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 D1.2:
Big Data Policies

Author(s):	Hans Lammerant, Antonella Galetta, Paul De Hert (VUB), Lorenzo Bigagli, Paolo Mazzetti (CNR), Stéphane Grumbach (INRIA)

Dissemination level:	Public
Deliverable type:	Final
Version:	1.0
Submission date:	Due 31 August 2014

Table of Contents

EXECUTIVE SUMMARY

This deliverable reviews policies concerning access to, linking of and (re-)use of big data. It explores, through a broad comparative scope, how private and public law frameworks have impact on the space for big data collection and processing and how policies of public and private actors further shape the data economy.
We have used a twofold understanding of policies. First, policy as governmental strategies and action plans, in this case governmental strategies concerning big data processing. Secondly, policies as regulation of data flows and the access and use of data. Big data processing refers to the possibility to obtain, combine and analyse diverse and large datasets from different sources. Access and use policies concerning data shape the space available for big data processing and how it can be done. In this context we look at legal frameworks that define or have an impact on such access and use of data. The selection criterion in our review of this policy landscape focused on whether the policies in question concerned the legal frameworks regulating access and use of data.
The main legal frameworks that have an impact on access, linking and re-use of data are intellectual property rights and licensing, privacy regulations and the regulation of access and use of public sector information. None of them is specific to big data processing. But big data poses challenges to these legal frameworks or changes their role.
Information held by public authorities is subject to a specific regime concerning access and use. The function of these regimes is changing from a transparency tool to an enabler of the data economy through open data policies. This shift also moves the attention from mere availability of information to interoperability of data sources. It widens the scope of issues to include elements concerning the usability and quality of the data.

The application of copyright to data and big data is not straightforward. Intellectual property rights create data enclosures, but it remains unclear when data is protected and which uses fall under the exclusive rights. Restrictive implementations of exceptions to these rights make data mining very difficult. Copyright law offers a default regime on which private actors can shape their relations through licenses. However, also the application of contract law on licenses in the digital market proves to be unclear. And the patchwork of jurisdictions makes it difficult to create a stable environment for the combination of diverse sources of data. This problem also affects open content licenses, the tools developed to enable easy access and use of data.
Private actors can shape their own access and use policies concerning the data they control, using the private law tools of IPR and contract law. Access and use gets controlled with terms of use and privacy policies. The comparison of the terms of two important global actors, Twitter and Google, revealed how both shape their own distinct and partially open, partially closed data ecosystem in a way favourable to their business model. These business models define their platform politics or the organizational interoperability they allow with their services. Through their terms they define how the data under their control can be linked up to provide other services.
The same private law tools are used for making open content licenses. These licenses do not aim to shape data enclosures but to open their data. They also use intellectual property law but to waive the protection offered by this law. Open content licenses are instrumental in shaping a much more open data ecosystem. However, fundamentally they attempt to fix the problems created by an unadapted intellectual property framework, which through too strong protections leads to enclosures.
These two opposite approaches clearly demonstrate how big data processing is dependent on legal interoperability of data sources and how the legal framework can be used to define the

data ecosystem in which such big data processing can take place. It also demonstrates how this is linked with organizational interoperability. Big data companies use the legal tools to allow uses fitting their business model while locking other out. Open data policies use open content licenses to build a much wider ecosystem and allow a wide range of business processes, in order to support positive network effects from diverse data sources.

The European Commission developed a broad vision on the data economy in its Communication 'Towards a thriving data-driven economy'. It puts forward the central role of (big) data in the knowledge economy. Early policy concerning ICT was often reactive and concerned the adaption of old economic models and their regulation to new technology or treated the internet sector as a distinct economic sector. The development and adoption of digital technology has now reached a stage that it leads to changes in business processes and organization in more traditional sectors of the economy and government. This gets reflected in more comprehensive visions on the data economy and data value chains. The real value of big data is not just a more developed technology, but that it enables to reap positive network effects from combining and re-using data sources.
A similar comprehensive vision can be found in the European Interoperability Framework (EIF), with its definition of interoperability levels: legal, organizational, semantic and technical interoperability. Making data usable and building data value chains is not just a technical question of big data processing, but includes overcoming legal barriers and adding precise definitions of meaning to data through vocabularies and metadata. Organizational interoperability concerns how organizations cooperate to achieve their goals. They need to rebuild their business processes around the data exchange.
The action plan in the data economy communication also lists some regulatory issues, which include privacy and intellectual property rights. The European IPR framework clearly poses problems for big data processing. The exceptions in the copyright framework can be problematic for data mining. It can also be questioned if the sui generis protection of databases is not over-protective and harmful for the data economy.
The legal framework on access to documents and re-use of PSI reflects the shift of a transparency tool towards market regulation and the opening up of PSI as an economic resource. Through promoting open data policies the EU stimulates the Member States to participate in this data ecosystem and to play an enabling role in the data economy.
The current European approach to data protection clearly poses challenges to big data processing and vice versa. Big data processing puts into question the sustainability of the basic concept of data protection: personal data, being distinguishable from other, non-personal data. Also the application of the data protection principles in actual big data processing proves difficult. However, the response of the data protection authorities shows that the data protection principles can be applied on big data processing without such contradiction.
Data protection creates a patchwork of jurisdictions, which can be a barrier to combining diverse sources in big data processing and lead to distinct data ecosystems. Methods exist to overcome these barriers. The Safe Harbor-regime is an important example thereof, as well as how through such brokerage elements of EU law get a wider territorial application. Binding corporate rules are another instrument to eliminate barriers with a corporate group. These methods lead to a re-ordering of the borders of the data ecosystem and are therefore very important for big data processing.

We reviewed a sample of sectorial policies as well as strategies in member states. The INSPIRE-directive, regulating the access, re-use and sharing of geospatial data, reflects the data ecosystem approach as it concerns more than data availability. The starting point is the

sharing of data between public services, but it includes this in a wider data ecosystem with the private sector and general public by defining the available network services. The directive also reflects a broad approach to interoperability, with strong attention to semantic interoperability and usability of the data. The sector of geospatial data can therefore be considered as a testing bed for an approach to data that got reflected in the EIF.
Each of the reviewed institutions that produce large datasets (GMES/Copernicus, CERN, ESA, EUMETSAT) has adopted open access policies in general or partially, although sometimes limited for security concerns or IPR restrictions. All are data providers for a wide range of purposes and include a strong attention for semantic interoperability.
Big data policies vary a lot across Member States. The UK is a trendsetter in all aspects of big data policy. Although it has no distinct big data policy document, we find that all elements of the EU policy discussion are raised or were influenced by the UK discussion. This is surely valid for the IPR and privacy discussions. Similarly Germany is pioneering on the industrial use of big data and IoT. It is also an important player in the big-data related discussions on data protection.
In the other EU-countries considered we did not find such in-depth discussion. In France, big data figures as part of an industrial innovation strategy. As a result, it has recently developed a big data plan, which includes an evaluation of its regulatory framework concerning big data. Italy has a lot of attention to big data in the context of data protection and on the level of research infrastructure. The smaller countries Belgium and Sweden do lack attention to big data in their policies.
Open data policies shows similar variation. Open data policy in the UK has passed its start-up phase and attention widens to include data quality. The open data policy is embedded in a clear vision on how it has to support the development of a data economy. This resulted in the user-driven approach to create the data ecosystem. France also has a well-developed open data policy with user consultations.
Such clear vision is in general lacking in the open data policies in the other countries. Interesting is the pioneering role the INSPIRE-framework has in developing the experience to develop such vision.
Legal interoperability has been addressed in all countries except Sweden, as the governmental licenses are all made compatible with the Creative Commons license.

A review of US legal frameworks shows clear differences in the IPR framework and the privacy regulations, which make big data operations easier in the US compared to the EU. Legal frameworks concerning access and re-use of PSI are relatively similar in the EU and US.
The copyright framework in the US has fewer problems with new technologies, including big data. It has a weaker protection of databases compared to the EU, which makes re-use of data easier. And its fair use-system of exceptions to copyright protection is more compatible with data mining. Licensing has similar problems as in the EU when it becomes more than a mere authorization and turns into a contract.
Open data policies have become a policy topic around the same period in both regions. In the US the rationale was more focused on transparency and public control of the government, where the EU policy takes clearly an economic view. The data is seen as an enabler of democratic control, not of business processes. The result is that open data functions less as part of a market policy or a vision concerning value creation. In the policy documents on open data, we find strong attention for semantic interoperability, but much less for involving stakeholders.
Privacy law consists in the US of sectorial regulations. A general data protection framework does not exist. Important also is that such privacy regulations are, in general, a part of

consumer protection. The constitutional protection of privacy only applies to governmental intrusions of privacy and not to those of private actors. Result is that data (re-)use gets much less limited by privacy law compared to the EU. Of course this wider freedom comes with a price. Privacy concerns have led to recent high-level attention for privacy risks linked with big data. Although the EU is generally seen as too strict in its approach to privacy, the Consumer Privacy Bill of Rights would introduce a much more general privacy protection and bring the US privacy regulation much nearer to that in Europe. In the academic and political debate we see the same doubts raised about the compatibility of Fair Information Practice Principles with big data processing.

A review of data policies of other third countries (Australia, China, Japan) show similar discussions and evolutions in the two countries (Japan, Australia) with a well-developed big data strategy. Both make work of the framework conditions for big data. IPR law is changed or changes are under discussion. Also the tension between big data processing and privacy law gets raised. China seems to run behind in the development of legal frameworks and privacy, although it has an important home market of data and homegrown industrial players. Open data is in all three countries a clear policy choice but its implementation is in a start-up phase.

