Project acronym:	BYTE
Project title:	Big data roadmap and cross-disciplinarY community for addressing socieTal Externalities
Grant number:	619551
Programme:	Seventh Framework Programme for ICT
Objective:	ICT-2013.4.2 Scalable data analytics
Contract type:	Co-ordination and Support Action
Start date of project:	01 March 2014
Duration:	36 months
Website:	www.byte-project.eu

Deliverable D4.1:
Horizontal analysis of positive and negative societal externalities

Author(s):	Hans Lammerant, Paul De Hert, Vrije Universiteit Brussel Nelia Lasierra Beamonte, Anna Fensel, STI Innsbruck Anna Donovan, Rachel Finn and Kush Wadhwa, Trilateral Research and Consulting Stéphane Grumbach, Aurélien Faravelon, INRIA
Dissemination level:	Public
Deliverable type:	Final
Version:	1.0
Submission date:	31 August 2015

Table of Contents

EXECUTIVE SUMMARY

In this deliverable we present a horizontal analysis of the case study reports from D3.2. The case studies made an inventory of positive and negative externalities in the use of big data and were drawn from the domains of crisis informatics, culture, energy, environment, healthcare, maritime transportation and smart cities.

First we reviewed the big data practices and their similarities and differences encountered across the case studies. We assessed the big data practices in the case studies along the range of technical challenges through which big data is often characterised: volume, velocity, variety, veracity. To further clarify how these technical challenges are encountered in the different case studies, we map these challenges along the Big data value chain: data acquisition, data analysis, data curation, data storage and data usage. Next we compared the societal externalities encountered and checked which externalities are present across the sectors and which ones are sector-specific. We review the externalities according to the main categories of externalities presented in D2.1: economic externalities, social and ethical externalities, legal externalities and political externalities.

The big data assessment shows that big data does not stand for the same practice in every sector, but covers a wide variety of datasets and data. Volume is present in all case studies, but is quite different in scale. Velocity is not present in all case studies and some datasets are relatively static, but gets introduced when big data is based on data acquisition through sensors or on user interactions. Variety proves to be a challenge in all case studies, resulting from the combination of different sources or from the multidimensionality or unstructured character of the data. Veracity often captures a wider range of problems like validity, data granularity or resolution, uncertainty about meaning. These issues show up differently in the case studies. Mapped across the Big data value chain the technical challenges raised in the case studies are mostly observed in the data collection, data curation and in a lesser extent in the data usage phase. Often these challenges are not purely technical challenges, but the translation of societal externalities.

The comparison of the societal externalities shows a range of positive economic and societal impacts. We have observed positive economical externalities in terms of innovation and in improvements in efficiency. This also leads to changes in business models and the appearance of new business models. Such changes are not by definition positive and can lead to dominance of and dependence on a few technological players. On the other hand, even in such a case there can be enough space for smaller niche players, while open source and open data are methods to counter dependence relations. Further, despite these positive economic impacts the role of public funding proves to be important into kick-starting a data economy. This can be explained by the lack of sustainable business models for data platforms and data sharing.

In all case studies positive social externalities were reported, similar to improved efficiency or innovation but for social, non-economical aims. In several case studies the potential for improved, evidence-based decision-making and/or participation was mentioned. On the other hand the risk for negative impacts on important social values could also be observed. In most case studies (potential) negative effects on privacy were reported, while several case studies mentioned the risk for equality and new risks for discriminatory practices. In most case studies trust problems can be found, where the risk for manipulation and exploitation leads to distrust and withdrawal, thereby negatively affecting the potential positive impacts of big

data. This points to the need for developing practices, including but not limited to legal frameworks, which can assure a proper balance and thereby establish trust.

Two major legal frameworks, data protection and intellectual property rights, prove to have an important impact on big data and act as a barrier. Data protection, and to a lesser extent intellectual property rights, were raised as important to protect other societal values from negative impacts of big data. But in general both frameworks were considered outdated and too restrictive for big data.

Mostly issues related to political economics came forward as political externalities. These were generally noticed as negative externalities, creating vulnerabilities for the operations or for big data processing in particular. Mainly two issues came forward: on the one hand the relation between public sector or non-profit organisations and the private sector, on the other hand the fear of losing control to actors abroad, and in particular US-based actors, which sometimes translates in protectionist attitudes and requirements to store data within national territories. However, the attitudes towards and perceptions of these two issues varied among the case studies.

The overall picture shows positive benefits but also the potential to negatively affect other important social or ethical values. Important is that big data is not just a technical issue but has an impact on organisational borders and the 'business ecology' in general. This leads to uncertainty and conflict in a range of areas, translating in distrust and reluctance by all sorts of actors and conflicts on political and legal level. Organisational borders need to be redefined or redrawn, while also social norms and legal frameworks need to be clarified again based on a proper balancing of all interests.

Appendix 2 on the last page gives an overview of the externalities encountered in a summary table.

1 INTRODUCTION

In this deliverable we present a horizontal analysis of the case study reports from D3.2. The case studies made an inventory of positive and negative externalities in the use of big data and were drawn from the domains of crisis informatics, culture, energy, environment, healthcare, maritime transportation and smart cities.

First we will review the big data practices and their similarities and differences encountered across the cases. Next we will compare the societal externalities encountered and check which externalities are present across the sectors and which ones are sector-specific. Further we will compare these observations with the results of the earlier research. The aim of this horizontal analysis is to identify how these externalities are connected to the big data practices and to each other. A better view on the interconnection of externalities and big data practices establishes a footing for the later work on how to address and evaluate these externalities.

2 METHODOLOGY

In this deliverable we will make a horizontal analysis of the positive and negative externalities encountered across the case studies. The case studies concerned the following sectors: crisis informatics, culture, energy, environment, healthcare, maritime transportation and smart cities. In each sector one or more actors using big data was researched:

- Crisis informatics: a research institute using social media, especially Twitter data, to support humanitarian relief efforts during crisis situations with crisis 'maps'.
- Culture: a pan-European public initiative providing access to digitised cultural heritage works.
- Energy: operators and suppliers using big data in the exploration and production of oil & gas in the Norwegian Continental Shelf.
- Environment: an earth observation data portal coordinating access to earth observation data
- Healthcare case study: a health institute using big data for diagnostics of rare genetic diseases.
- Maritime transportation: the use of big data by a range of actors in the shipping industry
- Smart city: value creation based on data from urban infrastructure like transport and energy

The case study results were presented in D3.2. In this deliverable we provide a comparative analysis of these results.

In a first step we will review the big data practices themselves. Big data is not a very well-defined term and often used as a buzz word. Therefore it is important to take stock of the actual big data practices in the case studies. To do so we compare the use of big data in the case studies along the range of technical challenges through which big data is often characterised and which have been included in the definition of big data in D1.1: volume, velocity, variety, veracity. To further clarify how these technical challenges are encountered in the different case studies, we map these challenges along the Big data value chain, as defined in the BIG project. The data value chain has been characterised as consisting of the following steps:

- Data Acquisition is the process of gathering, filtering and cleaning data before it is put in a data warehouse or any other storage solution on which data analysis can be carried out.

- Data Analysis is concerned with making raw data, which has been acquired, amenable to use in decision-making as well as domain specific usage.
- Data Curation is the active management of data over its life-cycle to ensure it meets the necessary data quality requirements for its effective usage.
- Data Storage is concerned about storing and managing data in a scalable way satisfying the needs of applications that require access to the data.
- Data Usage covers the business goals that need access to data and its analysis and the tools needed to integrate analysis in business decision-making.[1]

The technical challenges encountered will be situated in these steps and compared across the case studies. Both steps allow us to get a comprehensive view on the similarities and differences between the big data practices found in the different sectors.

In the second step we review the societal externalities encountered in the case studies. As this analysis is based on the case studies, we adopt the same definition for externality as used in the case study methodology:

- Positive externalities occur when a product, activity or decision by an actor causes positive effects or benefits realized by a third party resulting from a transaction in which they had no direct involvement.
- Negative externalities occur when a product, activity or decision by an actor causes costs (or harm) that is not entirely born by that actor but that affects a third party, e.g. society. It is generally viewed as a failure of the market because the level of consumption or production of the product is higher than what the society requires.[2]

A problem is that the boundary between internal and external is often an arbitrary delineation to look at the positive results of big data practices. In D3.1 it was already noted that "there exists no clear cut definition of the system boundaries" and that "the arbitrariness in choice of system boundaries automatically means there is no uniform understanding of who could be the affected third parties".[3] Therefore we will in some instances also look at the internal impact and not just the externalities. Further this boundary plays an important role in the capacity of actors, be it companies or citizens, to address the impact of big data. This impact can be perceived and valued differently when it originates from an internal or an external practice.

