototype app, including a business model into an annual EU Open Data Competition. We will organise two such pan-European competitions that bring the

Project	Organisation (Consortium lead partner)	Summary
		best of the local and national competitions into a European limelight and introduce them to a business network that will increase their potential. The categories of this EU open data competition are: Government, Science and Culture. The best local and EU contestants are invited to pitch their plans during a Pitch & Prize event for investors, incubators, international companies, entrepreneurial non-profits and social investors. Additionally, the project spawns five support actions on themes that are identified as key areas of importance to organising successful open data competitions: Business and Policy; Data and Licensing; Media exposure for start-ups; Technical tooling and Community Building.
BIG: Big Data Public Private Forum http://www.big-project.eu	Atos Spain Corporation	Big data is an emerging field where innovative technology offers alternatives to resolve the inherent problems that appear when working with huge amounts of data, providing new ways to reuse and extract value from information. Three main dimensions characterize big data: huge variety of data format, often time-sensitive and large. Big data offers tremendous untapped potential value for many sectors but no specific intelligent-large-data-handling/brokering industrial sector exists. Furthermore, from an industrial adoption point of view, Europe is lagging behind US in big data technologies. A clear strategy to align supply and demand is needed as a way of increasing competitiveness of European industries. Building an industrial community around big data in Europe will be the priority of this project, together with setting up the necessary collaboration and dissemination infrastructure to link technology suppliers, integrators and leading user organizations. Big Data Public Private Forum (BIG) will work towards the definition and implementation of a clear strategy that tackles the necessary efforts in terms of research and innovation, but it will also provide a major boost for technology adoption and supporting actions from the European Commission in the successful implementation of the big data economy. As part of this strategy, outcomes of this project will be used as input for Horizon 2020 and will be sustained beyond the project duration. Foundational research technologies will be analysed and assessed in BIG and technology and strategy roadmaps created, so that the business and operational communities understand the potential of these technologies and are enabled to implement appropriate strategies and technologies for commercial benefit. To maximize the success of this initiative BIG has selected a balanced set of partners representing Academia and specially Industry.
BigFoot: Big Data Analytics of Digital Footprints http://bigfootproject.eu	EURECOM	The amount of digital information in our world has been exploding and new technologies and services will continue to fuel exponential growth of large pools of data that can be captured, stored, and analysed. Nowadays, however, tools and services to store, process and interact with data are still in their infancy, represented by scattered solutions that fall short in having a unified vision, that lack common interfaces, and that only offer best-effort services. The aim of BigFoot is to overcome current drawbacks by designing, implementing and evaluating a Platform-as-a-Service solution for processing and interacting with large

Project	Organisation (Consortium lead partner)	Summary
		volumes of data. The BigFoot stack – which builds upon and contributes to the Apache Hadoop ecosystem and the OpenStack project, in addition to creating new open source components – features automatic and self-tuned deployments of storage and processing services for private clouds, going beyond best-effort services currently available in the state-of-the-art. BigFoot takes a novel, cross-layer approach to system optimization, which is evaluated with a thorough experimental methodology using realistic workloads and datasets from two representative applications, namely ICT Security and Smart Grid data analytics. In addition, BigFoot aims at making data interaction easy by supporting high-level languages (for batch oriented analytic tasks) and by taking a service-oriented approach to support and optimize latency sensitive queries.
BYTE: Big data roadmap and cross-disciplinarY community for addressing socieTal Externalities http://byte-project.eu	Trilateral Research & Consulting, LLP	The Big data roadmap and cross-disciplinarY community for addressing socieTal Externalities (BYTE) project will assist European science and industry in capturing the positive externalities and diminishing the negative externalities associated with big data in order to gain a greater share of the big data market by 2020. BYTE will accomplish this by leveraging the BYTE advisory board and additional network contacts to conduct a series of big data case studies in actual big data practices across a range of disciplinary and industrial sectors to gain an understanding of the economic, legal, social, ethical and political externalities that are in evidence. BYTE will supplement these case studies with a horizontal analysis that identifies how positive externalities can be amplified and negative externalities can be diminished. YTE moves beyond current practices to consider how big data will develop to the year 2020 using foresight tools to identify future practices, applications and positive and negative externalities. This will allow BYTE to develop, in collaboration with expert stakeholders, a vision for big data in 2020 that includes meeting the relevant goals of the Digital Agenda for Europe. In collaboration with expert stakeholders, the consortium will then devise a research and policy roadmap that will provide incremental steps necessary to achieve the BYTE vision and guidelines to assist industry and scientists to address externalities in order to improve innovation and competitiveness. BYTE will culminate in the launch of the big data community, a sustainable, cross-disciplinary platform that will implement the roadmap and assist stakeholders in identifying and meeting big data challenges. Furthermore, BYTE will disseminate project findings and recommendations and publicise the big data community to a large population of stakeholders to encourage further innovation and economic competitiveness in Europe's engagement with big data.
CROSSOVER: Bridging Communities for Next Generation Policy-Making http://crossover-	Athens Technology Center SA	In an increasingly interdependent, unstable and complex world, as the financial crisis has shown, policy-makers lack the tools to anticipate emerging problems and the impact of possible solutions. At the same time, citizens are becoming more vocal in monitoring and influencing policy decisions. Emerging ICT tools for governance and policy modelling show great opportunities for decision-making in a complex