In general, this review finds that big data policies are very much in a developmental phase, both in Europe and around the world. While some countries and sectors are more advanced than others, no specific context has a comprehensive policy on big data that addresses all of the important aspects identified here. Furthermore, this report also finds that big data policy is a complex arena that includes elements associated with intellectual property, privacy and data protection, open data and economic development. As such, it raises a question as to whether such a comprehensive policy is possible at the moment. Considering big data developments across a range of different policy areas may be the most effective way to ensure that the possibilities and pitfalls of big data are considered across the European policy landscape. But it remains a necessity to ensure that localized approaches do not lead to limited tunnel visions, and are brought in contact with each other. The fact that formulating a comprehensive policy is difficult points to the need to set up areas for more comprehensive policy learning.

1 INTRODUCTION

This deliverable will research policies concerning access to, linking of and (re-)use of big data.

Big data is often characterized by the 3V-definition presented by Gartner: volume, velocity, variety. These Vs present how traditional database and data processing techniques are challenged by big data. However, these are the technical challenges. Additional hurdles to big data processing can be erected by legal frameworks, private regulation and government policies. These hurdles are not specific to big data, although big data processing can create specific problems not foreseen by these legal frameworks.
Another approach to discern these hurdles is looking at how they influence the major steps in the data value chain. The data value chain has been characterized 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]

Again, these steps are not specific to big data, but big data will require specific technical solutions in all these stages. Also, all these steps in the data value chain can be influenced by access and use policies concerning data.
Hurdles can present themselves in the data acquisition phase by limiting access to or disallowing the combining of data. Such access can be limited by intellectual property rights (IPR) or contractual arrangements. Open data policies can, on the other hand, make data available for large-scale analysis. Also, other legal frameworks like data protection limit the collection of data. These frameworks also have an influence in the later stages. IPR can be an obstacle for large-scale data analysis like data mining. Data protection influences how personal data can be used, while also posing specific requirements and constraints concerning the curation and storage of this data. Open data policies also focus on improving the usability and interoperability of the data by adding metadata.
This deliverable will review how such policies, reflected in legal frameworks, private regulation and government policies, support or hinder big data processing.

This demands some methodological delineations concerning what qualifies as a relevant policy.
Researching governmental strategies concerning big data processing raises immediately some issues. Big data is a very recent term and strategies specific to big data are still very rare. On the other hand, limiting ourselves to these rare strategies mentioning the word 'big data' would provide a too limited view on policies which affect big data processing. Therefore we have used a broader, twofold understanding of policies. First, policy as governmental

[1] Becker, Tilman, et al, *D2.2.2 Final Version of Technical White Paper*, Big Data Public Private Forum (BIG), 14/05/2014, pp. 1-2

strategies and action plans, in this case governmental strategies concerning big data processing. Secondly, policies as regulation of data flows and the access and use of data. Big data processing refers to the possibility to obtain, combine and analyse diverse and large datasets from different sources. Access and use policies concerning data shape the space available for big data processing and how it can be done. In this context we look at legal frameworks which define or have an impact on such access and use of data. Often such legal framework are not specific for big data, but concern the use of data and information in general. But big data processing has to take them into account, is impacted by them or raises specific issues concerning these frameworks. For instance, access policies to government-held information are often developed with citizens and journalists asking documents as target public, but this also affects the availability of such information for big data processing.
Therefore in this deliverable we reviewed legal frameworks concerning access to, linking of and (re-)use of data and considered governmental strategies when and for so far they concern these legal frameworks. We reviewed those strategies for their impact on big data processing, and where relevant, how the advent of big data changed these policies or engendered specific policies.

Further, big data points to a specific technology, but which builds upon others (like cloud computing as enabling technology), while the term itself is not clearly defined. Therefore policies concerning these enabling technologies are also relevant. Policies on big data build often on earlier policies concerning cloud computing, open data or are part of wider strategies concerning the digital economy. Such strategies are often much more comprehensive and wider than the issue of access to, linking and (re-)use of data. For instance they address issues like consumer protection in the digital market or the necessary skill base. Or the shift to big data policies points to a shift in attention. E.g. cloud computing is an important enabling infrastructure for big data processing, but the shift from policy plans on cloud computing to plans on big data points to a shift from attention for this specific infrastructure to attention for availability of data, data analysis capacity or for data as central element in value creation. In order to make a clear selection of policies, our criterion in reviewing the policy landscape were policies which concerned or affected the legal frameworks regulating access and use of data.
Our review ended on 1 August 2014, although some country studies (Germany, Italy and Sweden) were added later and contain material up to 1 July 2015.

In the next chapter we give a short and general overview of the legal frameworks regulating the access to, linking of and (re-)use of data. In order to assure this deliverable can be read on its own, this will contain some unavoidable overlap with the consideration of legal issues in D2.1. On the other hand, chapter 2 presents only the basics of the relevant legal frameworks while D2.1 will go into more depth concerning the problems big poses for the application of these frameworks and vice versa.
In the following chapters 3 to 5 and 7 we review governmental policies on big data and concerning these legal frameworks. We also point out relevant differences in the legal framework when necessary to understand the specific policies. Chapters 6 concerns policies of private actors. We show how these policies are shaped within the reviewed legal frameworks. On the one hand, we consider some large big data companies (Google, Twitter) and review how they regulate the access to their big data assets. On the other hand, we show a different approach regulating such access by open content licenses.

2 OUTLINE OF THE LEGAL FRAMEWORK OF THE DATA ECONOMY

In the first part we consider how private law affects access and use of data. This concerns in the first place intellectual property rights (IPR) and further also how right holders can shape such access and use further through contract law.
In the second part we look to relevant public law regulations. One important framework concerns the access to and re-use of public sector information, which also provides the legal framework for open data policies. Further privacy law and data protection regulates how personal data can be collected and used. Both have an important impact on big data processing and on the space within which a data economy can develop.

2.1 INTELLECTUAL PROPERTY RIGHTS AND CONTRACTUAL RELATIONS

A first level influencing the possibility to combine data sources into big data are intellectual property rights. Intellectual property rights does provide a default property regime regulating access to and control of data. Intellectual property law is a public regulation mechanism defining the playing field for the market in information and data. From this default regime rights holders can deviate through contract law. Through licenses they can give permission to access and use their data, but also add a range of conditions and limitations to such use.

The applicability of intellectual property rights on datasets and on uses of data is not straightforward. In principle, when a right holder has exclusive rights on data or the database certain activities with those data cannot be done without the authorization of the right holder. This results in 3 questions. Is the data and/or the database protected under one of the intellectual property rights? And, is the use made of the data part of one of the exclusive rights of the right holder? The answer to this last question is further fine-tuned by a range of exceptions to these exclusive rights, which can give space for a specific use without that a preliminary authorization by the right holder is needed (e.g. the right to quote). The legislator can thereby determine strongly how a data economy can function by the extent of the property rights it grants on data. Granting property rights affects the legal interoperability of datasets as their use requires a preliminary authorisation.

The most relevant intellectual property rights to look at are copyright and database rights. Intellectual property rights are defined in national laws, but have been harmonized through international treaties which define a minimal scope of the rights and protection mechanisms. The main treaty on copyright is the Berne Convention, originally from 1886 but which has been regularly revised. The WIPO Copyright treaty from 1996 and the TRIPS-treaty from 1994 refer to the Berne convention for copyright, but also specify a related copyright protection for databases as collections.
Important is that copyright does not protect data as such. In general copyright protects intellectual creations and expressions, but not facts or ideas. It is therefore linked with an originality requirement defining which level of intervention by the creator is needed, before copyright protection is granted. The international treaties leave it up to the parties to define this threshold or to define extra forms of protection, and here is where policy choices are visible.
We will see that the EU has chosen to give a much stronger protection than the US or other countries. The US has kept copyright protection on data and databases limited to those which can fulfil the originality requirement. It refused to give protection by copyright or a specific

IPR to investments in databases without any originality. The EU on the contrary has chosen to grant extra protection to databases based on investment. This difference in protection is further augmented by the much stricter regime of exceptions in the EU compared to the US.

Linked to this regime of exceptions is the possibility to deviate from it by using contractual arrangements or licenses. It is important to keep in view the differences between property law and contract law. The property owner is the holder of certain rights concerning a good, which give him the rights to exploit this good at exclusion of others. The ownership is a legal fact anyone else has to respect and is limited by. The holder of property rights can claim his rights and enforce them against anyone else. The holder of property rights can transfer these property rights, for instance by selling them. Who buys these property rights becomes the new owner and can now enforce his claims not just against the former owner but also against everyone else. Property law forms a specific set of rights for which by public law a default regime has been created. The public law character is the source for the possibility to claim respect for these rights against everyone. However, it is also possible to allow certain uses of property rights, without transferring these, to someone else by contract. One example is the right to access and use of a good for a limited time and under certain conditions. Contract law offers a mechanism to change the default rule of copyright law and to create private ordering deviating from this default regime. But the private ordering through contract law has other characteristics than property law. Contract law is based on an agreement between partners establishing legal obligations between them. Contrary to property law, these obligations are only binding for the parties in the contract, not for others. And by consequence, these obligations can only enforced to the parties in the contract.