We will compare and check which externalities are present across the sectors and which ones are sector-specific. We review the externalities according to the main categories of externalities presented in D2.1: economic externalities, social and ethical externalities, legal externalities and political externalities. For each of these categories we check which externalities are present across sectors or only in specific case studies. The coding of observations is used to sort the observations and to link them with specific externalities. In appendix 1 we give an overview of which codes were used for each externality. The codes are not exclusively used for one externality, but whenever it is relevant for a specific externality. When useful we also point to observations made in case study reports which were not coded. Further we check if these observations concur with the externalities registered in the literature study in D2.1 and point out new externalities found in the case studies.

[1] Zaveri, Amrapali, Andre Freitas, Andreas Thalhammer, et al., *D2.2.2 Final Version of Technical White Paper*, Big Data Public Private Forum (BIG), 14 May 2014, pp. 1-2

[2] Guillermo Vega-Gorgojo, Grunde Løvoll, Thomas Mestl, Anna Donovan and Rachel Finn, *D3.1 Case study methodology*, BYTE, 30 September 2014, p. 9.

[3] *Ibid.*, p. 10.

A summary table of the externalities encountered in the case studies in included in Appendix 2.

3 BIG DATA IN THE CASE STUDIES

3.1 BIG DATA ASSESSMENT

We make an initial big data assessment across sectors by comparing the assessment in the case studies of the characteristics of big data: volume, velocity, variety, veracity.

The maturity of big data processing varied a lot along the case studies. The maritime case study showed a potential for big data processing, especially for suppliers of subsystems and equipment and for navigation data. But in general this potential was not used, due to a lack of short term return on investment leading to an unwillingness to invest. The environment case study on the other hand showed a mature big data ecosystem, where the presence of big data is taken for granted. Even in this context there is still a lot of potential for further development. The other case studies showed varying levels of early development. They presented clear cases of big data processing, but they also remained limited or in an early stage, while being presented by a range of technical challenges and societal barriers.

When we assess the characteristics of big data present in the case studies, we can notice quite some differences. In all case studies we encounter large volumes of data, but big can still differ a lot. Highest volumes we encounter in the environment case study, where raw satellite data can be in the order of PB's while processed raster coverage is sized in the order of GB's. Similar volumes can be encountered in the energy case, where seismic data is the typical big data product of which 1 raw set can have a size of 100s GB. Total holdings of data measure in the PB range. Also sensor-driven data can produce large datasets. E.g. a subsea factory can produce 1 TB/day. The further deployment of sensor-driven applications will lead to a fast growing amount of data. Similar sensor-driven applications are possible in the maritime case, but remain a not developed potential. Also in the smart city case such high-volume applications remain underdeveloped.

In the other case studies volumes are lower. The crisis informatics case deals with several 100000's tweets/day, which comes down to several GB's/day. The culture case study concerned a set of 190 million metadata records, but growing much more slowly. The healthcare case concerned genetic samples providing datasets of 15-20 GB. The research institute had about 20-25 TB of genetic data stored. Also a lot of applications in the environmental, energy or smart city case studies will have similar volumes. This overview has to be considered as a snapshot of actual use in the case studies, and not of the potential use.

Case study	Raw dataset or typical volume used – order of magnitude:	Total volume of data holdings – order of magnitude:
Environment	PB	not available
Energy	TB	PB
Crisis informatics	GB (daily volume)	not available
Smart city	GB	not available
Healthcare	GB	TB
Culture	small	GB
Maritime	KB-GB	not available

Table 1: Volume of data – order of magnitude

In all these cases there is a potential for applications involving much larger datasets by combining datasets or through the use of more fine-grained data acquisition. E.g. combining genetic datasets with clinical data or comparisons with repositories of genetic data in healthcare, more use of user data in the culture or of finer-grained user data in the smart city context. More sensor-driven applications in the energy, maritime or smart city sectors can push typical volumes used into the TB range.

In terms of velocity, not all applications involve a high velocity of data acquisition. In the culture case it concerns a relatively static and slowly growing dataset. In other cases it concerns the results of a distinct survey, like the seismic surveys in the energy case or the genetic samples in the healthcare case. Once the data acquisition becomes more sensor-driven or user data gets used, the velocity of data acquisition becomes an important issue. E.g. social media data for crisis informatics, operational monitoring with sensors of equipment or the environment in the energy case, user data or infrastructure use monitoring in the smart city case. The potential applications, by monitoring user data or sensors, introduce a velocity challenge in other areas as well.

Case study	Velocity
Environment	high
Energy	high
Crisis informatics	high
Smart city	high
Healthcare	low
Culture	low
Maritime	low

Table 2: Velocity of data acquisition – order of magnitude

Variety is a challenge is all case studies. It is a challenge when combining different data sources, like sensory data with other data. And certain data is challenging due to its multidimensionality, like genetic data, or due to its unstructured nature like social media data. Variety being or becoming one of the main challenges was remarked in several cases, like in the crisis informatics, culture, healthcare or the energy case. It is thus clear, that value created out of data requires, across sectors, data integration from multiple sources. For that, as reported in D1.4[4], semantic technologies can be efficiently used to provide with structured data and further linked and integrated data. Particularly, exposing data as Linked Data and the utilization of normalized vocabularies can contribute to the successfully integration of Big Data.

Veracity was mentioned in several, but not all, case studies as a challenge, and the actual challenge varied a lot. Several issues are mixed in this topic: the correctness of the data (validity), the resolution or granularity of the data, the uncertainty concerning the meaning or the meaningfulness of the data (veracity). In crisis informatics working with social media data poses problems concerning validity. In the focus group it was mentioned that no action could be taken upon social media data except after verification and validation, as it contains a lot of rumours or duplication of old information. An automated treatment of such social media data is also confronted with a veracity challenge. The research institute in the case study has developed a crowdsourcing solution to train its algorithms to recognize meaningful information. In the culture case study veracity concerned problems of data quality. Main challenge is to assure the quality of metadata, as metadata is often developed for specific uses and can suffer from concept drift and divergent use. Challenges concerning data quality, validation and veracity were also mentioned in the environment and the smart city case study.

This comparison shows that big data, as an umbrella, captures datasets and data practices between which the characteristics and challenges can still vary a lot. Volume is present in all case studies, but big can still be quite different in scale. Velocity is not present in all case studies and some datasets are relatively static, but gets introduced when big data is based on data acquisition through sensors or on user interactions. Variety proves to be a challenge in all case studies, resulting from the combination of different sources or from the multidimensionality or unstructured character of the data. Veracity often captures a wider range of problems like validity, data granularity or resolution, uncertainty about meaning. These issues show up differently in the case studies.

This overview has to be considered as a snapshot of actual use in the case studies, and not of the potential use. Important to notice is the potential for applications involving much larger datasets by combining datasets or through the use of more fine-grained data acquisition raised in all of these case studies. The barriers to this potential use can be technical but are often caused by societal factors, ranging from economic factors, like the lack of funding, of fast return on investment or of adequate business models, to other factors like limitations due to privacy or intellectual property rights.

[4] Sebnem Rusitschka, Alejandro Ramirez, D1.4 Big Data Technologies and Infrastructures, BYTE, 31 August 2014, pp. 28-29.

3.2 TECHNICAL CHALLENGES

Across the case studies we notice that the technical challenges faced by the big data practices in the case studies partly differ. Here we map them across the Big data value chain: data acquisition, data analysis, data curation, data storage and data usage.

- Data acquisition

Although not a general observation, in several case studies the lack of sensors or the need for more fine-grained sensors was mentioned. This is partly a technical challenge, partly an economic challenge of making the necessary investments. In the energy case study it was mentioned that the oil and gas sector started the last decade to install sensors in every piece of equipment. In this case sensors have become not so much a challenge any more. But data acquisition still encounters economic barriers especially in the area of exploration, as seismic tests are very costly. The environment case study showed a wide range of sensors used. Main technical challenge is in improving the resolution (and the frequency and range of the data acquisition, which are generally a trade-off with resolution). Improvement of data acquisition is needed to make environmental information available at different scales from local to global and time scales from minutes to years. In the maritime case study the lack of sensors came clearly forward. Data acquisition happens often still manually. Also the SCADA systems in use are not designed for large scale data acquisition through sensors and the capacity for real-time access is lacking. Solution of this technical challenge came forward as an economical challenge, dependent on investments. The need for more sensors or more fine-grained sensors was also mentioned as a general problem in the smart city case study. On the one hand this is an economical challenge, as more investments are needed. On the technical level the challenge is linked to the sharing of the data: "how the data is being collected and stored massively influences how easily it can be shared". We can conclude that a turn to sensor-driven big data practices is both a technical and an economical challenge.