Project	Organisation (Consortium lead partner)	Summary
project.eu		world, through the dramatic and combined growth of data availability, analysis and simulation tools, participative and behavioural change technologies. However, they still remain at the margins of policymaking: research and deployment are fragmented between academic fields, policy domains, geographical areas, technological layers and type of stakeholders. As a result, resources are spent in «reinventing the wheel» and reach sub-optimal efficiency and effectiveness. The CROSSOVER project aims to consolidate and expand the existing community (built largely within FP7) by: • bringing together and reinforcing the links between the different global communities of researchers and experts: it will create directories of experts and solutions, and animate knowledge exchange across communities of practice both offline and online; • reaching out and raising the awareness of non-experts and potential users, with special regard to high-level policy-makers and policy advisors: it will produce multimedia content, detailed case studies, roadmap, workshops and high-level policy conference with competition for prize; • establishing the scientific and political basis for long-lasting interest and commitment to next generation policymaking, beyond the mere availability of FP7 funding: it will focus on use cases and a demand-driven approach, involving policy-makers and advisors in high-level conference It will pursue this goal through a combination of content production, ad hoc and well-designed online and offline animation; as well as strong links with existing communities outside the CROSSOVER project and outside the realm of eGovernment.
CROWD-LAND: Harnessing the power of crowdsourcing to improve land cover and land-use information http://erc.europa.eu/harnessing-power-crowdsourcing-improve-land-cover-and-land-use-information	Internationales Institut für Angewandte Systemanalyse	Information about land cover, land use and the change over time is used for a wide range of applications such as nature protection and biodiversity, forest and water management, urban and transport planning, natural hazard prevention and mitigation, agricultural policies and monitoring climate change. Furthermore, high quality spatially explicit information on land cover change is an essential input variable to land use change modelling, which is increasingly being used to better understand the potential impact of certain policies. The amount of observed land cover change also serves as an important indicator of how well different regional, national and European policies have been implemented. However, outside Europe and outside the developed world in particular, information on land cover and land cover change in poorer countries is hardly available and no national or regional dense sample based monitoring approaches such as LUCAS exists which deliver sufficiently accurate land cover and land cover change information. Moreover in particular in developing countries, there is no or very little information on land-use and crop management. Only very limited data available from FAO and an incomplete coverage of sub-national statistics (e.g. IFPRI) are available. This research project will assess the potential of using crowdsourcing to close these big data gaps in developing and developed countries with a number of case studies and different data collection methods. The CrowdLand project

Project	Organisation (Consortium lead partner)	Summary
		will be carried out in two very different environments, i.e. Austria and Kenya. The overall research objectives of this project are to 1) test the potential of using social gaming to collect land use information 2) test the potential of using mobile money to collect data in developing countries 3) understand the data quality collected via crowdsourcing 4) apply advanced methods to filter crowd sourced data in order to attain improved accuracy.
DAIAD: Open Water Management - from droplets of participation to streams of knowledge http://daiad.eu	Athena Research and Innovation Center	Efficient water management is a challenging issue with the potential to affect the long-term well-being, economy and security of society. Policies for sustainable water management have been established in the EU. However measures to support efficient water use for citizens are currently lacking. Consumers have limited means to accurately monitor their water consumption and thus stimuli to modify their behaviour towards a sustainable lifestyle. A potentially ground-breaking approach for efficient water use and reuse lies within the empowerment of consumers. The principles of open knowledge and participation have provided solutions and driven innovation in similar challenging and complex issues. We believe that a similar bottom up method, in which citizens can voluntarily adopt low cost water monitoring services, self-induce behavioural changes, and accordingly demand better services, can be a catalyst for large-scale changes in efficient water management. The DAIAD project constitutes an innovative approach for addressing the challenge of efficient water management through real-time knowledge of residential water consumption, bringing together leading members of the water and ICT domains. Our goal is to research and develop innovative low cost, inclusive technologies for real-time, high granularity water monitoring and knowledge extraction. We will devise multi-modal feedback interfaces, recommendation, and analysis services to communicate knowledge and incur behavioural changes to consumers in residential settings. We will apply big data management and analysis technologies to provide efficient management and analysis of real-time water consumption data, as well as multiple relevant data sources. This will enable water stakeholders to gain novel insight and explore the hidden correlations of the parameters that shape water demand strategies and water pricing, thus leading to more efficient water management.
DATA SIM: Data Science for Simulating the Era of Electric Vehicles http://www.uhasselt.be/UH/datasim/WOG-Home.html	UHasselt-IMOB	DATA SIM aims at providing an entirely new and highly detailed spatio-temporal micro simulation methodology for human mobility, grounded on massive amounts of big data of various types and from various sources, with the goal to forecast the nation-wide consequences of a massive switch to electric vehicles, given the intertwined nature of mobility and power distribution networks. While the increasing availability of big data about human activities provides radical new ways of understanding the social and ecological universe, it is our ambition in this project to complement this information with behaviourally rich data, pertaining to the purpose of human travels. In terms of interdependencies, our advanced integrated methodological environment allows for more realistic and consistent linkages across travel choices made by the

Project	Organisation (Consortium lead partner)	Summary
		individuals in the course of a day than conventional models, with the goal of simulating tens of millions of individual agents, each with their detailed prediction of every activity-travel schedule, enabling more detailed segmentations based on user profiles of the agents, different activity types, trip duration and driving ranges. Significant breakthroughs can be gained from the project, which lead to novel dimensions of use, along the milestones that were set forward in the European Industry Roadmap for the Electrification of Road Transport from today till 2020. Many scientists have already pointed out that the goal of social sciences is not simply to understand how people behave in large groups, but to understand what motivates individuals to behave the way they do. This fundamental insight, which can be gained from this project, is a step forward towards the solution of this important challenge,, and it can help us to better understand the dynamics of our society and, in the longer run, to have an impact on overall and wider societal well-being.
ePolicy: Engineering the POlicy-making LIfe Cycle http://www.epolicy-project.eu/node	Alma Mater Studiorum-Universita di Bologna	The e-POLICY project presents a multi-disciplinary research effort whose goal is to support policy makers in their decision process across the engineering of a policymaking life-cycle. The life-cycle integrates, in a unique way, global and individual perspectives on the decision process, bringing to the policy maker's attention both global concerns (e.g., impacts, budget constraints and objectives), and individual concerns (i.e. opinions, reactions), giving guidance towards better policy implementation strategies. An innovative game theoretical approach guides the interaction and conflict management between these two levels. In addition, the e-POLICY project proposes the evaluation of the economical, societal and environmental impacts of policy at both the global and individual levels. Societal impacts are derived from data retrieved by e-participation and Web 2.0 tools. Both policy maker and citizens are assisted in the decision-making and participation processes through visualisation tools. From a technological perspective, leading-edge optimization and decision support techniques contribute to global policymaking, while agent-based simulation tools will perform a social simulation at the individual level, with game theory being used to manage conflicts and regulate the interaction between the two levels. Subsidiary, but no less important, techniques will aid the decision-making process by using opinion-mining from e-participation derived data and advanced visualisation tools. The project proof of concept will be an open source decision support system where the above components are implemented, integrated and evaluated with a realistic target: that of a real-world regional energy plan. One of the key barriers to the progress of renewable energy projects is recognized as the planning and consenting process. The e-POLICY project goes towards resolving this issue by aiding both processes.
IMPART: Intelligent Management Platform for Advanced Real-Time	Universitat Pompeu Fabra	IMPART will research, develop and evaluate intelligent information management solutions for 'big data' problems in the field of digital cinema production. It will find new ways of managing, visualising and analysing very large multimodal data sets so that creative personnel can review three-dimensional scene representations on the set, understand