Property law has been conceived for the material world. Also intellectual property law was originally mainly based on the linkage of an intellectual product with a material carrier. Digitalization has subverted the functioning of this legal framework. The fact that a copy of a digital product can be created at no or negligible cost broke the natural barrier turning an intellectual product into a commodity. Commodification formed a natural enclosure upon which intellectual property rights could be grafted.[2] Attempts have been made to adapt the intellectual property frameworks to the digital world by developing specialized intellectual property rights like software copyright, database rights, etc. But also another evolution was visible, turning products into services and making them available through restrictive licenses. This mechanism of end user license agreements (EULA) is best known for software, but can also be used for sources of data like online databases. Both evolutions turned out to be much more restrictive than the traditional intellectual property law. For instance, traditional copyright and patent law contain the exhaustion of intellectual property rights. Once sold, the patent holder cannot restrict the further sale or the use of the material product containing the intellectual product (e.g. a book, CD, computer, car, ...). But in the digital environment it is much more difficult to make a distinction between making a new exemplar of a product and transferring the product in a second hand sale. Similarly with allowing access to other users. Therefore exhaustion was cancelled in the adaptation of IPR to the digital environment. Both the new forms of intellectual property rights and the licensing practices prove to be much more restrictive.
This evolution sparked another reaction: the use of licenses to open up content. Open content licenses try to counter the restrictive nature of IPR and its automatic application. Instead they try to guarantee access to the content without the need to ask authorisation for each use. To

[2] De Filippi, Primavera. *Copyright Law in the Digital Environment: Private Ordering and the regulation of digital works*, 30 June 2012, pp. 23-26. http://halshs.archives-ouvertes.fr/docs/00/71/34/03/PDF/Thesis_-_Private_Ordering_-_De_Filippi.pdf

do so they also use contractual mechanisms to deviate from the default rules of intellectual property law.

Both IPR and licensing mechanisms are the building blocks through which private actors can regulate the access and use of their data assets by others. Legislators and policy makers can influence the shape of the data economy through the legal protection they give to data and databases through the IPR framework and the space they leave for licensing mechanisms.

2.2 PUBLIC LAW REGULATIONS

The access and use of data is also influenced by public law regulations. We limit ourselves in this part to a short overview and treat these regulations in more detail in the specific chapters.

A first important regulatory area concerns the collection and use of personal data. The protection of personal data has evolved a lot as an answer to technological evolutions and the resulting societal changes. Privacy torts were an answer to the emergence of mass media. The advent of databases in the computer age received a regulatory reaction in data protection laws. The major question is now if the current data protection principles can still cope with big data processing. In this deliverable we will look at how data protection frameworks are shaped in distinct jurisdictions and which responses have been given so far to big data.

Among the regions and countries considered the legal frameworks dealing with personal data are very different in their foundations and grounded in different constitutional cultures. This results in a quite different environment to deal with big data.
Protection of personal data is based on the fundamental right to privacy, but has evolved into a framework of rights and duties which exceeds the right to privacy and has acquired in Europe the status of an autonomous fundamental right in itself. Both rights do partially overlap, but function with a different logic. Gutwirth and De Hert point to two distinct constitutional or legal tools to limit and control power.[3] One set of constitutional tools are opacity tools, which set limits to the interference of power in individual matters. These tools shield certain areas and prohibit the interference of power. The other set are transparency tools, which guarantee transparency and accountability of the powerful. These tools regulate and organise the exercise of power, in order for it to be legitimate. The right to privacy can be seen as an opacity tool, while the right to protection of personal data is a transparency tool. In practice these distinctions are not absolute, as also the data protection framework contains opacity elements. But in general both rights function according to different logics. Both set of tools get used in very different ways in the EU and the US. Main difference is that in the EU all processing of personal data needs a legal ground, where in the US such processing is free except when some specific legislation forbids it or subjects it to specific rules.

The general European data protection framework is provided by directive 95/46/EC[4], but it is rooted in earlier instruments like *Convention for the Protection of Individuals with regard to Automatic Processing of Personal data* (also known as Convention 108), adopted by the Council of Europe in 1981. The data protection framework had a profound impact on the fundamental right jurisprudence concerning the right to privacy. The right to protection of personal data developed into a fundamental in itself, distinct from the right to privacy. This

3 Gutwirth, S., De Hert, P., Regulating Profiling in a Democratic Constitutional State, in Hildebrandt, M. and Gutwirth, S. (eds.), *Profiling the European Citizen: Cross-Disciplinary Perspectives*, Springer Science + Business Media B.V. 2008, 271-293

4 European Parliament and the Council, Directive 95/46/EC on the protection of individuals with regard to the processing of personal data and on the free movement of such data, 24.10.1995

data protection framework provides that all processing of personal data requires a legal ground. In other words, all processing of personal data is regulated and subject to a set of rules guaranteeing the accountability of the processor and the transparency of the processing. The European data protection framework applies to all processing of personal data. Personal data is defined very broadly as "any information relating to an identified or identifiable natural person". Also the range of activities to which the directive applies is very broad. Processing is defined as "any operation or set of operations that is performed upon personal data, whether or not by automatic means". This means that whenever data in a big data-context contains information linked to an identifiable natural person, the processing has to be according to the data protection principles and mechanisms have to be implemented to allow data subjects to exercise their rights. The only possibility to escape this framework is by anonymisation of the data.

The US framework does not subject all processing of personal data to legal rules guaranteeing more transparency and control for data subjects. Personal data can be freely used unless it is forbidden. The basic structuring of the legal framework is based on opacity tools. The 4th Amendment to the US Constitution protects people "in their persons, houses, papers, and effects" against the government. Searches are only allowed with a warrant and upon probable cause. This 4th Amendment protection only applies towards the government and not towards private actors. Outside this limited area processing of personal data is in principle allowed, except when specific laws forbid it or subject it to certain rules. Privacy law between private actors was first established through tort law. Four privacy tort actions are recognized in the Second Restatement of Torts and can be considered as opacity tools between private actors, but these have no practical relevance for big data.
This does not mean that data protection has no place in US law. The growing use of computers and the surveillance scandals from the Nixon and FBI director Hoover-era led to the formulation of Fair Information Practice Principles (FIPP).[5] These FIPP are similar to the principles underlying data protection in Europe, but have only been put into law in specific areas.

This comparison shows a fundamentally different situation in which big data processing using personal data can take place. This US constitutional framework gives free space for such big data processing, as long as no other specific law provides constraints. The EU framework does only allow unconstrained big data processing with anonymized data. When using personal data, big data processing has to be able to fulfil the requirements of data protection law.

Another important area is the regulation of access and use of public sector information (PSI). Governmental authorities and institutions collect and produce a lot of data as part of their functioning. But this data can also be useful for a lot of other uses by both the public and private sector. Freedom of information or access to documents-regulation do exists already for a long time, as part of transparency and public control of the functioning of the government. Also governments have produced statistics to inform their policies.
But the advent of the data economy has also brought the economic value of the information held by governments under attention. The transparency aim remains present in the open government policies, but economic objectives like more governmental efficiency and the creating of a local data economy are important drivers. This has shifted the attention to release of data on a larger and regular scale, contrary to control through specific requests. The

5 US Department of Health, Education & Welfare, *Records, Computers and the Rights of Citizens, Report of the Secretary's Advisory Committee on Automated Personal Data Systems*, July 1973.

changes in legal frameworks reflect the changing role of governments and their relation to information. Secrecy and confidentiality rules assured control of information and were dominant in bureaucracies with a strong top down command and control culture. Freedom of information and transparency regulations opened these bureaucracies to feedback as part of quality control by the public. Recently the digital economy brought the growing role of positive networks effects under attention, where an extra user increases the value for all. This remains valid in the data economy, where qualitative data sources find additional uses by combining them with new data sources or applications.
Re-use of PSI regulations and open data policies reflect this new attention to the political economy of data. Re-use of PSI regulations are not additional transparency instruments, but rather forms of market regulation. Open data policies in general do not change the legal framework, but reflect the more active role of the government in building and enabling a data economy.
Attention goes to a broad range of aspects concerning the data value chain, not just access to the data but also the usability and interoperability. This can be seen in the 5-star scheme concerning open data, proposed by Tim Berners-Lee:[6]

		Licence
★	Make the data available on the web (whatever format)	Open Licence
★★	Make it available as structured data (e.g. excel instead of image scan of a table)	Open Licence
★★★	Make it available in an open, non-proprietary format (for example, CSV or XML instead of Microsoft Excel)	Open Licence
★★★★	use Uniform Resource Locators (URLs) to identify things, so that people can point at your data	Open Licence
★★★★★	link your data to other people's data to provide context	Open Licence

This scheme is now also used in Open data strategies, like the UK Open Data White Paper.[7]

Governmental policies are forward-looking, but also reflect the changing role of ICT and the internet. Early policy concerning ICT was often reactive and concerned the adaption of old economic models and their regulation to new technology. For instance the changes made to the IPR framework concerning software or to guarantee the functioning of the music and movie industry in the age of much easier digital reproduction. Also innovation policies were directed to the internet sector as a distinct economic sector. Attention went to enabling infrastructure, like broadband technology or more recent cloud computing, and to creating stable legal frameworks for this new sector, e.g. through e-commerce regulation.
The development and adoption of digital technology has now reached a stage that it leads to changes in business processes and organization in more traditional sectors of the economy and government. This gets reflected in more comprehensive visions on the data economy and data value chains.

6 Tim Berners-Lee, "Is your Linked Open Data 5 Star?" (addition 2010), in: Tim Berners-Lee, "Linked Data", 18 June 2009, http://www.w3.org/DesignIssues/LinkedData.html

7 UK Government, Open Data White Paper. Unleashing the Potential, 28 June 2012, p. 24. https://www.gov.uk/government/publications/open-data-white-paper-unleashing-the-potential

2.3 CONCLUSION

In this preliminary exploration we identified the main legal frameworks which have an impact on access to data, the linking of and re-use of data. These are intellectual property rights and licensing, privacy regulations and the regulation of access and use of public sector information. None of them is specific to big data processing. But big data poses challenges to these legal frameworks or changes their role.