Another challenge in terms of data acquisition, present in all case studies, is the access to other data sources. In the crisis case study, Twitter data was the main source as this was easily accessible, while other sources pose commercial and legal barriers. Commercial secrecy plays a role in impeding data sharing in the energy, the maritime and the smart city case study. Intellectual property rights are an important barrier in the culture case study. Privacy and protection of personal data pose a barrier in the healthcare, the smart city, the crisis and the environment case study, while exploiting the potential of user data in the cultural sector will raise similar issues. Again, these barriers to access other data sources are in the first place not a technical issue. They are legal, commercial or security concerns which lead to the creation of isolated data silo's. But allowing a restricted access while assuring that these concerns are respected turns it also into a technical challenge (such as privacy by design or security by design), as was noticed in the health case study.

- Data analysis

Technical challenges in the data analysis phase were only occasionally mentioned. This is surprising seen the fact that variety is one of the main challenges for analytics[5] and it is mentioned as a challenge across all case studies. This does not mean that no improvements on this level are possible, but rather that the more obvious challenges present themselves in other

[5] *Ibid.*, pp. 28-29.

areas. One such area linked to variety is the interoperability of datasets, which we consider under data curation.

The need for new analytics was raised in the energy case, where a lack of good methods to analyse the data was noticed. Data-driven methods were still under-performing compared to older methods based on physical models. In the crisis informatics case study analytical tools were still experimental and under development. In this case the challenge presented itself less as a need to develop good analytical methods, but rather to make a useful tool. In the healthcare sector the issue in the analytical phase is one of available processing capacity. Again this is mainly an economic challenge in terms of investments.

Once integration models are developed, predictive and prescriptive model analytics will serve to extract further insights out of Big Data. As described in D1.4, research at the technical level for data analysis involves new machine learning techniques and particularly on the so called “deep learning” techniques which enable the realization of predictive analytics.

- Data curation

Interoperability challenges were at evidence in all case studies. In the crisis informatics case study the need for more standardization was mentioned. Different formats made it difficult to analyse all data in time-sensitive situations. The institute studied in the culture case study focused on curating metadata and assuring its interoperability by promoting its data model, including an open access license. Achieving this objective is confronted with several challenges concerning data quality. One is the multilingual nature of cultural data, which has been dealt with by incorporating methods of translation into the data model. Issues with metadata quality also follow from the different uses made of this metadata and the different requirements it engenders. Also in the environmental case study problems with interoperability, data quality and data integration were raised. Data quality is partly linked with accuracy and resolution problems raised under data acquisition. But a lack of standardization, particularly of the data format, reinforces the challenges. Open access on the one hand mitigates problems by getting rid of authorization and authentication technicalities. On the other hand it makes maintenance of quality control more difficult. The research institute in the healthcare case study mentioned problems to make data interoperable across all its databases and some external databases. In the maritime case study interoperability is also generally a problem, with the exception of AIS navigation data, due to the lack of standards. The smart city case study mentioned a need for platform for intermediation between citizens, resource and infrastructure operators through which data can be shared. Interoperability issues further originate from the use of closed systems and proprietary formats. Such interoperability challenges present themselves as technical challenges but are also linked with societal challenges, as they generally need investments to solve them and are therefore linked with the availability or lack of incentives to do so.

Veracity was also a challenge mentioned in several case studies. In the crisis informatics case study crowdsourcing was used. Thousands of digital volunteers do the first analysis of samples of data, which could be used to train the machine learning algorithms. In the focus group similar solutions were mentioned for verification and rumour control. In general data from social media could only be used with some sort of verification. In the environment case study the lack of shared veracity, value and validity criteria were raised. Similarly, validation was raised as a problem area in the smart city case study.

Also privacy and data protection and data security lead to technical challenges for data curation. In the health case study maintaining anonymity was raised as a technical challenge in itself. In the research institute it was solved with developing new database solutions limiting access based on a need to know. But in general it is considered impossible to assure anonymity with genetic data or rare diseases. Also in the environment and smart city case studies privacy issues do present specific technical challenges. Data security was mentioned as a technical challenge in the maritime, environment and smart city case studies, while commercial secrecy and IPR was raised in the energy and maritime case study. In general, these external societal concerns engender technical challenges for data curation. This concurs with the observation in D1.4 of the increasing importance of legal and security aspects as emerging trend in big data research.[6]

- Data storage

Data storage was mentioned as a challenge in the crisis informatics case study. The need for more capacity was raised. Cloud solutions were considered as a primary need. Public-private partnerships around such hosting are seen as a solution and actively solicited, but create again their own challenges in terms of dependence on external, often US-based, infrastructure. Humanitarian organisations are also hesitant about the unpredictable engagement of private companies outside the crisis situation, as these companies use this engagement during crisis situations in the first place to promote their brand. The open source nature of the products now offered by the research institute in this case study makes it now more trustworthy.

In the environment case study sustainability was an issue. The availability of the data and continuous access has to be guaranteed. Also long-term maintenance of the infrastructure can pose a problem. Another technical challenge concerns increasing both storage and transfer velocity.

These challenges are partially technical challenges. The need to store vast amounts of data has driven research to investigate new solutions for data storage. Solutions to address Big Data storage challenges includes non-relational databases such as Graph DBs, key-value stores, Columnar DBs. Also, solutions such as *Hadoop* are considered by researchers and industry for management of massive data (see D1.4 for more information). But in general challenges concerning data storage showed up as linked to societal externalities: concerns about funding and of loss of control and dependency when opting for the cheapest solutions offered by private companies.

- Data usage

In the crisis informatics case study usage challenges concerned organisational cultures, which made the integration of these new information sources into existing workflows and decision-making mechanisms difficult. Possible solutions concerned turning the data into more familiar information products, but this requires additional data processing work. In the energy case study effective usage was considered low and not reaching its potential. The mindset is still focussed on more classical uses of data, while a lot of reluctance exist to choose for newer big data practices which still have to prove its effectiveness. The earlier raised problems of data silos also creates similar challenges for data usage, resulting in a lack of integrated usage of different data sources. Similarly conservatism was mentioned as the main

[6] *Ibid.*, pp. 29-30.

reason for not adapting big data practices, although main challenge is not technical but a lack of short term return on investment. Data usage challenges were also raised in the environment case study. The institutional gap between mapping authorities and scientists engenders interpretation problems. Also, industrial competitors are seen to use the lack of standards or standard violations to strengthen their position. Again, most of these perceived technical challenges are social challenges to adapt to new technologies. Organisations and people need to adapt work processes and mindsets to the new technological potentialities. Partly this returns as a technical challenge through the need to develop standards for data and their interpretation.

From this overview it appears that challenges are mostly observed in the data collection, data curation and in a lesser extent in the data usage phase. Often these challenges are not purely technical challenges, but the translation of societal externalities. Improving data acquisition through more and better sensors or improving interoperability and data quality needs investments and is therefore dependent on the availability of public funding, the return on investment or the availability of business models which allow to capture the benefits. Answers on social and legal concerns like privacy and security present themselves also as technical challenges. We can conclude that the case studies show a narrow link between most of the technical challenges and societal externalities. Big data technologies, or better the processes using big data, are not just the cause of societal externalities, but these societal externalities shape as well the technical challenges. All case studies show that there is still a large and unused potential for big data applications and that the barriers to this potential can in a large degree be found in the societal externalities.

4 SOCIETAL EXTERNALITIES

In the following section we analyse societal externalities encountered in the case studies. We compare and check which externalities are present across the sectors and which ones are sector-specific. We review the externalities according to the main categories of externalities presented in D2.1: economic externalities, social and ethical externalities, legal externalities and political externalities. For each of these categories we check which externalities are present across sectors or only in specific case studies. The coding of observations is used to sort the observations and to link them with specific externalities. In appendix 1 we give an overview of which codes were used for each externality. The codes are not exclusively used for one externality, but whenever it is relevant for a specific externality. References to the observations in the case study reports in D3.2 are made in the text by mentioning the codes. When useful we also point to observations made in case study reports which were not coded. Further we check if these observation concur with the externalities registered in the literature study in D2.1 and point out new externalities found in the case studies.

The maritime case study was excluded from this comparison because minimal applications of big data were found. It showed a sector which did not envision its introduction soon and had therefore difficulties to reflect on potential societal externalities. Using it in comparison with the other case studies to derive findings on societal externalities would therefore lead to a misrepresentation. However, the maritime case study shows some barriers for adoption of big data in this sector, especially on the economical level. When relevant we mention this in the comparison as a further corroboration of the results.