Project	Organisation (Consortium lead partner)	Summary
media processes http://impart.upf.edu		the data, identify errors, evaluate the quality of the shot and take creative decisions in real time. We will research and test solutions for the use case of large-scale movie production, with very large research and evaluation datasets based on the latest industrial processes. Results will be applicable to film production and more generally to commercial and professional domains ranging from TV and on-line video to surveillance. The project outcomes will include: • Architectures and file structures for intelligent multimodal data and metadata management, presentation and browsing linking data, metadata and advanced applications • Methods for automatically selecting relevant subsets of data to allow rapid processing • Techniques for processing, information extraction and analysis in distributed systems • Multimodal content semantic and low-level analysis techniques for the automatic derivation of metadata to support visualisation, decision support, management and retrieval • On-set decision support, including tools for multimodal data visualisation, data summarization, layout, registration and cross-referencing, data and metadata search, virtual overlays and real-time previews of effects, creative and technical decision support • A demonstrator system for professional testing and evaluation, both on and off set • A documented xml interface that allows third parties to develop compatible tools and processes • A published test dataset of multimodal data and metadata captured in realistic film production conditions for use by other researchers.
IPaCSO: Innovation Framework for Privacy and Cyber Security Market Opportunities http://ipacso.eu	Waterford Institute of Technology (WIT)	Innovation drives new product realisation and development. Significant opportunities exist for innovation in the privacy and cyber security (PACS) technology space, yet complex market, regulatory, policy, commercial, and economic considerations create several barriers to transforming research outputs into market-centric product and service applications. In response, Innovation Framework for Privacy and Cyber Security Market Opportunities (IPaCSO) will develop a structured knowledge and decision-support innovation framework for identifying, assessing and exploiting market opportunities in the privacy and cyber security technology space. IPaCSO will support security innovators, policy makers and research spectrum stakeholders in identifying, assessing and exploiting new ideas and research assets using innovation and market assessment best-practice and guidelines and bringing them to market. Achieving project impact objectives can be challenging for ICT security research projects thus through the adoption and utilisation of the IPaCSO innovation framework we propose that projects can increase the impact of their project results. IPaCSO will address the following main goals: • Assess existing innovation processes used in the PACS domain via in-depth stakeholder engagement • Identifiy a set of innovation framework requirements,

Project	Organisation (Consortium lead partner)	Summary
		interleaving improved innovation practices and case study scenarios, that support PACs domain concerns • Assess existing economic barriers to innovation and identify appropriate economic incentives needed to increase security product and service adoption • Develop an appropriate knowledgebase and decision support approach that is transferrable to PACS technologies exploiting potential market opportunities. • Develop effective training, exploitation and dissemination of the resultant IPaCSO framework to target stakeholder groups, both during and beyond the project lifecycle. The proposed work will focus on key thematic areas within the PACs domain, for example, security concerns across different OSI layers, emerging mobile and telecoms security, security monitoring and incident response, emerging notions of privacy and identity, embedded security, and emerging managed security services models. Such concerns will be considered in the context of wider ICT factors such as cloud computing, big data, emerging internets of services and things, and similar IT macro factors. IPaCSO project activities will lead to a structured understanding of such thematic market areas, a clearer understanding of how existing market barriers can be addressed within innovation frameworks, providing practical support for PACs stakeholders in producing improved market focussed outputs in the future.
LeanBigData: Ultra-Scalable and Ultra-Efficient Integrated and Visual Big Data Analytics http://leanbigdata.eu	Universidad Politecnica de Madrid	LeanBigData aims at addressing three open challenges in big data analytics: 1) The cost, in terms of resources, of scaling big data analytics for streaming and static data sources; 2) The lack of integration of existing big data management technologies and their high response time; 3) The insufficient end-user support leading to extremely lengthy big data analysis cycles. LeanBigData will address these challenges by: • Architecting and developing three resource-efficient big data management systems typically involved in big data processing: a novel transactional NoSQL key-value data store, a distributed complex event processing (CEP) system, and a distributed SQL query engine. We will achieve at least one order of magnitude in efficiency by removing overheads at all levels of the big-data analytics stack and we will take into account technology trends in multicore technologies and non-volatile memories. • Providing an integrated big data platform with these three main technologies used for big data, NoSQL, SQL, and Streaming/CEP that will improve response time for unified analytics over multiple sources and large amounts of data avoiding the inefficiencies and delays introduced by existing extract-transfer-load approaches. To achieve this we will use fine-grain intra-query and intra-operator parallelism that will lead to sub-second response times. • Supporting an end-to-end big data analytics solution removing the four main sources of delays in data analysis cycles by using: 1) automated discovery of anomalies and root cause analysis; 2) incremental visualisation of long analytical queries; 3) drag-and-drop declarative composition of visualisations; and 4) efficient manipulation of visualisations through hand gestures over 3D/holographic views. Finally, LeanBigData will demonstrate these results in a cluster with 1,000 cores in four real industrial use cases with real data, paving the way for deployment in the context of realistic

Project	Organisation (Consortium lead partner)	Summary
		business processes.
LIAISE: Linking Impact Assessment Instruments to Sustainability Expertise http://www.liaise-kit.eu	Stichting Dienst Landbouwkundig Onderzoek	Existing research points out that the full potential of Impact Assessment (IA) for delivering sustainable development is not being realised. Many tools to support IA are not yet being fully employed by policy makers. This is symptomatic of a large and deep gap between the two broad communities of IA researchers and IA practitioners. Practitioners tend to look for tools that are simple and transparent while the researchers are more interested in the sophistication and innovative aspects of assessment tools. The main purpose of LIAISE is to identify and exploit opportunities to bridge between these two communities in a way that leads to an enhanced use of IA tools in policymaking. Its centrepiece will be a shared toolbox simultaneously accessible and useful for policy makers as well as for the research community. The LIAISE consortium will: Unite the multi-disciplinary competences of a core of large European institutes, that in turn consolidate the expertise from large FP6 projects. Combine researchers that analyse current policy needs and link them in innovative ways to those who maintain and develop IA tools; Develop a roadmap towards a virtual centre of excellence on IA, that can operate as the durable hub of existing academic and practitioner networks relevant to the themes of the NoE; Maintain the flexibility to support real life IA processes, informed by a structured dialogue with the IA user community; Develop a business plan to ensure durable operation, scientific credibility and efficient usability of the shared toolbox, also in the post-project period. Resulting in: A shared toolbox: a durable and flexible infrastructure to support IA with improved tools; A continuously updated shared research agenda; Capacity building and training components to spread the results of research activities to target groups in communities of IA users and IA researchers; A virtual centre of excellence on IA.
LinkedUp: LinkedUp: Linking Web Data for Education Project – Open Challenge in Web-scale Data Integration http://linkedup-project.eu	Leibniz Universität Hannover	LinkedUp aims to push forward the exploitation of the vast amounts of public, open data available on the Web, in particular by educational institutions and organizations. This will be achieved by identifying and supporting highly innovative large-scale Web information management applications through an open competition (the LinkedUp Challenge) and dedicated evaluation framework. The vision of the LinkedUp Challenge is to realise personalised university degree-level education of global impact based on open Web data and information. Drawing on the diversity of Web information relevant to education, ranging from Open Educational Resources metadata to the vast body of knowledge offered by the Linked Data approach (31 Billion RDF statements as part of the Linked Open Data cloud alone), this aim requires overcoming substantial challenges related to Web-scale data and information management involving big data, such as performance and scalability, interoperability, multilinguality and heterogeneity problems, to offer personalised and accessible education services. Therefore, the LinkedUp Challenge provides a focused scenario to derive challenging requirements, evaluation criteria, benchmarks and thresholds which are reflected in the LinkedUp evaluation