The application of copyright to data and big data is not straightforward. Intellectual property rights create data enclosures, but it remains unclear when data is protected and which uses fall under the exclusive rights. Restrictive interpretations or implementations of exceptions to these rights make data mining very difficult. Copyright law offers a default regime on which private actors can shape their relations through licenses. However, also the application of contract law on licenses in the digital market proves to be unclear. And the patchwork of jurisdictions makes it difficult to create a stable environment for the combination of diverse sources of data. This problem also affects open content licenses, the tools developed to enable easy access and use of data.
The protection of personal data is a strong limitation of data collection and data use and defines a specific regime of processing and access rights. In the following chapters we will look into the diverse privacy regulations and how they affect big data.
Information held by public authorities is subject to a specific regime concerning access and use. The function of these regimes is changing from a transparency tool to an enabler of the data economy through open data policies. This shift also moves the attention from mere availability of information to interoperability of data sources. This widens the scope of issues to include elements concerning the usability and quality of the data.
Governmental policies evolve from reactive adaption to new technologies over a sectorial internet policy into comprehensive strategies which include a visions on how big data changes business processes in general.
This outlines the elements we will explore in the next chapters in a comparative approach.

3 DATA POLICIES IN THE EUROPEAN UNION

3.1 INTRODUCTION – EU POLICY ON THE DATA ECONOMY

Developing a functioning data economy is a EU policy goal embedded in the long term EU socio-economic policies. The Europe 2020 strategy sets out a vision on how the EU has to develop its social market economy. This vision is functions as a coordinating umbrella vision for more specific policy initiatives. Part of the Europe 2020 strategy are 7 flagship initiatives and one of them is the 'Digital Agenda for Europe'. The main focus of this Digital Agenda is "a digital single market based on fast and ultra-fast internet and interoperable applications".

Here we will give an overview of relevant actions and initiatives developed as a result of the Digital Agenda. We also review the most recent policy initiative of the European Commission building on the Digital Agenda: the Communication of 2.7.2014 on a data driven-economy.
In the next part we will review the main legal frameworks defining the European data economy. The last part reviews sectorial legal frameworks and data policies.

Digital Agenda for Europe

The Digital Agenda contains a comprehensive agenda concerning the digital economy. It identified a wide range of obstacles: fragmented digital markets, lack of interoperability, rising cybercrime and risk of low trust in networks, lack of investment in networks, insufficient research and innovation efforts, lack of digital literacy and skills and missed opportunities in addressing societal challenges. The actions defined in answer to these obstacles are as wide ranging. Here we will focus on the actions and resulting policy initiatives which concern the access, linking and use of data.

A first important action in the Digital Agenda, within the aim to create a vibrant digital single market, concerns the opening up of access to content. The main problem is that the European digital market is still very fragmented, both concerning private and public data or content. Action points identified are simplifying copyright clearance, management and cross-border licensing.[8] Part of this has been the review of the PSI Directive and the adoption of Directive 2014/26/EU on collective rights management and multi-territorial licensing of rights in musical works for online uses, but also the ongoing review of the data protection framework with the proposed General Data Protection Regulation (GDPR) and e-commerce related legislation.[9]
The Commission plans continued action on e-commerce related issues and intellectual property rights. Relevant concerning data policies is also the stated intention to make proposals to strengthen the European data industry, specifically on "issues such as common licensing conditions and the implementation of charging rules to enable public data to fuel the development of online content".[10]

[8] European Commission, A Digital Agenda for Europe, COM(2010)245, 19.5.2010, p. 9.
[9] European Commission, The Digital Agenda for Europe – Driving European growth digitally, COM(2012)784, 18.12.2012, p. 5.
[10] Ibid., p. 6.

Also relevant in this context is the structured stakeholder dialogue Licences for Europe held in 2013, which addressed cross-border portability of content, user-generated content (UGC), data- and text mining, access to audiovisual works and cultural heritage institutions.[11]
As part of this opening up of content the Commission also focused on public data. It presented its policy in the Communication COM(2011) 882 of 12.12.2011 on 'Open data. An engine for innovation, growth and transparent governance'. Public sector information is seen as a resource. With an active open data-policy this resource is made available for the European economy.

A second important action area linked to data policies in the Digital Agenda is the focus on interoperability and standards. This concerns a wide range of hardware, software, IT services but it can also concern data. Standardization has always been an important instrument in the single market and it also plays a key role in creating a functioning data economy. When content remains locked up in incompatible formats, licenses, etc., the data economy remains very fragmented. The focus on making data sources more interoperable is mostly present in the effort to enhance the interoperability between public administrations.

Further a lot of attention in the Digital Agenda concerned trust and security. The main issue raised in this context which can have an impact on the access and use of data is data protection, and involves the review of the data protection framework.

Big data as such was not mentioned in the Digital Agenda, but it contained attention for cloud computing as part of its innovation strategy, with the development of "an EU strategy for cloud computing notably for government and science" as specific action. Cloud computing is an important infrastructure to enable big data processing, and attention for cloud computing is the only element specific to big data in this Digital Agenda. Following the Digital Agenda, the European Commission outlined a more specific policy agenda on cloud computing in its Communication Unleashing the Potential of Cloud Computing in Europe.[12] It presented 3 key actions: enhance standards and certification, establishing safe and fair contract terms and conditions (through model contracts and contractual clauses, and a code of conduct for cloud computing providers) and the launch of the European Cloud Partnership. Especially the action on contracts has an important effect on the access and use of data, even when this concerns infrastructure for big data processing. It can create a more predictable and safe environment in terms of data security and data protection and avoid abuse by cloud providers of data on their servers. The state of work and results are presented in the Report on Implementation accompanying the Communication on a data-driven economy.[13]

Towards a thriving data-driven economy

The European Commission presented an updated version of its vision on the data economy in its Communication of 2.7.2014 on a data-driven economy.[14] It builds upon the ideas first

[11] European Commission, On content in the Digital Single Market, COM(2012)789, 18.12.2012; Results can be found on http://ec.europa.eu/licences-for-europe-dialogue/en
[12] European Commission, Unleashing the Potential of Cloud Computing in Europe, COM(2012) 529, 27 September 2012.
[13] European Commission, Report on the Implementation of the Communication 'Unleashing the Potential of Cloud Computing in Europe' Accompanying the document Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions 'Towards a thriving data-driven economy', SWD/2014/0214, 2 July 2014.
[14] European Commission, Towards a thriving data-driven economy, COM(2014) 442, 2 July 2014.

formulated by Commission Vice-President Neelie Kroes in a strategic initiative on the data value chain, which was launched in the ICT2013 conference in Vilnius in November 2013.[15]

This policy agenda aims to "provide the right framework conditions for a single market for big data and cloud computing". It puts data, and data-driven innovation, forward as the central element in the future knowledge economy. Data-driven innovation is defined as "the capacity of businesses and public sector bodies to make use of information from improved data analytics to develop improved services and goods". In other words, improved data analytics are seen as key to more efficient business and production processes.
The Communication presents a vision of a data-driven economy characterized by:

- Availability of good quality, reliable and interoperable datasets and enabling infrastructure
- Improved framework conditions that facilitate value generation from datasets

This refers to an adequate skills base and cooperation between players.

- A range of application areas where improved big data handling can make a difference

The action plan consists of:

- Community building: This involves setting up a contractual Public-Private Partnership (cPPP) on data to develop a data community, encourage exchange of best practices, steer R&I activities and implement H2020 in this field, and to develop incentives to share datasets. Further initiatives concern digital entrepreneurship and an open data incubator supporting the development of data value chains, the development of the skill base, a data market monitoring tool to measure the European data market and efforts to identify sectorial priorities for R&I with stakeholders and research communities.
- Developing framework conditions through the availability of data and interoperability: This involves fostering open data policies, H2020 efforts concerning data handling tools and methods, and the support of new open standards.
- Developing framework conditions through enabling infrastructures
- Regulatory issues

The regulatory issues concern:

- Personal data protection and consumer protection: The reform of the data protection framework is a first objective. After the adoption the Commission plans to work on guidance for issues like data anonymization, data minimization, tools for consumer awareness, etc. and to support R&I on privacy by design. It will consult and support R&I on user-controlled cloud-based technologies for storage and use of personal data. Further regulatory work concerns ensuring the application of consumer law on big data technologies.
- Data mining and its relation to the copyright framework
- Security: The Commission plans to explore the security risks related to big data technologies and propose risk management and mitigation measures.
- Data location requirements, taken as a security measure, form a barrier to the single market for cloud computing and big data. Also the need for guidance on issues of data ownership and liability, especially in the context of the Internet of Things, will be investigated.

[15] European Commission, *A European strategy on the data value chain*, 12 December 2011. http://ec.europa.eu/information_society/newsroom/cf/dae/document.cfm?action=display&doc_id=3488

We can conclude that with this Communication the European Commission develops a broad vision on the data economy. Although less outspoken in its vision as the DG Connect paper on the data value chain, it clearly puts forward the central role of (big) data in the knowledge economy. This makes it in international comparison one of the rare comprehensive big data strategies. However, its proposed actions clearly build on the earlier initiatives.

3.2 GENERAL FRAMEWORK CONCERNING THE USE OF INFORMATION AND DATA

3.2.1 Intellectual property rights

The EU has done some important efforts to provide for legal interoperability by harmonizing intellectual property. Main instruments interesting in the big data-context are the database directive and the harmonization of copyright by the InfoSoc directive.

Directive 2001/29/EC on the harmonization of certain aspects of copyright and related rights in the information society, also known as the InfoSoc directive, harmonizes copyright in the EU according to the WIPO Copyright treaty and the TRIPS treaty. It states that member states have to provide authors or others producers of intellectual products with exclusive rights on reproduction, on communication to the public and on making available to the public by wire or by wireless means, and on distribution. The directive also sets aside the principle of exhaustion in these online cases of communication and making available to the public, as well as for certain cases of distribution.
Important is that the directive does not provide a definition of the originality requirement or the amount of creativity needed in this requirement.[16] This is left to the member states and by consequence remains different, leading to a possible different application on certain data.
The directive provides further for a range of exceptions. Important for big data practices can be the exception on the right of reproduction for temporary acts of reproduction which are transient or incidental, are an essential part of technological processes like transmission or other lawful uses and have no independent economic significance.[17] This exception was meant for caching and temporary storage during digital communication, but can also be used for text and data mining. All these exceptions are limited by the 'three-step test': they can "only be applied in certain special cases which do not conflict with a normal exploitation of the work or other subject-matter and do not unreasonably prejudice the legitimate interests of the rightholder"[18].