The analysed externalities are classified according to the following categories: (1) Economic externalities, (2) Social and ethical externalities, (3) Legal externalities and (4) Political externalities. Some of the externalities can be considered under several of these groups and are sometimes differently categorised across the case studies. Here we regroup those observations together under the most relevant category, which can therefore differ from the original categorisation in the case study.

4.1 Economic externalities.

The following economic externalities are discussed: improved efficiency, innovation, changing business models, employment and the role of public sector funding. The role of public sector funding was not raised in D2.1 nor specifically coded in the case studies, but it showed up as relevant in the specific observations found in the case study reports.

- Improved efficiency

First we look at how big data processes resulted in forms of improved efficiency. Under improved efficiency we understand the whole of improvements of existing processes. Cost-efficiency can result from more efficient use of resources, improved quality control, faster decision making, better targeting of services and so on. In general it results from improved management based a better information position which allows faster and more fine-grained command and control over business processes. The emergence of new value chains we discuss under 'innovation', here we look at improvements of existing processes. Improved efficiency of internal processes cannot be categorized as an externality, as it is often the actual reason for which investments in big data are made. However, such improved efficiency does not have to be limited to the actual processors, but can as well lead to such efficiencies elsewhere. Therefore improved efficiency can be classified both as an internal improvement as an externality. Further the internal improvements of efficiency can also lead to external benefits in a broader sense. These benefits can be both of an economic as a social nature, like improved healthcare or avoiding detrimental environmental effects, and therefore also show up under positive social externalities.

Positive results in terms of improved efficiency were found in most of the case studies, although also some negative results. Big data utilisation leads to better services in the crisis, environment, healthcare and smart city case studies. In the crisis informatics case study the information product of the research institute allowed humanitarian organisations to deliver better services to the public, as it improved situational awareness and led to better services providing relief faster and allocating resources to where the need is highest (E-PC-BM-2). Therefore services are better targeted (E-PC-BM-3) and resource allocation is improved (E-PC-BM-4). Resource efficiency also improves through a better work division between organisations with technical capacity that are analysing the data and the humanitarian organisations focusing on relief (E-PC-BM-4). The big data tools also let to a better crisis preparedness by using the tools to predict trends and needs (E-PC-TEC-1). However, some negative effects on resource efficiency were suggested. The increasing use of big data requires more resources to enable this. The infrastructural needs linked with big data also lead to a money flow towards large technology companies to provide services (E-PC-BM-4). Under the ethical externalities also a negative externality linked to management efficiency was mentioned: the risk for misinterpretation of data, especially when the data gets separated from the contextual knowledge about its creation (E-OC-ETH-13). Similarly in the environment case study a better utilisation of current services was signalled, which extended the operational life of satellites (E-PC-BM-2). Further, the use of big data was seen having strong impacts on economies through providing reliable environmental data (e.g. sea data for fishing nations and weather data for tourism) (E-OO-BM-7). The data also leads or will lead to improved decision making for sustainability, and possibly reduced disaster risk and more environmental safety (E-PC-BM-1). Also here, some possible negative implications were noticed. The high energy consumption contradicts the aim of sustainability. Also inefficiencies caused by excess of pre-computed data (E-OO-BM-7) or excessive trust in data-intensive applications has been highlighted as a possible negative implication (E-OC-ETH-12). In the healthcare case study the potential for cost saving for healthcare organisations was highlighted. Such cost saving results from more accurate and timely diagnoses and efficient treatments. This also allows to allocate resources more effectively. (E-PC-BM-2) In the smart city case study a large efficiency increase potential was indicated along the dimensions of time, costs, and resources through improved situational awareness on the city and cross-optimization of resource networks (E-OC-TEC-2).

Improved efficiency was not signalled in the energy or the culture case study. But we find other positive economic externalities in those case studies. And we have to take into account that this means no efficiency gains were found as externality, but it does not exclude a positive effect on efficiency for the internal operations. E.g. in the energy case study improved efficiency in the internal operations is clearly provided by conditions-based maintenance of equipment or improved operational and well monitoring. Similarly, in the maritime case study improved efficiency as an externality was in general not on the mind of the interviewed persons. On the other hand, the example of an information service collecting and providing data on prices for container shipping showed that it could have the external effect of making the market more transparent and consequently also the market mechanism more efficient. Improved efficiency of internal operations was envisioned from condition-based maintenance of equipment by suppliers, but its return on investment was in general considered too low to do the effort of investing.

- Innovation

Big data cannot only improve efficiency of existing processes, but also engender innovation. Under innovation we understand in this context the formation of new value chains or major transformations of existing ones.

The crisis case study shows innovation in the building of a new data value chain, based on social media. The research institute is an essential actor on this value chain, but it also results in an innovative use of the results in the humanitarian organisations. It further shows the importance and potential of open data and open source code for innovation. The data value chain is made possible exactly because the Twitter data is open, and other sources cannot be used because of their closed character. Further by publishing its code as open source the research institute hopes to involve external expertise in the development of its code (E-OC-DAT-2).

The culture case study also concerns an actor which allows the building of a new data value chain, by aggregating cultural metadata and making it available for other uses. Innovation also result from the institutes work through the spread and adoption of its data model by other actors, which further improves the interoperability of cultural data. Again, this case study shows the important role of open data for establishing new data value chains, while also signalling the barrier posed by copyright issues. The major impact of this institute lies in making the metadata open and accessible for re-purposing, allowing a wide range of new applications. No direct economic value can be derived from the metadata, but it allows innovative re-use. The indirect economic value through an enlarged visibility encourages providers to make their data available. (E-PC-DAT-1, E-PO-BM-2) The closure of data behind copyright is on the other hand mentioned as one of the main barriers. Partly also because it results in a lot of uncertainty and misunderstanding leading to a 'copyright paranoia'. (E-PO-LEG-2) Potential for further innovation can be found in interaction data, which is still underused despite potential economic benefits. Reason mentioned is the public positioning of cultural sector, focussing on deriving cultural and social value instead of commercial value, leading to less attention for this potential. (E-PC-TEC-2)

In the energy case study several innovative uses of big data were mentioned, resulting in in new services or new methods at a diverse range of actors. Examples were condition-based maintenance of equipment and new techniques and services monitoring well integrity, drilling and operations. (E-OO-BM-1, E-OO-BM-2). Also in this case open data plays a role, as Norway's regulator obliges the companies to send certain data and makes it available as open data. Through this open data policy the regulator establishes a certain information ecology to drive competition. (E-PO-BM-1).

Also in the environment case study the advent of new innovative services and uses based on big data was observed (E-PO-BM-2). Access policies and data sharing play an important role in this, both within the public sector as with the private sector. The group studied in the case study coordinated access and interoperability of a wide range of earth observation data, which was in large part provided through open access. Open data is seen as an important enabler of innovation (E-PO-BM-1, E-OO-BM-1), while restrictive IPR legislation is seen as an impediment (E-PO-LEG2). Another important enabler are investments in standardization (E-OC-DAT-1).

The healthcare case study observed a range of opportunities through innovation, like the development of marketable treatments and therapies, the innovation of health data technologies and tools (E-PC-BM-2, E-PO-BM-2). The smart city case study pointed to potential innovations in service delivery and utilisation of resources, but also showed important barriers for such innovation. These barriers were formed by the need of data platforms allowing data sharing. Again, the issue of access to data and open data pops up. But another important barrier is a lack of investment in such platforms, and public investment is seen as key to develop an innovative digital sector around a city. (E-PC-DAT-1 , E-OC-DAT-2, E-OO-BM-2).

To conclude, we can see big data in all case studies leading to actual innovation, also by external actors, or a large potential for such innovation. Data access and data sharing is key to enable such innovation. In several case studies the availability of open data leads to actual innovation, in other case studies it is the lack of such availability or data access which poses a barrier. Again, the maritime case study provides an exception where innovation is lacking. Reason is mainly a lack of return on investment and a conservative mindset. But also the lack of an adequate business model providing incentives to share data.

- Changing business models

These opportunities for innovation also lead to the advent of new business models. A business model is here understood as an organisational structure aimed at capturing value. A new business model implies the formulation of a specific value proposition towards customers and a specific activity to do so in the most cost-efficient and therefore competitive way. New business models based on big data can emerge through the development of new services based on a new use of data or by specialisation in specific services needed as part of the data value chain (e.g. specific analytics or data curation services).