Project	Organisation (Consortium lead partner)	Summary
		framework. Information management solutions have to apply data and learning analytics methods to provide highly personalised and context-aware views on heterogeneous Web data. Building on the strong alliance of institutions with expertise in areas such as open Web data management, data integration and Web-based education, key outcomes of LinkedUp include a general-purpose evaluation framework for Web-data driven applications, a set of quality-assured educational datasets, innovative applications of large-scale Web information management, community-building and clustering crossing public and private sectors and substantial technology transfer of highly innovative Web information management technologies.
Live+Gov: Reality Sensing, Mining and Augmentation for Mobile Citizen–eGovernment Dialogue http://liveandgov.eu	Universität Koblenz	Bringing policy makers closer to real life of citizens is the objective of Live+Gov. We will develop a mobile government solution that allows citizens to accurately express their needs to government by using a variety of mobile sensing technologies available in their smartphones (GPS; image/audio recording; compass, orientation sensors, etc.) next to established means of mobile eParticipation such as textual input and output. Sensing and mining data from the real world, such as geo-located images of street damages or GPS track recordings during traffic disruption, government achieves accurate analyses of citizens' needs and opinions and can take advantage of a collective citizen intelligence. Live+Gov develops novel policy models and links data contributions by citizens to existing administration processes. The Live+Gov key feature is that it fully exploits the capabilities of widely proliferated smartphones for mobile eParticipation, rendering an augmented reality of governmental actions and plans, such that citizens obtain fast and comprehensive feedback on issues they encounter. Anonymization, data randomization and access control will safeguard against misuse and ensure privacy for all involved. Live+Gov will be delivered in a Software-as-a-Service model packaging its fully developed software components such that take-up is possible at dramatically lower cost of operations. Success will be measured through three live field trials (mobility, urban maintenance, urban planning) conducted by partners already operating end-user platforms in these domains and reusing existing governmental data. The methodology, solution, and experiences made in field trials will be published through a combination of open source, public reports and open workshops. Field trials and exploitation is supported by a committed advisory board from four governmental institutions ensuring widespread take-up and long-term sustainability of the project results.
MAPPING: Managing Alternatives for Privacy, Property and INternet Governance http://www.mappingtheinternet.eu	Rijksuniversiteit Groningen	Building on the results of several EU FP7 projects including CONSENT (covering on-line consent and privacy in social networks), SMART and RESPECT (which cover smart and on-line surveillance, etc.) MAPPING's goal is to create an all-round and "joined-up" understanding of the many and varied economic, social, legal and ethical aspects of the recent developments on the internet, and their consequences for the individual and society at large. MAPPING would specifically investigate and debate the existing innovation policies, business models and legal framework related to the implementation of the Digital

Project	Organisation (Consortium lead partner)	Summary
		Agenda for Europe and the changes needed to set up an improved governance structure for the EU innovation ecosystem. The key to MAPPING's success would be its planned mobilisation of a wide spectrum of ICT-related stakeholders and social actors from both EU Member States and associated countries, including academics, law and policy makers, ISPs, international and EU internet governance bodies, NGOs and civil society organisations. The project would provide these actors with a forum for informed discussion of issues related to the digital transition, such as problems of personal data and IPR protection online, business models and e-government applications based on the use of personal data, economic exploitation of IPRs and open innovation. MAPPING would then move to create an Action Plan and put forward workable policy guidelines based on a multidisciplinary perspective on the latest and foreseeable developments in ICTs taking into account conflicting interests, perceptions and practices of different societal actors that shape the EU's technological future. MAPPING would thus significantly contribute to creating an enabling framework for completing the digital transition and improving the innovation climate in the EU.
MOSIPS: Modeling and Simulation of the Impact of Public Policies on SMEs http://www.mosips.eu	Anova IT Consulting	The aim of the project is to develop a user-friendly policy simulation system allowing forecasting and visualizing the socio-economic potential impact of public policies. This will allow policy makers to make experiments with different socio-economic designs, with the participation of citizens and potentially impacted stakeholders, before a public policy is settled. It will also allow the different stakeholders to participate into the process and better understand different scenarios possible and their side effects. Focus will be set on SME-oriented policies in order to manage the scope of R&D activities. They have been selected given the importance of SMEs in EU Economy, and the R&D potential given by the nature of policies and impacts addressing this collective. Combining suitable data, models, artificial intelligence and interactive visualisation tools, the final goal is to develop a "policy wind tunnel". MOSIPS system will be suited to craft policy options, giving social and economic stakeholders a decision arena visualizing and illustrating policy insights, providing valuable decision-support, as it will make counter-intuitive feedback, cascading and side effects understandable. MOSIPS will be able to significantly improve the quality of policies implemented by public authorities at any administrative level, and at the same time, facilitate access to meaningful information to citizens about the expected consequences of policies designed for the region/locality/country/area they live at, promoting their engagement and active participation in the policymaking and implementation process. Concrete S&T objectives are related to the consecution of this project aim, mainly dealing with the areas of modelling, multi-agent based intelligent and dynamic simulation, visual technologies for data analysis and real-time interaction with simulation engines, citizens' empowerment, software certification and the paradigm of Software as a Service and related ICT areas.