Data is of course not always subjected to copyright. When the data is purely factual and does not fulfil the originality requirement, others are free to use the data. Also, even if some copyright applies on a certain expression, this copyright does not prevent the use of the fact covered by the expression. Further, copyright concerns certain uses, like reproduction, while some of these uses are excluded from protection by exception. If text or data mining always involve such reproduction and is not covered by an exception is a question under debate.[19]

---

[16] Truyens, Maarten & Patrick Van Eecke, "Legal aspects of text mining", *Computer Law & Security Review*, Vol. 30, no. 2, 2014, p. 156.
[17] European Parliament and the Council, Directive 2001/29/EC of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society, art. 5 §1.
[18] Ibid., art. 5 §5.
[19] Triaille, Jean-Paul, Jérôme de Meeûs d'Argenteuil and Amélie de Francquen, *Study on the legal framework of text and data mining (TDM)*, March 2014.
http://ec.europa.eu/internal_market/copyright/docs/studies/1403_study2_en.pdf; Hargreaves, Ian et al, *Standardisation in the area of innovation and technological development, notably in the field of Text and Data*

The European Court of Justice (EUCJ) has made several decisions based on these directives which are relevant for big data practices. In the Infopaq-decision of 16 July 2009 the court looked into the application of copyright law on a search engine of newspaper articles, providing summaries of articles. It stated that the protection by copyright according to directive 2001/29 applies only when the data "is original in the sense that it is its author's own intellectual creation"[20]. This originality requirement also needed to be checked when the discussion concerned reproduction in part. In this case string of 5 words before and after the keyword were stored. The Court considered a word in isolation not to be the intellectual creation of the author, but that such creation could be achieved "through the choice, sequence and combination of those words"[21]. Words were therefore not covered by protection, but strings of 11 words could be and this needed to be checked by the national court.
Further the court stated that the copies made of the newspaper for the search for keywords could be considered a temporary and transient act of reproduction which fell under the exception if those copies were indeed automatically deleted at the end of the process. This exception could not apply for the further storage or printing of the strings of 11 words, when these fell under protection.
In a second decision in the same case the EUCJ further clarified the exception. The acts of reproduction were all part of the same technological process, of which the purpose was lawful (making summaries). And these acts could be considered as not having an independent economic significance if their implementation did not enable the generation of an additional profit, going beyond that derived from lawful use of the protected work and if these acts did not modify the work.[22]
Other exceptions were not checked. The most relevant is the exception for quotes, but it is part of national law and varies a lot depending on jurisdiction.[23] The exception for temporary storage did draw attention also in other decisions.[24]

This decision makes clear that the characteristics of text and data mining methods are legally relevant in the context of copyright law. Methods based 'bag-of-words' sets, making a frequency distribution of words in a text, can probably avoid the applicability of copyright protection. But the use of longer strings can make copyright law applicable and will demand efforts to make sure it remains in the space left by the exemptions. Otherwise authorization will be required from all right holders on texts in the dataset.

Database protection is provided by directive 96/9/EC of 11 March 1996 on the legal protection of databases. This directive contains 2 forms of protection for databases, one as copyright, another as a sui generis right. These protections can coincide.
A database is defined as "a collection of independent works, data or other materials arranged in a systematic or methodical way and individually accessible by electronic or other means"[25].

Mining. Report from the Expert Group, Publications Office of the European Union, Luxembourg, 2014. http://ec.europa.eu/research/innovation-union/pdf/TDM-report_from_the_expert_group-042014.pdf
[20] EUCJ, C-5/08, *Infopaq International A/S v Danske Dagblades Forening*, 16 July 2009, §37.
[21] Ibid., §45.
[22] EUCJ, C-302/10, *Infopaq International A/S v Danske Dagblades Forening*, 17 January 2012,
[23] Truyens and van Eecke, op. cit., 2014, p. 158.
[24] EUCJ, C-360/13, *Public Relations Consultants Association Ltd v. Newspaper Licensing Agency Ltd and Others*, 5 June 2014 (aka the Meltwater decision); EUCJ, C-403/08 and C-429/08, *Football Association Premier League Ltd*, 4 October 2011.
[25] European Parliament and the Council, Directive 96/9/EC of 11 March 1996 on the legal protection of databases, art. 1 §2.

A first form of protection is given by copyright protection. Protected are databases which, by reason of the selection or arrangement of their contents, constitute the author's own intellectual creation. The copyright protection of databases does not extend to their contents. It remains distinct from and does not affect copyright or other rights on the contents. That implies that the intellectual creation involved in producing the individual data items is not taken into account to judge the intellectual creation involved in making the database.
The author of a database is granted exclusive rights on reproduction, adaption, distribution, communication, display or performance to the public. Exception to these exclusive rights is the performance of these acts by the lawful user of a database to access the contents of the database. Normal use of the contents by the lawful user does not require the authorization of the author of the database. The directive gives member states the option to provide other exceptions traditionally foreseen under copyright law. This as long as the 3-step test is respected, that is exceptions are specific, may not unreasonably prejudice the right holder's legitimate interests or conflict with the normal exploitation of the database.

This copyright protection remains linked to the protection of an intellectual creation, implying that a minimal creativity is required. It can be difficult to prove the creativity in the selection of arrangement, while building the database can involve a substantial economic effort. Therefore the directive provides for a second form of protection: the Sui generis-right. This sui generis-right protects the maker of a database "which shows that there has been qualitatively and/or quantitatively a substantial investment in either the obtaining, verification or presentation of the contents"[26]. What is protected is the economic investment in creating the database. To assess this substantiality, the cost for obtaining, creating or updating the individual data items cannot be taken into account. Only the costs associated with the actual making and maintenance of the database itself are relevant.[27]
The maker of the database is given the right to prevent extraction and re-utilization of the whole or of a substantial part of the contents of that database. Extraction stands for "the permanent or temporary transfer of all or a substantial part of the contents of a database". Re-utilization is any form of making available to the public, like the distribution of copies or renting by on-line or other forms of transmission. A limited form of exhaustion is foreseen on the right to control resale of a copy of the database, after the first sale of a copy by the right holder or with his consent.
This right does not prevent lawful use, consisting of extracting or re-utilizing insubstantial parts of database contents. Also this substantiality can be assessed both quantitatively or qualitatively. Further may this use not conflict with the normal exploitation of the database or unreasonably prejudice the legitimate interests of the maker of the database. This lawful use does not affect other limitations posed by copyright or other rights residing in the contents. The sui generis-right provides less exceptions than copyright, but possibly relevant is the exception for non-commercial scientific research.
Another important difference is that this sui generis-right is only available for a citizen or a resident in a EU member state, or a company with its principal place of business in the EU. Extending this right to databases made in third countries depends on agreements between the EU and that state. Does this state has a similar sui generis protection or does it respect this EU protection? As this sui generis protection is not known in most countries outside the EU it will often not be the case. For instance the US does not have a similar protection of databases. Only protection of databases by copyright exists in the EU. This has the consequence that no sui generis protection can be claimed in the US, also not on databases

[26] Ibid., art. 7 §1.
[27] Truyens and van Eecke, op. cit., 2014, p. 160.

originating from the EU and which receive such protection in the EU. Another consequence is that no sui generis protection is granted inside the EU for databases originating from the US, as the needed reciprocity does not exist.

It is clear that database rights can be a limit to big data practices. Combining data sets and data mining can be made difficult by the limitation to 'normal' or insubstantial use of the contents of databases. It can also be questioned if the sui generis protection of databases is not over-protective and harmful for the data economy. The strong protection dates from an earlier assumption that strong property rights stimulates investments and by consequence the economy. However, in the data economy such protection can be a barrier to big data practices. Ian Hargreaves points out in his review of intellectual property rights that the evaluation of this directive by the European Commission in 2006 found out that database creation had declined.[28] The US continued to see a growth.

Intellectual property law is a default regime from which right holders can depart through contract law. Licenses for the use of databases or protected sources of data can be more restrictive than copyright law and put aside some of the exceptions. The database directive limits this possibility by stipulating that contractual provisions limiting the normal or insubstantial use by lawful users are null and void. But the other exceptions are subject to the freedom of contract, and licenses can forbid practices like screen scraping and data mining.
Courts have dealt with practices like screen scraping. But it is clear that European intellectual property law is barely adapted to the new digital environment, as exceptions are clearly based on practices pre-dating the internet. Big data practices which have become common, like the Google search engine, are based on practices contrary to EU copyright law. For instance, it works with an opt out-principle through the robots.txt-file, but copyright law points to an opt-in, as was decided in the Belgian Google News-case.[29] This forces courts to look for creative solutions, like the German Bundesgerichtshof which allowed thumbnail pictures based on an implied license.

Recent EU policy documents on intellectual property focus mostly on copyright of digital content like music, video, etc. Problems related to databases and big data are not present in the recent communications of the European Commission on intellectual property or in the Digital Agenda[30], with 2 exceptions: digitization of cultural heritage and lately also text and data mining.
The first problem concerns the limits on digitization of cultural heritage due to high transaction costs for rights clearing. This problem was already recognized in 2009.[31] Problem is that libraries or archives do not have a blanket exception allowing them to digitize their entire collection. This needs prior authorization of right holders, leading to high costs for rights clearing. Problems exacerbate for archive pieces like private documents, business records, etc. and for so-called orphan works for which the right holders cannot be located.