The potential for new business models was observed in the culture case study, e.g. for tourism or education, in the healthcare or in the environment case study (E-PO-BM-2, E-OO-BM-1). The energy case study saw a range of new business models around data generation or data analysis (E-OO-BM-2). Also commercial partnerships develop around data between operators and suppliers. Several suppliers of equipment develop data-enabled services around the equipment, like conditions-based maintenance (E-OO-BM-1). But also problems exists with developing adequate business models. In general companies are reluctant to share data, especially when it threatens existing business models. The lack of clear business models also translates into a reluctance to invest, while specific partnerships are contract-based and result in data silos. (E-OO-BM-3) The obligatory data flow towards the regulator and its open data policy provides some counterbalance against this reluctance (E-PO-BM-1).

The smart city case study mostly showed the barriers for new business models to develop. This is dependent on the establishment of platforms on which data can be shared. Such platforms do not have a clear business case and therefore investments are lacking. Where such platforms exist in the culture and the environment case study based on public funding, also in the smart city case study public funding is often seen as a solution to establish the necessary enabling infrastructure. (E-OO-BM-2) Some experts also made clear that data markets were not a suitable business model, as raw data has no value in itself while it is very difficult to keep track of data through complex machine learning algorithms.

The maritime case study provides a negative example concerning changes in business models, which also helps clarifying the issues at play. Generally no investments are made due to a lack of return on investment. But such lack also follows from the lack of an adequate business model providing incentives for the sharing of data. Generally the only examples of big data applications effectively coming under consideration are therefore in supplier-ship owner relations, where the investment decision turns out negatively.

However, negative changes were also perceived. Some existing business models can be threatened, as was brought up in the environment (E-OO-DAT-1) and the smart city case study (E-OC-BM-5). Rent-seeking by the private sector on public investments was brought up in the culture, environment and smart city case studies. Both in the culture and the environment case study the inequality between public institutions providing open data and private actors using that open data but having no obligation to share their data was pointed out. Rent-seeking can also be more indirect, e.g. search engines like Google use interaction data with cultural websites for targeted advertising. (E-PO-DAT-1) The risk for dependency on large companies and monopolies was brought up in the environment, smart city, as well as the crisis case study (E-OC-BM-8). In the environment case study the fear was raised that big data favours big players and would drive SMEs out (E-OO-DAT-1). This can result in the creation of a few dominant players leading to reduced market competition, as well as new private data silos (E-OC-BM-7). On the other hand, counterbalancing opinions were raised as well, seeing niche opportunities for small players, especially concerning data analysis. The question of concentration also came up in the smart city case study. It was pointed out that the need for platforms requiring a certain investment pointed towards concentration and quasi-monopolistic structures. The solution put forward was to let the public sector make this investment and subsequently opening it up as a commodity or utility. Open source and open platforms was also put forward as an answer to prevent monopolistic structures. (E-OO-BM-2, E-OO-BM-3, E-OO-BM-5)

To conclude, in general we see changes in business models and the introduction of new business models. Innovation based on big data changes the 'information ecology', or the opportunities to capture value from data. It can therefore also lead to a new 'business ecology' of organisations when new tasks get outsourced or new services acquired. This can lead to new business models, but also risks for dependencies on new dominant players or the 'creative destruction' of older business models.

An important question is how the opportunities to capture value from data get translated into organisational structures to do so. Big data practices are dependent on the availability of data and therefore on the cooperation of others to provide data. If the value gets captured in a way that results in benefits for other companies or citizens, these will be more willing to cooperate. If the big data practice is perceived as (potentially) captive or rent-seeking with strong negative effects or leading to negative externalities, other actors will be reluctant to share data. In that case big data practices will only be taken up for internal processes or in limited client-supplier relations heavily regulated by contract and with closed or proprietary data flows. These circumstances present a push for more integration of organisational structures (vertical integration instead of outsourcing of specialized tasks) instead of the emergence of new business models. When the negative effects are considered limited and outweighed by the benefits of using the service, an uptake remains possible.

- Employment

The growing use of big data also has its effect on employment. In the energy and smart city case study the need for more data scientists was observed (E-OC-BM-3). In the environment case study opportunities for new jobs and business were raised (E-PO-BM-1). Potential negative employment effects were also mentioned in the smart city case study, as certain jobs will become unnecessary and disappear (E-OC-BM-5). The employment effects were only mentioned in 3 case studies. But, seen the conclusion on the change in business models, effects on employment are a logical consequence. On the other hand, the lack of observations in other case studies can imply that such changes are in those cases not or not yet very strong.

- The role of public sector funding

A further observation is that the role of the public sector is important to kick-start a data economy. In the culture and environment sector the public sector establishes a data platform on which the private sector can build applications (E-PO-DAT-1), while the necessary enabling role of the public sector was strongly put forward in the smart city case study (E-OO-BM-2). In all these cases the availability or lack of public funding is thereby an enabling or limiting factor. The healthcare case study concerned a big data application and not a platform, but makes clear that the availability of public funding plays a role as well. Similarly, although the data source came from a private platform, the building of specific crisis informatics applications is very dependent on public funding. In the energy and the maritime sector public funding plays a less important role. In these sectors public authorities are important in their regulatory role. Public funding plays a role in terms of their ability to do so, but not in terms of investments in platforms or actual applications. Although not general, we can therefore conclude that public finding plays an important role to provide initial investments to kick-start a data economy.

Conclusions

We can conclude in general that big data delivers a range of positive economic impacts, also as externalities. We can observe improvements in efficiency of existing processes and innovation. This also leads to changes in business models and the appearance of new business models. Such changes are not by definition positive and can lead to dominance of and dependence on a few technological players. On the other hand, even in such a case there can be enough space for smaller niche players, while open source and open data are methods to counter dependence relations. Further, despite these positive economic impacts the role of public funding proves to be important into kick-starting a data economy. This can be explained by the lack of sustainable business models for data platforms and data sharing.

4.2 Social and ethical externalities

We consider first the positive social externalities, which mostly consist of improved or new services, but also issues like transparency and participation. Further we consider negative externalities. On the one hand the positive effects can be diminished or erased through faulty uses of and excessive trust in big data practices. On the other hand big data can also have a negative impact on other important values. Privacy will be treated under the legal externalities. Here we consider the concerns about equality and the potential for discrimination. Further we consider a range of issues related to trust.

- Social benefits

In all the case studies social, or non-economical, benefits were raised. More specific, social benefits raised were improved humanitarian services and safety (crisis informatics), enlarged access to culture and heritage for citizens and researchers (culture), improved human and environmental safety (energy, environment), treatment opportunities and better diagnostics for and a greater understanding of rare genetic disorders, improved natal counselling (healthcare) and improved or optimized resource utilisation, services or security (smart city). (E-OC-ETH-1, E-PC-ETH-2) Such benefits can result indirectly through improved awareness and preparedness or improved decision making (crisis E-PC-TEC-1, environment E-PC-BM-1, E-PC-TEC-1, healthcare E-PC-BM-2, E-PC-TEC-1, smart city E-OC-ETH-1, E-PC-ETH-2). Enhanced transparency and accountability of the public sector, was put forward in the environment case study, as decisions need to be based on measurable evidence (E-PC-LEG-1). Also participation was brought up in several case studies. In the culture case study making cultural data accessible through aggregated metadata is seen as enlarging the participation of citizens in culture. In the environment case study participation is enhanced by democratization of knowledge, which can make communities more prepared and resilient, while also the opportunities for crowdsourcing are enhanced (E-CC-TEC-1). The crisis case study contained a crowdsourcing practice and others were mentioned in the focus group. In the smart city case study participation was seen as enhanced by the development of citizen-centric services with immediate feedback (E-PC-ETH-1).

However, also several potential negative effects putting these beneficial effects at risk were raised. As mentioned already in the part on efficiency, in the crisis and environment the risk of misinterpretation and decisions based on unreliable data (e.g. information from social media) was raised. Both in the environment and the healthcare case study the adverse effect of excessive trust in data-intensive applications was raised. The belief that the environmental dynamics can be captured quantitatively can encourage to overlook important qualitative aspects. In the healthcare sector it can lead to the over-medicalisation of an otherwise healthy population.

Big data practices can also have a negative impact on other important values. We will consider the threat to privacy, which was raised in several case studies, under the legal externalities. Further concerns about equality and the potential for discrimination was raised.