Project	Organisation (Consortium lead partner)	Summary
NOMAD: Policy Formulation and Validation through non moderated crowdsourcing http://nomad-project.eu	Panepistimio Aigaiou	Collaboration and crowdsourcing are the realities of today's public internet. The so-called "Web 2.0" represents a precious repository of thematic information, thanks to the heterogeneous content that is inserted daily and spontaneously updated by its users. Very recently, a big commercial interest has started to arise – especially within industries that manufacture consumer goods and services – in acquiring, classifying and managing all product related information that emerges out of Web 2.0 channels, thus going beyond the known capabilities of consolidated search engines. It would be possible to use this insight to information at multiple stages of the policy-life cycle to support the definition of the political agenda, the creation, the implementation and the monitoring of policy proposals. In this context, modern politicians could test, detect and understand how citizens perceive their own political agendas, and also stimulate the emergence of discussions and contributions on the informal web (e.g. forums, social networks, blogs, newsgroups and wikis), so as to gather useful feedback for immediate (re)action. In this way, politicians can create a stable feedback loop between information gathered on the Web and the definition of their political agendas based on this contribution. The ability to leverage the vast amount of user-generated content for supporting governments in their political decisions requires new ICT tools that will be able to analyze and classify the opinions expressed on the informal Web, or stimulate responses, as well as to put data from sources as diverse as blogs, online opinion polls and government reports to an effective use. To this end, NOMAD aims to introduce these different new dimensions into the experience of policymaking by providing decision-makers with fully automated solutions for content search, selection, acquisition, categorisation and visualisation that work in a collaborative form in the policy making arena.
OpenCube: Publishing and Enriching Linked Open Statistical Data for the Development of Data Analytics and Enhanced Visualization Services http://opencube-project.eu	Ethniko Kentro Erevnas Kai Technologikis Anaptyxis	The ultimate goal of OpenCube project is to facilitate (a) publishing of high-quality linked statistical data, and (b) reusing distributed linked statistical datasets to perform advanced data analytics and visualisations. Towards this end, OpenCube project will develop open source software tools in terms of (a) standalone applications and (b) extensions of two linked data management platforms, namely Swirrl's PublishMyData and fluidOps' Information Workbench, for publishing and reusing high-quality linked statistical data. Although the Open Government Data (OGD) movement promised to provide a number of benefits, recent studies show that its full potential has not yet realized. The difficulty in exploiting open data seems surprising if we consider the huge importance data have in modern societies e.g. in business intelligence, research models and evidence based policymaking. The main tenet of OpenCube project is that the real value of OGD will unveil from performing data analytics and advanced visualisations on top of open statistical data that were previously resided in disparate sources and can now be combined to provide unexpected and unexplored insights. However, statistical data need to be formulated and integrated as linked Data Cubes characterized by

Project	Organisation (Consortium lead partner)	Summary
		dimensions, slices and measures in order to unveil their full potential and value. Towards this end, the OpenCube project aims at developing software tools to address the particular requirements of Data Cubes, thus enabling the efficient and easy publishing and reuse of open statistical data. In particular the OpenCube project will support (a) public bodies to publish linked statistical data of high quality in an easy way and (b) European SMEs to combine linked open statistical data with other data (e.g. enterprise's own data) and perform advanced data analytics and visualisations in a cost effective manner. The OpenCube foreseen results include: • The OpenCube toolkit comprising standalone tools for publishing, visualizing and performing data analytics on top of linked statistical data. • The OpenCube extension of the PublishMyData platform. • The OpenCube extension of the Information Workbench platform. • The OpenCube Pilots in three public authorities and businesses to validate and prove the usability and effectiveness of the developed tools and guidelines.
Optique: Scalable End-user Access to Big Data http://www.optique-project.eu	Universitetet i Oslo	Scalable end-user access to big data is critical for effective data analysis and value creation. Optique will bring about a paradigm shift for data access by • providing a semantic end-to-end connection between users and data sources • enabling users to rapidly formulate intuitive queries using familiar vocabularies and conceptualisations • seamlessly integrating data spread across multiple distributed data sources, including streaming sources • exploiting massive parallelism for scalability far beyond traditional RDBMSs and thus reducing the turnaround time for information requests to minutes rather than days. These objectives will be achieved by bringing together leading experts and cutting edge technology from diverse communities to develop an extensible platform that will provide a complete and generic solution to the data access challenges posed by big data. The platform will: (i) Use an ontology and declarative mappings to capture user conceptualisations and to transform user queries into complete, correct and highly optimised queries over the data sources; (ii) Integrate distributed heterogeneous sources, including streams; (iii) Exploit massively parallel technologies and holistic optimisations to maximise performance; (iv) Include tools to support query formulation and ontology and mapping management; (v) Use semi-automatic bootstrapping of ontologies and mappings and query driven ontology construction to minimise installation overhead. Development of the platform will be informed by and continuously evaluated against the requirements of complex real-world challenges, with two large European companies providing the project with comprehensive use cases, and access to user groups and TB scale data sets. Experience from the use case deployments will be used to develop high quality tutoring and training resources, and to engage in an aggressive dissemination and exploitation program aimed at achieving the widest possible uptake of Optique technology.