[28] Ian Hargreaves, *Digital Opportunity: Review of Intellectual Property and Growth*, May 2011, http://www.ipo.gov.uk/ipreview-finalreport.pdf, p. 19.
[29] Court of Appeal Brussels, 5 May 2011, *Google Inc. v. Copiepresse et al*, http://jure.juridat.just.fgov.be/oldf/view_decision?justel=F-20110505-18&idxc_id=252985&lang=fr
[30] European Commission, Copyright in the Knowledge Economy, COM(2009)532, 19.10.2009; European Commission, A Single Market for Intellectual Property Rights – Boosting creativity and innovation to provide economic growth, high quality jobs and first class products and services in Europe, COM(2011)287, 24.5.2011; European Commission, A Digital Agenda for Europe, COM(2010)245, 19.5.2010; European Commission, The Digital Agenda for Europe – Driving European growth digitally, COM(2012)784, 18.12.2012; European Commission, On content in the Digital Single Market, COM(2012)789, 18.12.2012
[31] European Commission, op. cit., 2009, p. 4.

Directive 2012/28/EU of 25 October 2012 on certain permitted uses of orphan works now provides for such exception after a diligent search for right holders, while Directive 2014/26/EU on collective rights management and multi-territorial licensing of rights in musical works for online uses should lower the transaction costs for rights clearing.
The Commission also considered in 2011 in its Intellectual Property Strategy the possibility of a comprehensive codification in a European Copyright Code, which would include an evaluation of directive 2001/29/EC.[32] This review is ongoing and a decision is aimed at in 2014.[33]

Relevant concerning data policies is also the stated intention in the Digital Agenda to make proposals to strengthen the European data industry, specifically on "issues such as common licensing conditions and the implementation of charging rules to enable public data to fuel the development of online content"[34]. In the Communication on 'Content in the Digital Single Market' this was further developed through the structured stakeholder dialogue Licences for Europe, which was held in 2013. This stakeholder dialogue addressed cross-border portability of content, user-generated content (UGC), data- and text mining, access to audiovisual works and cultural heritage institutions.[35] This dialogue led to 10 pledges to bring more content online. One pledge concerned specifically non-commercial text and data mining in scientific research and contains 3 commitments: to implement a clause allowing such data mining in agreements, to develop the mine-ability of content and to develop platforms allowing researchers to integrate subscribed journal content and open access licensed content for text and data mining.[36]

We can conclude that, notwithstanding the harmonization effort with the InfoSoc-directive, the European IPR framework clearly poses problems for big data processing. The exceptions in the copyright framework can be problematic for data mining, although the EUCJ did an effort to widen the interpretation. It can also be questioned if the sui generis protection of databases is not over-protective and harmful for the data economy. The majority of the IPR reforms point to a protection of older business models. Without putting the legitimacy of these interests into question, it is as legitimate to reconsider the balance struck in the IPR framework and create more space for the functioning of a data economy. This is partly a question of modernization of the copyright framework, which can be done by legislative work or interpretation by courts. But is also a political question involving different interests and finding an appropriate balance between them.

3.2.2 Access to and re-use of public sector information

[32] European Commission, A Single Market for Intellectual Property Rights, op. cit., 2011, p. 11.
[33] European Commission, On content in the Digital Single Market, op. cit., 2012, p. 5.
[34] European Commission, The Digital Agenda for Europe – Driving European growth digitally, op. cit., 2012, p. 6.
[35] European Commission, On content in the Digital Single Market, op. cit., 2012; Results can be found on http://ec.europa.eu/licences-for-europe-dialogue/en
[36] A Statement of Commitment by STM publishers to a roadmap to enable text and data mining (TDM) for non-commercial scientific research in the European Union, http://ec.europa.eu/licences-for-europe-dialogue/sites/licences-for-europe-dialogue/files/10-Text-data-mining.pdf

The EU makes a difference between access and re-use of public sector information (PSI). After presenting the basic legal notions of access, re-use and sharing, we will discuss the main instruments in more detail.

It has no competence to regulate access to PSI in member states, except on environmental information. The general framework for accessing information held by EU institutions is the Access to documents-regulation 1049/2001. This regulates access on request (passive transparency). Access to environmental information is regulated in the Access to environmental information-regulation 1367/2006 for EU-institutions and the Access to environmental information-directive 2003/4/EC for information held by member state institutions and private companies fulfilling public duties.
Re-use of information is regulated by the PSI directive 2003/98/EC, and concerns PSI held by member states.
More specific sectorial rules exist on access and re-use, like the Inspire-directive.

Limitations of access and re-use can be found in the refusal grounds of the Access to documents-regulation and the Access to environmental information-directive. These are often further developed in specific legislation: data protection, statistical confidentiality, secrecy laws, …
The term 'access' is nowhere defined in EU law. The nearest to a definition comes article 10 of Regulation 1049/2001, which states "The applicant shall have access to documents either by consulting them on the spot or by receiving a copy, including, where available, an electronic copy", although this article rather defines the modalities of access to documents.
But the several instruments regulating or granting the right of access to documents held by EU institutions or national authorities make clear that the right of access consists of a right to see and to take knowledge of the content of the documents. This right of access does not go without restrictions and is regulated by national and EU law.
Access concerns the availability of information. The literature often makes a difference between active transparency, meaning the making available of information by initiative of the public authority, and passive transparency, which concerns granting access to information on request.

This right to access official documents is enshrined in the constitutions of several member states and in the Treaty on the Functioning of the European Union (TFEU), which states "Any citizen of the Union, and any natural or legal person residing or having its registered office in a Member State, shall have a right of access to documents of the Union's institutions, bodies, offices and agencies, whatever their medium, subject to the principles and the conditions to be defined in accordance with this paragraph"[37].
This right is seen as part of the principle of openness, which allows citizens to participate and to keep the government accountable. It evolved from a principle of good governance into a fundamental right necessary for the functioning of a democracy. As such it is also included in the Charter of Fundamental Rights of the European Union: "Any citizen of the Union, and any natural or legal person residing or having its registered office in a Member State, has a right of access to European Parliament, Council and Commission documents"[38]. This article grants a right to access to documents held by the EU institutions, not by the member states. It limits the beneficiaries to natural and legal persons with citizenship or residence in the EU, not to non-citizens residing outside the EU.

[37] Treaty on the Functioning of the European Union (TFEU), art. 15(3).
[38] Charter of Fundamental Rights of the European Union, art. 42.

Also the European Court of Human Rights has developed the right to access PSI as a fundamental right based on the freedom of expression in article 10 of the European Convention of Human Rights, more specifically as part of the freedom to receive information. This applies to the member states.

The availability of public sector information and the right to access official documents does not imply to right to use the information contained in those documents without restrictions. EU-law differentiates the right of access from the right of re-use of public sector information.
A definition of re-use can be found in the PSI-directive: "the use by persons or legal entities of documents held by public sector bodies, for commercial or non-commercial purposes other than the initial purpose within the public task for which the documents were produced. Exchange of documents between public sector bodies purely in pursuit of their public tasks does not constitute re-use"[39].
In other words, re-use concerns the further use of the information, after having received knowledge of it. The use of this information by public authorities for the original public task for which the information was gathered is not considered re-use. When these authorities also make a further use of this information, which is not part of their public task, it is considered re-use. E.g. the commercialization of certain data in order to recuperate part of the costs, like publishing and selling maps.
The use by other public sector bodies of this information is also not considered re-use, as long as this further use is part of their public task. In the literature this is sometimes categorized separately as 'sharing'.[40] The purpose of the use of PSI, part of their public task or not, is delineating re-use from other uses by public sector bodies of information.[41]

Access to documents - Regulation 1049/2001/EC

This regulation implements the right of access to documents held by the European Parliament, Council and Commission, as foreseen in art. 15(3) TFEU. It concerns both passive and active transparency. All 3 institutions have also additional transparency clauses in their rules of procedure.

Art 15(3) TFEU grants this right of access to "Any citizen of the Union, and any natural or legal person residing or having its registered office in a Member State". The regulation has in article 2.1 the same scope of beneficiaries, but allows in article 2(2) the institutions to grant similar access to natural or legal persons outside this scope. The Commission and the European Parliament have granted access to persons outside the EU in decisions concerning their rules of procedure[42], the Council's Rules of Procedure do not mention such persons.

This regulation concerns documents, not information. This difference means that information is delivered in the state as it is in possession of the institution. The institution is not obliged to process the content into a certain form, produce new information or new compilations with the content.

[39] European Parliament and the Council, Directive 2003/98/EC on the re-use of public sector information (PSI-directive), 17 November 2003, art. 2(4).
[40] Janssen, Katleen, *The EC Legal Framework for the Availability of Public Sector Spatial Data. An examination of the criteria for applying the directive on access to environmental information, the PSI directive and the INSPIRE directive*, ICRI, Leuven, 4 December 2009, p. 63.
[41] Ibid., p. 65.
[42] Although the EP keeps room for more limited access by adding 'as far as possible'.

A document is defined as 'any content whatever its medium (written on paper or stored in electronic form or as a sound, visual or audiovisual recording) concerning a matter relating to the policies, activities and decisions falling within the institution's sphere of responsibility'[43]. Content in electronic form includes data stored in a database.
The right of access concerns in principle all documents held by the institutions, also documents not created by them but which are received from third parties.

The regulation foresees 3 methods of providing access, namely "documents shall be made accessible to the public either following a written application or directly in electronic form or through a register"[44]. The first method concerns the passive transparency giving access on request, the second method is about active transparency giving access on initiative of the institution to the general public. The third method is a limited form of active transparency, meant to provide an instrument to make the passive transparency effective.