- Equality and discrimination

In the smart city case study it was pointed out that not all citizens will benefit equally from big data. There is the concern about the digital divide, affecting people lacking skills due to age or education or people lacking access due to a low income. (E-OC-ETH-4, E-OO-DAT-4) Therefore debate is needed on how to assure the benefits are shared 'equally enough'. It is important to formulate and define socially desirable outcomes first (E-PC-ETH-1). A similar concern of the effect of the digital divide was raised in the crisis focus group and in the environment case study. Regions with less developed infrastructure and digital skills can be more difficult to integrate in data sharing practices. Similarly, relying on social media favours those people with more access to digital devices and digital and (English) language skills. (E-OC-ETH-4)

Concerns about discriminatory practices were raised in the environment and healthcare case studies. In the environment case study this ranged from the use of data for profiling and targeted advertising to concerns about political abuse and prosecuting of specific groups (E-OC-ETH-4). In the healthcare case study concerns were raised about potential discrimination based on the stratification on genotype, or in relation to health insurance policies. In this context the need for new legal frameworks was felt. But the range of information that can be derived from genetic data on a person's actual and future health condition raises a range of ethical questions very specific to the healthcare case, like how to deal with incidental findings (inform the subject or not) and more generally how such may and must be taken into account in a wide range of decisions (E-OC-ETH-2). Potential political abuse was raised in the crisis case study. Raising attention to people’s tweets or who appears in pictures on social media and, mapping activities can expose people in ways they become more vulnerable or threatened. (E-OC-ETH-9) Such secondary use is also a data protection concern (E-OC-ETH-3).

- Trust

In most case studies trust problems proved to be a barrier or to potentially result in negative externalities. In some case studies concerns were raised about possible exploitation and manipulation. In the environment case study manipulation or the fear for it, due to privacy violations or data abuse, was raised as a problem which could hamper participation and engagement and make social media data less reliable or affect crowdsourcing. (E-PC-ETH-5, E-OC-ETH-11) Such fears can themselves lead to obfuscation or other manipulative practices. Such manipulation also can happen for other motivations (fraud, politics, fun) and affect especially data from social media. (E-OC-ETH-12) Also possible manipulation of visualizations was raised, pointing to a need for specific ethical principles. (E-OO-ETH-1, E-

OC-ETH-7) In the smart city case study the flow of personal data by a range of sensors was considered to raise trust issues, which can turn against participating in the services. On the other hand, computing methods can also be a guarantee against malpractice. Generally trust can become more difficult to establish due to the way big data enables understanding of individuals and optimization of their behaviour. (E-OC-ETH-7, E-CC-ETH-1, E-OO-DAT-4) This can be self-controlled, but can also be perceived as manipulation or exploitation, affecting a person's autonomy. These points back to the already mentioned need to define socially desirable outcomes first, if necessary within the legal framework. (E-PC-ETH-1, E-PC-LEG-3) This potential for negative externalities on the other hand also had the mitigating effect of an increased awareness and more attention for socially responsible and ethical data practices, as was raised in the crisis, environment and healthcare case study. (E-OC-ETH-2).

Trust issues can show up in other contexts as well. The energy case study uncovered it in operator-supplier relations as concerns about the reliability of data of others or from uncontrolled sources. This becomes more important when aggregating data or using data-driven models (E-OO-DAT-4). A similar trust concern about the 'data gap', or the use of data becoming divorced from the context and the knowledge how it is constructed, and the possible resulting misinterpretations, was raised in the crisis case study (E-OC-ETH-13).

We can relate these observations of negative externalities again to the problematic of how the opportunities to capture value from data get translated into organisational structures to do so. Similar to the fear for captive commercial practices or rent-seeking we encounter the fear of manipulation and discrimination. This is linked to a loss of control which can potentially result from big data practices and therefore leads to distrust and reluctance to participate and share data. Again, the key to enable the uptake of big data practices is incorporating safeguards against the negative externalities. This includes restoring transparency and control, which allows to 'internalise' the interests of the data subjects in the big data practices.

Conclusions

In all case studies positive social externalities were reported. These can be seen as similar to improved efficiencies or innovation but for social, non-economical aims. However, also fears for over-reliance on data were raised in a couple of case studies. In several case studies the potential for improved, evidence-based decision-making and/or participation was mentioned.

On the other hand the risk for negative impacts on important social values could also be observed. In most case studies (potential) negative effects on privacy were reported, while several case studies mentioned the risk for equality and new risks for discriminatory practices. In most case studies trust problems can be found, where the risk for manipulation and exploitation leads to distrust and withdrawal, thereby negatively affecting the potential positive impacts of big data. This points to the need for developing practices, including but not limited to legal frameworks, which can assure a proper balance and thereby establish trust.

4.3 Legal externalities

In this section we will consider mainly 2 major legal frameworks, the protection of personal data and intellectual property rights, on which most concerns about legal externalities were raised in the case studies. Other legal issues, like liability and accountability, were raised in a more sporadic and less structured way. Due process problems were not raised as such but are implicit in the discussion on discrimination, which we treated earlier. Jurisdictional issues were raised as a political externality and will be considered in the next section.

- Privacy and data protection

Concerns about privacy and data protection, which is both an ethical as a legal concern, were raised in most case studies, with the exception of the energy case study. In the crisis case study a large potential for violation of data protection and privacy was observed. This concerned mainly the secondary use without consent of personal data, including sensitive data, posted on social media by people themselves or by others (e.g. by posting pictures with other people on it). (E-OC-ETH-3, E-OC-ETH-9, E-PC-LEG-4) This was mitigated through a conscious and developed data protection practice by the research institute and humanitarian organisations (E-OC-ETH-2). As raised earlier, this also entails the risk for data abuse, which can lead to discrimination and political persecution (E-OC-ETH-4).

In the culture case study potential privacy issues resulting from sharing metadata and links to cultural data were noticed, e.g. people on pictures or information about them in metadata, but in practice it proved not to be a major concern. The institute considered the issue in its Terms of Use and promises to take material down when people see a problem. (E-PC-LEG-4) Data protection becomes much more important when the use of interaction data is envisaged.

In the environment case study risks for privacy resulting from big data and concerns about surveillance were raised, implying a need for protection (E-PC-LEG-4, E-CC-ETH-1), while also problems with the data protection framework as a barrier for big data were brought up (E-PO-LEG-2).

In the healthcare case study a similar mix of an awareness of large risks to privacy but problems with the existing data protection framework could be observed. As already noticed, privacy-related risks include a wide potential for discriminatory practices. Similar to the crisis case study we see therefore a well-developed practice in order to deal with the genetic data, involving pseudonymization and access control to data. (E-PC-LEG-4) The ethical and legal concerns mean that no or very limited re-use is made of genetic data, beyond the original purpose of patient care, although such re-use for research would otherwise be very beneficial. But in general re-use is very complicated because of the data protection framework and especially due to the strict consent requirements for sensitive data like health data. Genetic data is also impossible to anonymize and patients with rare diseases are very likely to be identified. (E-OC-ETH-10) The data protection framework is therefore seen as an important barrier to big data research in healthcare. Also in this case study the need to review data protection, taking the needs of medical research into account, was seen as a necessity.

Privacy and data protection are also important issues in the smart city case. Certain data involved, like spatio-temporal data linked to behaviour (e.g. mobility data), is near to impossible to anonymize properly while it also gives an intrusive understanding of a person's behaviour. This results in the need to build trustworthy structures, to counterbalance suspicion and withdrawal by users. (E-OC-ETH-7, E-CC-ETH-1) On the other hand, also in this case study the existing data protection framework was seen as too restrictive. The suggestion was brought up for shifting the attention of protection to a risk-based approach, with more attention to protecting people from risks of abuse and manipulation than protecting their data as such. In particular the legal principles of data minimization and purpose limitation are seen as barriers to big data. (E-PC-LEG-4, E-PC-LEG-3) It was also brought up that machine learning or other algorithms can serve the purpose of data protection or assuring compliance. Further, as machine learning algorithms also prescribe actions, making machine learning algorithms open source or open for audit becomes an important control measure (E-OO-TEC-1, E-OC-ETH-7 , E-CC-ETH-1).