Project	Organisation (Consortium lead partner)	Summary
PHEME: Computing Veracity Across Media, Languages, and Social Networks http://www.pheme.eu	The University of Sheffield	Social media poses three major computational challenges, dubbed by Gartner the 3Vs of big data: volume, velocity, and variety. Content analytics methods have faced additional difficulties, arising from the short, noisy, and strongly contextualised nature of social media. In order to address the 3Vs of social media, new language technologies have emerged, e.g. using locality sensitive hashing to detect breaking news stories from media streams (volume), predicting stock market movements from microblog sentiment (velocity), and recommending blogs and news articles based on user content (variety). PHEME will focus on a fourth crucial, but hitherto largely unstudied, challenge: veracity. It will model, identify, and verify phemes (internet memes with added truthfulness or deception), as they spread across media, languages, and social networks. PHEME will achieve this by developing novel cross-disciplinary social semantic methods, combining document semantics, a priori large-scale world knowledge (e.g. Linked Open Data) and a posteriori knowledge and context from social networks, cross-media links and spatio-temporal metadata. Key novel contributions are dealing with multiple truths, reasoning about rumour and the temporal validity of facts, and building longitudinal models of users, influence, and trust. Results will be validated in two high-profile case studies: healthcare and digital journalism. The techniques will be generic with many business applications, e.g. brand and reputation management, customer relationship management, semantic search and knowledge management. In addition to its high commercial relevance, PHEME will also benefit society and citizens by enabling government organisations to keep track of and react to rumours spreading online. PHEME addresses Objective ICT-2013.4.1 Content analytics and language technologies; a) cross-media analytics.
RETHINK big: Roadmap for European Technologies in Hardware and Networking for Big Data http://www.rethinkbig-project.eu	Barcelona Supercomputing Centre	The objective of the RETHINK big project is to bring together the key European hardware, networking, and system architects with the key producers and consumers of big data to identify the industry coordination points that will maximize European competitiveness in the processing and analysis of big data over the next 10 years. Specifically, RETHINK big will deliver a strategic roadmap for how technology advancements in hardware and networking can be exploited for the purpose of data analytics while also taking into consideration advancements in applications, algorithms and systems. In this project, we will not carry out actual research on Hardware optimizations for big data, but rather survey the landscape of opportunities and propose a strategic roadmap from that landscape. The outcome of the project will be a series of detailed (mentioning specific technologies), realistic (considering our 10-year timeline), verifiable (including target metrics) and coordinated technology development recommendations that would be in the best interest of European big data companies to undertake in concert as a matter of competitive advantage. Practically speaking, the roadmap will be produced as a result of area specific and cross-functional working groups meetings and congresses. We will initially identify and evaluate the existing competencies across European big data application domains and technology providers in Europe and then identify the key European stakeholders, or the

Project	Organisation (Consortium lead partner)	Summary
		established and up-and-coming institutions that possess or are developing the technologies, processes or services that map to these competencies. From these stakeholder institutions, we will select technology and business experts that will chart the technological advancements, their respective challenges and the potential business opportunities that they present. These experts will not only share an interest in defining a credible roadmap, but also hold the decision-making power within their respective institutions (and collectively) to implement that roadmap. At the highest level, this project will identify and evaluate the existing competencies across European big data Hardware and Networking technology sectors and application domains and prioritize the complementary interests and the shared opportunities that allow all key industrial stakeholder companies to unlock the highest return on their respective investments; it will result in a roadmap that would be irrational not to follow.
SMART: Sustainable Mobility and Robust Transportation http://smart-eu.org	Technion Israel Institute of Technology	Traffic Congestion has several effects on travellers, businesses, agencies, and cities. Congestion costs Europe about 1% of Gross Domestic Product (GDP) every year. Transportation networks are complex and large-scale dynamic systems with uncertainties. Efficient traffic control and management of large-scale transportation networks still remain a challenge both for traffic researchers and practitioners. Our main objective in this project is to develop smart transportation systems for large-scale traffic networks that aim at minimizing network delays and increasing sustainable mobility and accessibility, with the help of feeded data from advanced sensor and information technologies. To achieve this goal, we will develop a new interdisciplinary research approach that combines and fuses research areas developed in different fields in traffic theory, control theory, optimal control, optimization, big data mining and fusing, and estimation theory. The SMART project is unique in several aspects, as it is an interdisciplinary research and it is expected to make advances in the state of the art in different directions: (i) Developing a hierarchical scheme of integrated modelling and control for the intersection and network levels. (ii) Developing efficient control strategies that are robust under uncertainties. (iii) Developing implementable real time control based on feeded information for both levels including the monitoring and estimation procedures. (iv) A real field study implementation. We believe that the new interdisciplinary approach will result in a better understanding of complex and congested cities, and will allow integrating traffic control and management approaches to generate smart cities that are more livable and sustainable. The operational deployment of such smart networks would improve the life quality in Europe by improving traffic conditions in cities, reducing travel times and minimizing delays, and reducing pollution levels.
SPEEDD: Scalable ProactivE Event-Driven Decision making http://speedd-project.eu	National Center for Scientific Research "Demokritos"	SPEEDD (Scalable ProactivE Event-Driven Decision making) will develop a system for proactive event-based decision-making: decisions will be triggered by forecasting events - whether they correspond to problems or opportunities- instead of reacting to them once they happen. The decisions and actions will be real-time, in the sense that they will be taken under tight time constraints, and require on-the-fly

Project	Organisation (Consortium lead partner)	Summary
		processing of "big data", i.e. extremely large amounts of noisy data storming from different geographical locations as well as historical data. The effectiveness of the SPEEDD solution will be evaluated against concrete requirements in traffic management and credit card fraud detection. SPEEDD will contribute to the state of the art by:-Developing techniques for on-the-fly, low-latency, scalable, distributed monitoring and online distributed learning, given extremely large, geographically distributed, noisy event streams, and highly complex event patterns.-Developing novel methods for real-time event recognition and forecasting that are resilient to various types of uncertainty and supported by novel machine learning techniques for continuously improving recognition and forecasting accuracy.-Providing novel methods for event-based, real-time decision-making under uncertainty.-Developing techniques for real-time explanation and visualisation of big data. The expected outcome of SPEEDD includes: • Technology for real-time event recognition and forecasting under uncertainty. • Technology for event-based, real-time decision-making under uncertainty. • Visual analytics suite for real-time interaction with, and explanation of, big data, as well as proactive decision-making support. • A highly scalable proactive event-driven computing prototype integrating all SPEEDD components. The prototype will support a Human Factors validation by means of usability and effectiveness testing. • A suite of real-world demonstrations using live data in an operational environment.
SUMO: Supermodeling by combining imperfect models http://www.sumoproject.eu	Macedonian Academy of Sciences and Arts	Scientists develop computer models of real, complex systems to increase understanding of their behaviour and make predictions. A prime example is the Earth's climate. Complex climate models are used to compute the climate change in response to expected changes in the composition of the atmosphere due to man-made emissions. Years of research have improved the ability to simulate the climate of the recent past but these models are still far from perfect. The model projections of the globally averaged temperature increase by the end of this century differ by as much as a factor of two, and differ completely in regard to projections for specific regions of the globe. Current practice commonly averages the predictions of the separate models. Our proposed approach is instead to form a consensus by combining the models into one super model. The super model has learned from past observations how to optimally exchange information among individual models at every moment in time. Results in nonlinear dynamics suggest that the models can be made to synchronize with each other even if only a small amount of information is exchanged, forming a consensus that best represents reality. This innovative approach to reduce uncertainty might be compared to a group of scientists resolving their differences through dialogue, rather than simply voting or averaging their opinions. Experts from non-linear dynamics, machine-learning and climate science are brought together within SUMO to produce a climate change simulation with a super