Passive transparency implies that access is granted only on request. Such request can be made in any written form, including electronically. Reasons for the request or an interest in the document do not have to be given or proved.
As this is a right to access a document and not information, the applicant has to be sufficiently precise to enable the institution to identify the document. To enable applicants in requesting specific documents, the institutions have set up a register of documents, in which all documents have to be referenced (with some exceptions for sensitive documents).
The institution has 15 days to make a decision, which can be prolonged with another 15 days when it concerns a large amount of documents. In this case the regulation allows for consultations between applicant and institution.
The institution can charge the applicant for access, but this charge may not exceed the real cost of producing and sending the copies. No extra costs can be included, like to compensate part of the original cost for collecting the data. This does not prevent such higher charging for certain cases of re-use, which is foreseen in the PSI-directive. Consultation on the spot or sending copies less than 20 A4 pages is free of charge.
A document can also be a database, even a large one. But the procedure for passive transparency is clearly not developed for a regular or online checking of data. From the viewpoint of big data uses, in which official data is integrated or a dataset needs to be regularly checked, this procedure is not adapted to those applications. In principle it is possible to ask a copy of a database, but the integration of this data in big data applications is regulated by the PSI directive. This regulation remains important as the PSI directive refers to it for the grounds of refusal.

Active transparency is provided by making documents directly accessible in electronic form or through the registers. The Regulation lists documents which have to be made directly accessible or which have to be published in the Official Journal. This direct access to electronic documents or to the registers is free of charge.
In principle this information is available for integration in big data applications, but such re-use is regulated by the PSI-directive. Also technical hurdles can remain, as the information is often not provided in the most accessible electronic format (that is machine-readable), while the Terms of Service of the concerned EU-website can be an obstacle for uses involving a large amount of requests. The EU-administration is slowly moving towards more accessible formats for the data it provides. More on this in the part on the open data policy.

[43] European Parliament and the Council, Regulation 1049/2001 regarding public access to European Parliament, Council and Commission documents, 30 May 2001, art. 3(a).
[44] Ibid., art. 2(4).

This right of access to documents is not unlimited but has to be balanced with other interests. As this right is also a fundamental right, all limitations have to be provided by law.
Article 4 of Regulation 1049/2001 contains several grounds based on which access to a document can be refused if disclosure would undermine the protection of:
- the public interest as regards public security, defence and military matters, international relations, the financial, monetary or economic policy of the Community or a Member State
- privacy and integrity of the individual. Disclosure has to be in accordance with the data protection legislation
- commercial interests, including intellectual property
- court proceedings and legal advice
- the purpose of inspections, investigations and audits
The last 3 grounds are only relative, meaning that refusal has to be balanced with the public interest in disclosure. When this interest is large enough access has to be granted, even when there is a certain damage to these grounds. The first 2 grounds are absolute: access has to be refused, no balancing with the public interest in disclosure is needed.
Linked to the refusal grounds foreseen in article 4(1)(a), public security, defence and military matters, is a specific category of sensitive documents, which get a confidentiality classification (TOP SECRET, SECRET or CONFIDENTIAL). Only officials with the right to access these classified documents can decide on applications for access and decide about their declassification. Also referencing in the public registers can be limited or such reference can be entirely withheld. Decisions giving access to sensitive documents or recording them in the registers can only happen with consent of the originator.

Article 4(3) provides another ground of refusal: when access to a document would undermine the institution's decision-making process. Again this is a relative ground and it needs to be balanced with the public interest in disclosure. Two cases are foreseen:
- when the document relates to a matter where the decision has not yet been taken by the institution. This ground is by consequence limited in time.
- when the document contains opinions for internal use as part of deliberations and preliminary consultations within the institution concerned. Such access can be refused even after the decision has been taken.
Another proportionality mechanism is that documents can be partially released, when only part of the document is covered by the grounds for refusal.

The right of access also applies to documents not originating from the institution itself, but the third parties need to be consulted before taking a decision on the access and member states may request not to disclose without their prior agreement documents originating from them.
All grounds of refusal can be raised during a maximum period of 30 years, although when the grounds are privacy or commercial interests or in case of sensitive documents (see further) this period can be prolonged.

Re-use of public sector information - PSI directive 2003/98/EC

The directive 2003/98/EC on the re-use of public sector information, or PSI directive, regulates the re-use of documents or information. That is the further use of the information, after having received knowledge of it through access.

Access to such information is not regulated on EU level, except for environmental information, but by the member states themselves. The directive makes clear it does not affect national access regimes.
A definition of re-use can be found in article 2(4) of the PSI directive: "the use by persons or legal entities of documents held by public sector bodies, for commercial or non-commercial purposes other than the initial purpose within the public task for which the documents were produced. Exchange of documents between public sector bodies purely in pursuit of their public tasks does not constitute re-use".

Aim of the directive is to create an internal market of PSI. It wants to assure that all private actors can use public sector information in an equal manner. The documents to which this directive applies have to be available for re-use both for commercial and non-commercial purposes under the conditions stipulated by the directive. This directive does not obliges states to give access or to allow re-use, only that once the permission for re-use is given it has to be done under equal conditions for all players and in a transparent manner.

Document is defined in the PSI directive as: "(a) any content whatever its medium (written on paper or stored in electronic form or as a sound, visual or audiovisual recording); (b) any part of such content"[45]. The preamble clarifies that this definition does not include source code.
The PSI directive limits the range of documents to which it is applicable. It excludes documents whose supply falls outside the scope of the public task of the public sector body, documents for which third parties have intellectual property rights, documents to which access is excluded or restricted by the national access laws, documents containing personal data for which the further use would not be compatible with data protection law, etc. It also excludes documents held by public service broadcasters and educational, research or cultural establishments.
This implies that the scope of the PSI-directive is more limited than the Access regulation.

This directive only concerns re-use of documents held by member states. The EU institutions regulate the re-use of their documents themselves, as is done by the Commission in its Decision 2011/833/EU of 12 December 2011. It concerns documents held by public sector bodies. This has to be understood broadly as in EU public procurement law.
Re-use concerns the further use by natural persons or private legal persons, not by other public sector bodies.

The directive provides a procedure through which natural and legal persons can ask to re-use certain PSI or to be given a license to do so. When such permission or license is given, public sector bodies have to make their documents available in any pre-existing format or language. There is no obligation to provide a translation in another official EU-language when it does not exist already or to produce the data in a certain format if this is a disproportionate effort. But when available, it has to make the documents available in "open and machine-readable format together with their metadata. Both the format and the metadata should, in so far as possible, comply with formal open standards."[46]
Public sector bodies may allow re-use without conditions or may impose conditions, where appropriate through a license. Member States have to establish standard licenses for the re-

[45] PSI-directive, art. 2(3).
[46] PSI-directive, art. 5(1).

use of public sector documents, which can be adapted to meet particular license applications, and ensure these are available in digital format and can be processed electronically.

The conditions linked to the re-use of documents have to be non-discriminatory for comparable categories of re-use. This concerns the private actors but also public sector bodies when it is for activities outside their public tasks. When public sector bodies develop commercial activities outside their public task it has to be done in the same market conditions as the private actors.
This also implies that as a general rule no exclusive rights are granted but that the re-use remains open for all actors on the market. Exception is where an exclusive right is necessary for the provision of a service in the public interest. In this case such arrangements have to be transparent, made public and regularly review. A specific exception is given for the digitalization of cultural resources.

The public sector body can charge for the re-use of documents, but these charges have to be limited to the marginal costs incurred for their reproduction, provision and dissemination. The charge may not include extra costs linked to the original collection of data or production of the document.
The directive provides an exception for this rule and allows higher charges when public sector bodies are required to generate revenue to cover a substantial part of their costs relating to the performance of their public tasks in general or for specific documents relating to their collection, production, reproduction and dissemination. Also, this rule does not apply to libraries, including university libraries, museums and archives.
This charging has to be transparent. The conditions, calculation basis and the actual amount have to be pre-established and published.

Member states have to encourage public sector bodies to use standard licenses. They also have to arrange active transparency measures concerning which data is available for re-use, like asset lists and portals.

Re-use of public sector information - Commission Decision 2011/833/EU of 12 December 2011 on the reuse of Commission documents

Similar rules concerning re-use of public sector information (PSI) are made by the EU institutions. The Commission has done do in its Decision 2011/833/EU of 12 December 2011 on the reuse of Commission documents. It concerns the reuse of "public documents produced by the Commission or by public and private entities on its behalf"[47], both published or non-published.

Both 're-use' and 'document' are defined in similar terms as in the PSI-directive. The term 'public document' is not defined, but has to be understood as a document for which access is allowed according to regulation 1049/2001.
Also here the scope is more limited than in the Access regulation. It does not apply to software or to documents covered by industrial property rights, to documents for which the Commission is not in a position to allow their re-use in view of intellectual property rights of third parties, to documents for which no access can be allowed according to regulation 1049/2001, to data falling under statistical confidentiality, …

[47] European Commission, Decision 2011/833/EU on the reuse of Commission documents, 12 December 2011, art. 2(1).

The general principle is that all documents are available for re-use for commercial or non-commercial purposes, without charge, without the need to make an individual application and without restrictions or, where appropriate, with an open license or disclaimer setting out conditions explaining the rights of re-users. The decision provides an application procedure for exceptional cases.
Documents are made available in any existing format or language version, in machine-readable format where possible and appropriate. But this does not imply an obligation to create, adapt or update documents, nor to translate the documents into any other official language than already available.
The reuse of documents shall in principle be free of charge, but in exceptional cases, marginal costs incurred for the reproduction and dissemination of documents may be recovered. When a document is adapted in order to satisfy a specific application, the costs involved may be recovered from the applicant. Again, the charging has to be transparent. Any applicable conditions and standard charges have to be pre-established and published.
Any conditions linked to the re-use of documents have to be non-discriminatory for comparable categories of re-use. In principle no exclusive rights are granted and the re-use remains open for all actors on the market. Exception is where an exclusive right is necessary for the provision of a service in the public interest. In this case the arrangements have to be transparent, made public and regularly review. A specific exception are exclusive rights for publishers of scientific and scholarly journals for articles based on the work of Commission officials for a limited period.