- Intellectual property rights

Intellectual property rights and licensing were raised in the crisis, culture and environment case studies. In the crisis case study the lack of access to data, due to intellectual property rights and/or restrictive licensing conditions (e.g. through the Terms of Use of social media), posed an important barrier to using data from other social media than Twitter (E-PC-LEG- 5). In the culture case study IPR and licensing is a major concern. It can be a barrier to sharing and leads to high transaction costs. On the one hand the copyright situations are often unclear or badly understood, while also setting up licensing arrangements can be very arduous. Both cultural heritage institutions and users accessing cultural data need guidance. On the other hand, the copyright framework is very restrictive in itself. (E-PC-LEG-5, E-PO-LEG-2) Further, the intellectual property framework is still fragmented, a situations that also haunts open licenses (E-PP-LEG-2). Also in the environment case study a lot of problems with the IPR framework were brought up and it is generally considered to be too restrictive, very difficult and therefore costly to maintain compliance with and in need of an overhaul (E-PO-LEG-2, E-OC-LEG-4, E-PC-LEG-5, E-PP-LEG-2). On the other hand, also some concerns about the protection for content creators were raised, especially with platforms like Google or Facebook where users give up their rights (E-OC-LEG-4, E-PC-LEG-5). IPR-related problems were also raised in the healthcare focus group. These concern not the actual research on genetic data done by the research institute, but rather subsequent uses made like in drug therapies, and are therefore of less relevance for the topic of this horizontal analysis. The impact of property rights can also be seen through their absence, especially in the energy case study. The absence of such property rights and the resulting lack of control can lead to a reluctance to share data in a broader context than limited situations regulated by contracts. In other words, here not property rights but rather their absence results in a barrier to big data practices. This was reflected in positioning data ownership as a central concept and in need of additional legislative clarification. (E-PO-LEG-1) On the other hand, in the smart city case study data ownership was considered to be a faulty concept on which to build protection (E-OC-LEG-3).

The objections against both the data protection and the IPR framework were in general pleas for reform, not for abolition. The need and the demand for legal frameworks could be observed, as was remarked above in most case studies concerning privacy.

- Liability and accountability

This need for clarification through regulation or legislation, in order to enable big data practices, can also be observed concerning other legal issues. (E-PO-LEG-1) In the energy case study liability was mentioned as an area needing clarification. The plea in the smart city case study to shift protection from data to protection of subjects against abuse also comes down to build protection more on liability and accountability principles. In the environment case study a more general demand to regulate under which circumstances data can be used and by who, referring to the already mentioned discussion on data ownership but going broader than that. Although not made explicit as such, the issues raised above concerning potential discrimination in the crisis, environment and healthcare case studies point towards the need to assure accountability and due process in terms of procedures to object to certain actions or decisions.

Conclusions

Two major legal frameworks, data protection and intellectual property rights, prove to have an important impact on big data and act as a barrier. Data protection, and to a lesser extent IPR, were raised as important to protect other societal values from negative impacts of big data. But in general both frameworks were considered outdated and too restrictive for big data. Also other areas, like concerning liability and accountability, were mentioned in some case studies to be in need of legal adaptation or clarification.

4.4 Political externalities

In the case studies mostly issues related to political economics came forward as political externalities. These were generally noticed as negative externalities, creating vulnerabilities for the operations or for big data processing in particular. Mainly 2 elements came forward: on the one hand the relation between public sector or non-profit organisations and the private sector, on the other hand the fear to lose control to actors abroad, and in particular US-based actors.

In the crisis case study both these elements came forward. Partnerships between non-profit actors and private actors were also seen as a potential positive externality, as they can provide cost-effective solutions for humanitarian organisations. On the other hand, they can make humanitarian organisations dependent on these private technology providers, who are seen to be unreliable once the crisis situation is past and more interested in brand promotion than in real engagement. Open source solutions are seen as an answer to avoid dependency. The second element comes into play with the fact that most providers are US-based, making the datasets and the information of people contained in them subject to US law. (E-OC-BM-8) A further negative externality, already raised as an ethical externality but also clearly a political externality, is the potential risk for political abuse of these data services and using them as surveillance tools, potentially leading to political persecution (E-OC-ETH-9).

In the culture case study both elements were present as well. The fear to lose control over data held by national institutions was seen as a barrier to cooperation, which translates in a reluctance to share data under a CC0 license as open data but also laws requiring that cultural heritage is stored within the national territory. (E-PP-LEG-1, E-PO-LEG-1) This also translates in a tension around the US dominance over infrastructure. The trend towards outsourcing has been reversed towards developing own infrastructure and tools. Further questions on the relation between public and private sector were raised as well, both in positive terms (cultural data provides economic opportunities through commercial re-use) as in negative terms (rent-seeking on public investments by the private sector and inequalities in data sharing), as has already been discussed above. Lastly, specific for the cultural sector a potential negative externality affecting big data is how partisan politics with public funding can influence its operations.

The protective attitude towards certain data could also be observed in the energy case study. Several states regulate their oil & gas sector differently, making operating internationally for companies difficult. A common feature of these regulations seems to be the requirement that seismic data has to be kept in the country of origin, although the export of copies is normally allowed. (E-PP-LEG-2) The availability of data is an important issue in international project and creates dependencies. Normally this translates in commercial arrangements to buy data. (E-OO-DAT-2) Tensions on dependence on specific private players were not raised as such, seen that outsourcing specialised tasks is a normal business decision. It was observed that certain suppliers are becoming leaders in big data (E-OO-BM-5). As such this changes the 'business ecology' and influences business decision making and strategy.

In the environment case study concerns about dependence on private technological players was also raised. This mostly in terms of over-dependency on centralized services and the lock-in this creates when most applications get developed for that system (Google maps in particular) and impose implicitly a standard. The fact of being US-based was not raised, but rather the issue of dominance of a few big players. (E-OO-DAT-2) More general the concern about privatisation of infrastructure, public knowledge bases and so on was raised, creating a potential barrier for access. A solution was seen in advocating and if necessary imposing open access policies. (E-OC-BM-8) In relation with the state or other public authorities both positive and negative externalities were observed. The availability of environmental data supports decision-making and can make it more evidence-based. It also can improve the knowledge and awareness of citizens and improve their possibilities to hold authorities accountable. (E-PC-LEG-1, E-CC-TEC-1) On the other hand, such data can in specific circumstances be politically sensitive, in particular around disputed or otherwise sensitive regions. Protective attitudes towards data can play, like on natural resources as in the energy case, but more importantly it can be perceived as surveillance from abroad (E-PP-LEG-1).

In the smart city case study the 2 negative externalities merged into one: the dominance or monopoly of US-based technological companies and how this negatively affects the prospects for a European big data economy. A further negative externality posing an obstacle for the European data economy is the fragmentation of legislation. The need to adapt for national regulations makes scaling over a European market difficult. A clear need for harmonization of data legislation was put forward. A positive element mentioned, where policy has played an important role, was the fact that Europe has strong physical infrastructures and is much more connected. (E-OC-LEG-1, E-OC-LEG-2) A further positive externality is the potential for better decision-making and service delivery through more citizen-centric services and politics with immediate feedback (and that way resulting in a variant on evidence-based decision-making) (E-PC-ETH-1).

Political externalities played out differently in the healthcare sector. Both elements were not raised in this case study. On the one hand positive externalities were seen in the potential for improved decision-making in health-related issues. On the other hand, the earlier discussed potential for discriminatory practices based on health data create a real need to develop policies on these issues. (E-PP-LEG-3)

Conclusions

In the case studies mostly issues related to political economics came forward as political externalities. These were generally noticed as negative externalities, creating vulnerabilities for the operations or for big data processing in particular. Mainly 2 issues came forward: on the one hand the relation between public sector or non-profit organisations and the private sector, on the other hand the fear to lose control to actors abroad, and in particular US-based actors, which sometimes translates in protectionist attitudes and requirements to store data within national territories. However, the attitudes towards and perceptions of these 2 issues varied among the case studies. Further, in several case studies, the potential for better, evidence-based decision-making was raised. Other political externalities were rather specific or only raised in one or two case studies, like the risk of data being re-used for political abuse or surveillance (crisis, environment) or the need for considering and regulating wider implications of big data (genetic data in the healthcare case).

5 CONCLUSIONS

The horizontal analysis of the case studies shows that big data does not stand for the same practice in every sector, but covers a wide variety of datasets and data practices and the technical challenges these present. Volume is present in all case studies, but big can still be quite different in scale. Velocity is not present in all case studies and some datasets are relatively static, but gets introduced when big data is based on data acquisition through sensors or on user interactions. Variety proves to be a challenge in all case studies, resulting from the combination of different sources or from the multidimensionality or unstructured character of the data. Veracity often captures a wider range of problems like validity, data granularity or resolution, uncertainty about meaning. These issues show up differently in the case studies. Mapped across the Big data value chain the technical challenges raised in the case studies are mostly observed in the data collection, data curation and in a lesser extent in the data usage phase. Often these challenges are not purely technical challenges, but the translation of societal externalities.

The comparison of the societal externalities raised shows a range of positive economic and societal impacts, also as externalities. The specific look at externalities can deform the overall picture, as in some cases specific results can be observed in internal processes while in others they show up as externalities. The creation of data value chains is often dependent on obtaining data from a wide range of actors and therefore penetrates organisational boundaries. As we noticed that it is important to evaluate how such organisational borders are affected and possibly redrawn, it is important to take a broad perspective and not limit the analysis to externalities.