Project	Organisation (Consortium lead partner)	Summary
		model combining state-of-the-art climate models. The super-modelling concept has the potential to provide improved estimates of global and regional climate change, so as to motivate and inform policy decisions. The approach is applicable in other situations where a small number of alternative models exist of the same real-world complex system, as in economy, ecology or biology.
THINKBIG: Patterns in Big Data: Methods, Applications and Implications http://erc.europa.eu/patterns-big-data-methods-applications-and-implications	University of Bristol	The availability of huge amounts of data has revolutionized many sectors of society, enabling engineers to bypass complex modelling steps, scientists to find shortcuts to new knowledge, and businesses to explore novel business models. For all its success, this field is still very young, and in need of systematic attention. Both risks and opportunities are very significant at this stage. They can be organized into three interconnected areas, which need to be addressed in a coordinated way: methods, applications and implications. By this we mean the interconnected needs to 1) develop new technology to take advantage of this resource; 2) explore the domains where this technology can make a significant impact; and 3) develop a set of cultural, legal and technical tools to reduce the risks associated with the application of these technologies to science and society. This project is about understanding, exploiting and managing the paradigm shift that is under way. It will address these three areas at the same time, 1) developing new types of algorithms and software architectures to take full advantage of this opportunity; 2) exploring new areas of opportunity for big-data to make an impact, with particular attention to the growing field of computational social sciences; and 3) investigating the ethical and epistemological challenges that arise from the transition towards a data-driven way of running society, business and science. We build on a strong track record in each of these areas. We have secured access to a valuable resource for historians, the collection of all UK newspapers from the past 200 years, which we will analyse with our tools, and we will greatly expand our current work on social media mining, working closely with colleagues from other disciplines. It is our intention to impact the social sciences, the general public and the lawmakers, besides our field of engineering.
urbanAPI: Interactive Analysis, Simulation and Visualisation Tools for Urban Agile Policy Implementation http://www.urbanapi.eu	Fraunhofer Gesellschaft zur Förderung der Angewandten Forschung	In the context of European initiatives to improve policy as a more transparent and understandable process, the urbanAPI project aims to support activities such as issue identification, policy analysis, consultation, decision and evaluation in urban planning and land management policy. For this purpose, a policy metamodel, a formalised vocabulary, a set of rule languages to define data integration and abstract simulation models are introduced. Furthermore, the urbanAPI approach will transpose elements of agile ICT development to the urban policy making process: Multiple activities can run in parallel, and all activities are kept synchronized. In such a process, risks are identified earlier, conflicts are understood better, and that knowledge gained in one activity can directly be used in all other activities. On this formal basis, a novel ICT toolset is developed. The urbanAPI toolset allows the fast development and deployment of participative policy support applications. These applications can be used for decision support, conflict management, analysis and visualisation and rely on

Project	Organisation (Consortium lead partner)	Summary
		innovative interaction platforms. Concepts known from web 2.0 and semantic web technologies, such as linking, tagging, rating and adding of semantic annotations, are supported directly. Applications created with the urbanAPI toolset also make use of a comprehensive data integration system to make use of the vast data resources – geospatial and statistical datasets – related to planning. Local initiatives in the four urban regions working in this project will be encouraged to participate within the planning process, to contribute to the final solutions and understand and finally accept the expected impacts on environment and habitants. Based on the toolset, adapted urban planning applications will be created, deployed, evaluated and used to support policy makers, planners and stakeholders at different governance and spatial levels (urban quarter level, municipal level, urban region level).
VELaSSCo: Visualization for Extremely Large-Scale Scientific Computing http://www.velassco.eu	Centre Internacional de Metodes Numerics en Enginyeria	VeLASSCo will provide new visual analysis methods for large-scale simulations serving the petabyte era and preparing the exabyte era by adopting big data tools/architectures for the engineering and scientific community, leveraging new ways of in-situ processing for data analytics and hardware accelerated interactive visualisation. Regardless of its origin, in the near future the challenge will not be on how to generate data, but rather in how to manage big and highly distributed data to make it easily handled by users in their personal devices. In the case of simulation data, the extraction of the specific useful information for analysis requires the treatment of large amounts of the data generated by the simulation, such as data for each time step, entire volume results when only some cuts are needed, etc.. The global goal of VELaSSCo is to provide big data tools, for the engineering and scientific community in order to better manipulate simulation with billions of distributed records. In large-scale simulations the domain is partitioned across several thousands of nodes, and the data (mesh and results) is stored on those nodes in a distributed manner. The VELaSSCo platform to be developed in this project will access this distributed information, will process it and send the results back to the users for local visualisation by their specific visualisation clients and tools (GiD from CIMNE, iFX from Fraunhofer, or other). Based on the layered structure commonly used in the big data field, the architecture of VELaSSCo platform is based on three layers: • Service layer: it is the communication point between the platform and the user's visualisation client, responsible of classifying user queries and send the results back. It will ensure access to the distributed data and provide a unified view of it. • Speed layer: it will ensure nearly instant access and visualisation of large-scale data, so that the user can interact with the simulation results (sub-second response times would be achieved by using simplified version of the data). • Batch layer: it will perform the heavy data processing tasks, such as complex queries, creation of simplified versions of the data and interpolation of DEM data to continuum domain. This layer will ensure resolution of complex queries and provide processed data to the speed layer. VELaSSCo will take into account the different types of