We have observed positive economical externalities in terms of innovation and in improvements in efficiency. This also leads to changes in business models and the appearance of new business models. Such changes are not by definition positive and can lead to dominance of and dependence on a few technological players. On the other hand, even in such a case there can be enough space for smaller niche players, while open source and open data are methods to counter dependence relations. Further, despite these positive economic impacts the role of public funding proves to be important into kick-starting a data economy. This can be explained by the lack of sustainable business models for data platforms and data sharing.

In all case studies positive social externalities were reported. These can be seen as similar to management efficiencies or innovation but for social, non-economical aims. However, also fears for over-reliance on data were raised in a couple of case studies. In several case studies the potential for improved, evidence-based decision-making and/or participation was mentioned. On the other hand the risk for negative impacts on important social values could also be observed. In most case studies (potential) negative effects on privacy were reported, while several case studies mentioned the risk for equality and new risks for discriminatory practices. In most case studies trust problems can be found, where the risk for manipulation and exploitation leads to distrust and withdrawal, thereby negatively affecting the potential positive impacts of big data. This points to the need for developing practices, including but not limited to legal frameworks, which can assure a proper balance and thereby establish trust.

Two major legal frameworks, data protection and intellectual property rights, prove to have an important impact on big data and act as a barrier. Data protection, and to a lesser extent IPR, were raised as important to protect other societal values from negative impacts of big data. But in general both frameworks were considered outdated and too restrictive for big data. Also other areas, like concerning liability and accountability, were mentioned in some case studies to be in need of legal adaptation or clarification. Jurisdictional aspects appeared more as political externalities.

In the case studies mostly issues related to political economics came forward as political externalities. These were generally noticed as negative externalities, creating vulnerabilities for the operations or for big data processing in particular. Mainly two issues came forward: on the one hand the relation between public sector or non-profit organisations and the private sector, on the other hand the fear to lose control to actors abroad, and in particular US-based actors, which sometimes translates in protectionist attitudes and requirements to store data within national territories. However, the attitudes towards and perceptions of these two issues varied among the case studies. Further, in several case studies, the potential for better, evidence-based decision-making was raised. Other political externalities were rather specific or only raised in one or two case studies, like the risk of data being re-used for political abuse or surveillance (crisis, environment) or the need for considering and regulating wider implications of big data (genetic data in the healthcare case).

The overall picture shows positive benefits but also the potential to negatively affect other important social or ethical values. Important is that big data is not just a technical issue but has an impact on organisational borders and the 'business ecology' in general. This leads to uncertainty and conflict in a range of areas, translating in distrust and reluctance by all sorts of actors and conflicts on political and legal level. Organisational borders need to be redefined or redrawn, while also social norms and legal frameworks need to be clarified again based on a proper balancing of all interests.

APPENDIX 1: CLASSIFICATION OF EXTERNALITIES AND CODED OBSERVATIONS

A. Economic externalities

- improved efficiency
E-PC-BM-2 crisis, environment, healthcare
E-PC-BM-3 crisis
E-PC-BM-4 crisis
E-PC-TEC-1 crisis, environment, healthcare
E-OC-ETH-12 environment
E-OC-ETH-13 crisis
E-OC-TEC-2 smart city
E-OO-BM-7 environment

- innovation
E-PC-BM-2 healthcare
E-PC-DAT-1 culture, smart city
E-PC-TEC-2 culture
E-OC-DAT-1 environment
E-OC-DAT-2 crisis, smart city
E-PO-BM-1 energy, environment
E-PO-BM-2 culture, environment, healthcare
E-PO-LEG-2 culture, environment
E-OO-BM-1 energy, environment
E-OO-BM-2 energy, smart city

- changing business models
E-OO-BM-1 energy, environment
E-OO-BM-2 energy, smart city
E-OO-BM-3 energy, smart city
E-OO-BM-5 energy, smart city
E-OO-DAT-1 environment
E-PO-BM-1 energy, environment
E-PO-BM-2 culture, environment, healthcare
E-PO-DAT-1 culture, environment
E-OC-BM-5 smart city
E-OC-BM-7 environment
E-OC-BM-8 crisis, environment

- employment
E-OC-BM-3 energy, smart city
E-OC-BM-5 smart city
E-PO-BM-1 environment

- role public funding
E-PO-DAT-1 culture, environment
E-OO-BM-2 smart city

B. Social and ethical externalities

- beneficial impacts due to improved efficiency and innovation
E-PC-BM-1 environment
E-PC-BM-2 crisis, environment, healthcare
E-PC-BM-3 crisis
E-PC-ETH-2 smart city
E-OC-ETH-1 crisis, energy, environment, smart city

- improved awareness and improved decision-making
E-PC-TEC-1 crisis, environment, healthcare
E-PC-BM-1 environment
E-PC-BM-2 crisis, environment, healthcare
E-PC-BM-3 crisis
E-OC-ETH-1 smart city
E-PC-ETH-2 smart city
E-PC-LEG-1 environment

- participation
E-PC-ETH-1 environment, smart city
E-CC-TEC-1 environment

- privacy (see legal)

- equality
E-OC-ETH-4 crisis, environment, smart city
E-OO-DAT-4 smart city

- discrimination
E-OC-ETH-2 crisis, environment, healthcare
E-OC-ETH-4 crisis, environment, smart city
E-OC-ETH-9 crisis

- trust
E-PC-ETH-1 environment, smart city
E-PC-ETH-5 environment
E-OC-ETH-2 crisis, environment, healthcare
E-OC-ETH-7 environment, smart city
E-OC-ETH-11 environment
E-OC-ETH-12 environment
E-OC-ETH-13 crisis
E-OO-ETH-1 environment
E-OO-DAT-4 energy, smart city
E-PC-LEG-3 smart city
E-CC-ETH-1 environment, smart city

C. Legal externalities

- data protection and privacy

E-PC-LEG-3 smart city
E-PC-LEG-4 crisis, culture, environment, healthcare, smart city
E-OC-LEG-3 smart city
E-OC-ETH-2 crisis, environment, healthcare
E-OC-ETH-3 crisis
E-OC-ETH-4 crisis, environment, smart city
E-OC-ETH-7 environment, smart city
E-OC-ETH-9 crisis
E-OC-ETH-10 healthcare
E-CC-ETH-1 environment, smart city
E-PO-LEG-2 culture, environment
E-OO-TEC-1 smart city

- intellectual property rights
E-PC-LEG-5 crisis, culture, environment
E-OC-LEG-3 smart city
E-OC-LEG-4 environment
E-PO-LEG-1 culture, energy, environment, smart city
E-PO-LEG-2 culture, environment
E-PP-LEG-2 culture, energy, environment

- liability, accountability
E-PO-LEG-1 culture, energy, environment, smart city

D. Political externalities

- relations private sector vs. public and non-profit sector
E-OO-DAT-2 energy, environment
E-OO-BM-5 energy, smart city
E-OC-BM-8 crisis, environment

- losing control to actors abroad
E-OC-BM-8 crisis, environment
E-PP-LEG-1 culture, environment
E-PP-LEG-2 culture, energy, environment
E-PO-LEG-1 culture, energy, environment, smart city
E-OC-LEG-1 smart city
E-OC-LEG-2 smart city

- improved decision-making and participation
E-PC-LEG-1 environment
E-CC-TEC-1 environment
E-PC-ETH-1 environment, smart city

- political abuse & surveillance
E-PP-LEG-1 culture, environment
E-OC-ETH-9 crisis

APPENDIX 2: OVERVIEW SOCIETAL EXTERNALITIES

Externalities	+/-	Crisis Informatics	Culture	Energy	Environment	Healthcare	Smart city
Economic externalities							
improved efficiency	+	+			+	+	+
innovation	+	+	+	+	+	+	+
changing business models	+/-	-	+/-	+/-	+/-		+/-
employment	+/-			+	+		+/-
Dependency on public funding	-	-	-		-	-	-
Social and ethical externalities							
improved efficiency and innovation	+	+/-	+	+	+/-	+/-	+
improved awareness and decision-making	+	+			+	+	+
participation	+	+	+		+		+
equality	-	-			-		-
discrimination	-	-			-	-	
trust	-	–/+		-	-	–/+	–/+
Legal externalities							
privacy	-	-	-		-	-	-
IPR	-	-	-		-	-	
liability, accountability	-	-		-	-	-	-
Political externalities							
private vs. public and non-profit sector	-	-	-		-		-
losing control to actors abroad	-	-	-	-			-
improved decision-making and participation	+				+	+	+
political abuse & surveillance	-	-			-		