Project	Organisation (Consortium lead partner)	Summary
		simulation schemes (FEM, DEM, LB, etc.) and fields of application (Fluid and solid mechanics, electromagnetism, etc.) that could benefit from the use of its platform. The VELaSSCo consortium includes experts with relevant background in big data handling, advanced visualisation, engineering and scientific simulations. The team will be reinforced with the experience and feedback of a User Panel including research centres, SMEs and companies form key European industrial sectors such as aerospace, household products, chemical, pharmaceutical and civil engineering.
We.Gov: Where eGovernment meets the eSociety http://www.wegov-project.eu	University of Southampton	Social networking technology provides major new opportunities for policy makers (eGovernment) to engage with the community (eSociety). WeGov will develop a toolset so full advantage can be taken of well-established social networking sites (Facebook, Twitter, Bebo, WordPress etc.) when engaging citizens in two-way dialogs as part of governance and policy making processes. The tools will make it possible to detect, track and mine opinions and discussions on policy oriented topics including their origins, bias and evolution. The tools will allow discussions to be seeded and stimulated through injection of policy discussion points into relevant communities. Pseudo-anonymisation, audit trails, and access control will safeguard against misuse and ensure privacy for all involved. WeGov will be delivered as fully implemented software components packaged in a government toolbox broadly applicable to the interaction between eGovernment and the eSociety, including deployment on cloud infrastructure to dramatically lower the cost of operation. Success will be measured through three live field trials (diversity of opinion on public health, public spending in an era of financial uncertainty, and public policy research in the social sciences) conducted by three end-user partners. The trial results will be combined into a methodology, including a legal and ethical analysis and review by a confirmed external project Advisory Board of eGovernment stakeholders. The methodology, toolbox, and exemplar use cases will be published through a combination of open source, public reports and open workshops. Together with an exploitation strategy targeted at widespread take-up and use, long-term sustainability of the project results is assured. Policy makers will be freed from the inherent limitations of bespoke and dedicated platforms (e.g. government websites), and instead engage directly with citizens in their native online social environments to close the loop with this audience.

Technopolis, OII, CEPS (2015) based upon CORDIS and project websites

Appendix F References Chapter 2 (Literature review)

Aiden, E. and Michel, J.-B. (2014). *Uncharted: Big Data as a Lens on Human Culture*. New York: Riverhead Books.

Bengtsson, L., Lu, X., Thorson, A., Garfield, R. and Von Schreeb, J. (2011). Improved Response to Disasters and Outbreaks by Tracking Population Movements with Mobile Phone Network Data: A Post-Earthquake Geospatial Study in Haiti. *PLoS Med*, 8(8): e1001083.

Bollen, J., Van de Sompel, H. (2009). Clickstream Data Yields High-Resolution Maps of Science. *PLoS ONE*. 4(3): e4803.

Borgman, C. (2014). *Big Data, Little Data and Beyond*. Cambridge: MIT Press.

Boyd, D. and Crawford, K. (2012). Critical Questions for big data: Provocations for a cultural, technological and scholarly phenomenon. *Information, Communication and Society*. 15(5), 662-79.

Buelens, B., Daas, P., Burger, J., Puts, M. and Van den Brakel, J. (2014). *Selectivity of Big data*. Discussion paper. Statistics Netherlands. http://www.cbs.nl/en-GB/menu/methoden/onderzoek-methoden/discussionpapers/archief/2014/default.htm

Bulger, M., Taylor, G. and Schroeder, R. (n.d.), *Data-Driven Business Models: Challenges and Opportunities of Big Data*. http://www.nemode.ac.uk/?page_id=593

Clarke, A. and Margetts, H. (2014). Governments and Citizens Getting to Know Each Other? Open, Closed, and Big Data in Public Management Reform. *Policy and the Internet*, 6(4), 393–417.

Deetjen, U., Meyer, E., and Schroeder, R. (2015). *Big Data for Advancing Dementia Research*. Report for the OECD. http://bigdatadementia.oii.ox.ac.uk

Eagle, N. and Greene, K. (2014). *Reality Mining: Using Big Data to Engineer a Better World*. Cambridge MA: MIT Press.

Einav, L. and Levin, J. D. (2013). The Data Revolution and Economic Analysis. *NBER Working Paper Series* No. 19035: http://www.nber.org/papers/w19035

Generous, N., Fairchild, G., Deshpande, A., Del Valle, S.Y., Priedhorsky, R. (2014). Global Disease Monitoring and Forecasting with Wikipedia. *PLoS Comput Biol,* 10(11): e1003892.

Golder, S. and Macy, M. (2014). Digital Footprints: Opportunities and Challenges for Online Social Research. *Annual Review of Sociology*. 40:6.1–6.24

Greenleaf, G. (2012) Global data privacy laws: 89 countries, and accelerating. Queen Mary University of London School of Law. *Legal Studies Research Paper*. No. 98/2012.

Habermas, J. (1968). *Technik und Wissenschaft als "Ideologie"*. Frankfurt: Suhrkamp.

Kitchin, R. (2014). *The Data Revolution*. London: SAGE.

Lane, J., Stodden, V. (2014). Privacy, *Big Data, and the Public Good: Frameworks for Engagement*. Cambridge: Cambridge University Press.

Lazer, D., Kennedy, R., King, G., and Vespignani, V. (2014). The Parable of Google Flu: Traps in Big Data Analysis. *Science*. 343, no. 14 March: 1203-1205.

Mayer-Schonberger, V. and Cukier, K. (2013). *Big Data: A Revolution That Will Transform How We Live, Work and Think*. London: John Murray.

Meyer, E. and Schroeder, R. (2015). *Knowledge Machines. Digital Transformations of the Sciences and Humanities*. Cambridge: MIT Press.

Neuman, R.W., Guggenheim, L., Mo Jang, S. and Bae, S. Y. (2014). The Dynamics of Public Attention: Agenda-Setting Theory Meets Big Data. *Journal of Communication*. 64: 193–214.

Pasquale, F. (2005). *The Black Box Society: The Secret Algorithms that Control Money and Information*. Cambridge MA: Harvard University Press.

Pentland, S. (2014). *Social Physics: How Good Ideas Spread: The Lessons from a New Science*. London: Penguin.

Savage, M. and Burrows, R. (2007). The Coming Crisis of Empirical Sociology. *Sociology*. 41(5): 885-899

Schneier, B. (2015). *Data and Goliath*. New York: W.W. Norton.

Schroeder, R. (2014). Big Data and the brave new world of social media research. *Big Data and Society*. July-December: 1-11.

Schroeder, R. and Taylor, L. (n.d.). Big Data and Wikipedia Research: Social Science Knowledge across Disciplinary Divides, forthcoming in *Information, Communication and Society*.

Strassheim, H. and Kettunen, P. (2014). When does evidence-based policy turn into policy-based evidence? Configurations, contexts and mechanisms. *Evidence and Policy*. 10(2): 259-77.

Taylor L., Schroeder R. and Meyer E. T. (2014). Emerging practices and perspectives on big data analysis in economics: Bigger and better, or more of the same?. *Big Data and Society*. July-December: 1-10.

Taylor, L. and Schroeder, R. (2015). Is bigger better? The emergence of big data as a tool for international development policy. *GeoJournal*.

technopolis |group| The Netherlands
Spuistraat 283
1012 VR Amsterdam
The Netherlands
T +31 20 535 2244
E info.nl@technopolis-group.com
www.technopolis-group.com