The decision also provides to set up a data portal. This portal functions as a single point of access to structured data held by the Commission in a variety of web sources. The sources generate both the data and the related metadata, which are automatically harvested by the portal for indexing.

EU Policy concerning Open Data and data economy

As mentioned above, one of the actions in the Digital Agenda, concerns the opening up of access to content. This content also includes PSI. The Commission presented its policy in the Communication COM(2011) 882 of 12.12.2011 on 'Open data. An engine for innovation, growth and transparent governance'. Aim is to open up public sector information as a resource. The Commission points to estimations of the total market for PSI in 2008 at € 28 billion and of the total economic gains from further opening up PSI at € 40 billion a year for the EU27. Aside of the economic gains it also fosters public participation, scientific uses and addressing societal challenges.
The measures to overcome fragmentation and to open up PSI consists of three parts. First revisions of the legal framework for data re-use. This concerned the now completed revisions of the PSI directive and the Commission Decision on the re-use of its own information, but also sectorial initiatives. The Commission also mentioned a soft law approach to open access to scientific information. This led to the Communication 'Towards better access to scientific information: Boosting the benefits of public investments in research' and the Recommendation on 'access to and preservation of scientific information'.
Secondly support for R&D and innovation. This mainly through the Horizon 2020 program, as well as support for research infrastructures. The Commission also planned to set up 2 data portals. A first one through which it makes available its own data resources. This European Union Open Data Portal is now online at https://open-data.europa.eu/en/data/. A second would be a pan-European data portal with data from the Commission, member states and

public sector bodies. A prototype has been developed through the LOD2-project: http://publicdata.eu/.
Thirdly the Commission would continue to facilitate coordination and experience sharing with the Member states. Examples are a Member state's expert group called the PSI group, the ePSI-platform[48], ISA actions like on semantic interoperability, ...

The Digital Agenda also planned action on interoperability and standards. Making data sources more interoperable is mostly present in the effort to enhance the interoperability between public administrations. As part of the European eGovernment action plan 2011-2015 the Commission set as objectives:
"- By 2015, a number of key cross-border services will be available on line – enabling entrepreneurs to set up and run a business anywhere in Europe independently of their original location, and allowing citizens to study, work, reside and retire anywhere in the European Union.
- By 2015, 50% of EU citizens will have used eGovernment services."[49]

The focus on increased interoperability is in the first place aimed at cross-border exchanges of information between member state and EU public administrations, in order to enable European public services. Objective is to aggregate 'basic' public services and to make them Europe-wide accessible. This would result in the availability of the cross border services mentioned in the eGovernment action plan 2011-2015.
But such increased interoperability between public administrations would not only lead to more efficient and effective public administrations, but also have a strong impact on the data economy. Open data-policies have a limited effect when this data cannot be linked easily and remains locked in incompatible formats. Interoperability between open data sources turns these sources into big data.
As main policy tools the European Commission developed a European Interoperability Strategy (EIS)[50] and a European Interoperability Framework (EIF)[51], and promotes now the adoption of national interoperability frameworks by member states in line with the EIF[52]. The EIS combines a top-down approach through European policy development and coordination with a bottom-up, sectorial approach through projects and works through 3 clusters: trusted information exchange, interoperability architecture and the assessment of ICT implications of new EU legislation. The practical implementation of this sectorial approach is found in the program on Interoperability Solutions for European Public Administrations (the ISA program)[53]. This ISA program supports activities to facilitate cross-border digital collaboration between European public administrations, with the aim to make them more interoperable. This includes both public administrations from member states and EU institutions.
A further step in the top-down approach is the EIF, which defines an agreed approach to interoperability. It sets principles of and a conceptual model for European public services and describes interoperability levels, interoperability agreements and governance.

[48] http://www.epsiplatform.eu/
[49] European Commission, The European eGovernment Action Plan 2011-2015. Harnessing ICT to promote smart, sustainable & innovative Government, COM(2010)743, 15.12.2010, p. 4.
[50] European Commission, Towards interoperability for European public services, COM(2010) 744, annex I, 16.12.2010.
[51] European Commission, Towards interoperability for European public services, COM(2010) 744, annex II, 16.12.2010
[52] An overview of the progress can be found on http://www.daeimplementation.eu/dae_actions.php?action_n=26
[53] http://ec.europa.eu/isa/

Relevant for our subject of big data policies is especially the definition of interoperability levels: legal, organizational, semantic and technical interoperability. At all these levels barriers for interoperability can exist and solutions have to be developed.
Technical interoperability concerns the technical aspects of linking information systems. Organizational interoperability concerns how organizations cooperate to achieve their goals. In practice it implies aligning business processes and the related data exchange.
Legal interoperability concerns how to deal with differences in legislation. Administrations work within their national legal framework and such differences can make the collaboration between public services complex and even impossible. When information is exchanged across borders the legal validity of this information has to be maintained. In other words, public services encounter a similar problem as private actors trying to offer internet services across several countries under a license. Legal interoperability can imply the need to align legislation or the development of agreements between member states.
Especially relevant in the context of big data is semantic interoperability. The objective is to ensure that the precise meaning of exchanged information is understood and preserved throughout the data exchanges. It concerns developing descriptions and vocabularies concerning the exact format of information (grammar, format, schemas) as well as the meaning of data elements and their relations. In practice semantic interoperability is put in practice through developing sector-specific sets of data structures and data elements and making agreements on the meaning of the information these contain. Growing levels of semantic interoperability also enable to make better use of open government data, as it makes it easier to link otherwise isolated data sources.

Conclusion

The legal framework on access to documents and re-use of PSI reflects the shift of a transparency tool towards market regulation and the opening up of PSI as an economic resource. At the same time, the EU respected the political sensitivities surrounding openness and kept a soft approach towards the Member States. It left them the choice on how open they want to be. When information is opened, then the market logic demands equal conditions for all market players including public authorities.
This soft approach does not prevent the EU to take a proactive and stimulating role towards the Member States concerning enlarging the availability and quality of PSI. Through promoting open data policies the EU stimulates the Member States to participate in this data ecosystem and to play an enabling role in the data economy. This not only concerns data flows from the government to citizens and the private sector, but also between public authorities. Avoiding a fragmentation of the data economy and creating a single market implies stimulating interoperability in all its aspects.
The broad concept of interoperability developed in the EIF reflects that the newer concepts of the data value chain and data ecosystem. The real value of big data is not just a more developed technology, but that it enables to reap positive network effects. Effectively doing so is not just a question of technology or making data available, but also a question of making data usable (semantic interoperability) and of developing new business processes (organizational interoperability).
The link with organizational processes is less clear with the open data policies as such. Open data policies are not linked to a specific business process, but rather serve an enabling purpose. However, this link becomes visible again in the creation of user-driven release processes or stakeholder involvement in the management of data resources.

3.2.3 Data Protection

The current European data protection framework consists of the Data Protection-directive 95/46/EC concerning processing of personal data by private and public authorities in the member states. This directive is supplemented with the E-Privacy directive 2002/58/EC concerning electronic communications. Processing by the EU institutions is regulated by regulation 45/2001. These instruments are only applicable for what were first-pillar matters, and not when the processing occurs by public authorities as part of their activities concerning police, justice or external affairs. Exchange of personal data by police and justice authorities is regulated by the Council Framework Decision 2008/977/JHA on the protection of personal data processed in the framework of police and judicial cooperation in criminal matters.[54]
The general data protection-framework provided by directive 95/46/EC is currently under review. The Commission proposed a new General Data Protection Regulation (GDPR)[55], as well as a new directive for the processing of personal data by competent authorities in criminal matters[56].
Here we will limit our discussion to directive 95/46 and the GDPR, as these concern the applicable law for the case study sectors in the BYTE-project.

The legal framework concerning the protection of personal data has an important impact on policies concerning big data. This framework both limits and regulates the processing of personal data. Protection of personal data is based on the fundamental right to privacy, but has evolved into a framework of rights and duties which exceeds the right to privacy and has acquired the status of an autonomous fundamental right in itself. Both rights do partially overlap, but function with a different logic.
Gutwirth and De Hert point to two distinct legal tools to limit and control power. The first set are opacity tools, which set limits to the interference in individual matters. These tools shield certain areas and prohibit the interference. The right to privacy functions this way. The other set are transparency tools, which guarantee transparency and accountability of who interferes. These tools regulate and organize the exercise of power, in order for it to be legitimate. The data protection framework can be seen as such a transparency tool, although it also contains 'opacity elements', and regulates how the processing of personal data can be legitimate.[57]

Personal data

Trigger for the application of the data protection framework is the notion of personal data. In the data protection-directive 95/46 it is defined as “any information relating to an identified

[54] Council, Framework Decision 2008/977/JHA on the protection of personal data processed in the framework of police and judicial cooperation in criminal matters, 27 November 2008.
[55] European Commission, Proposal for a Regulation of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation), *Official Journal of the European Union*, C 102, COM (2012) 11, 5 April 2012. http://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX:52012PC0011 The legislative procedure and linked documents can be followed on http://eur-lex.europa.eu/procedure/EN/201286
[56] European Commission, Proposal for a Directive of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data by competent authorities for the purposes of prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and the free movement of such data, *Official Journal of the European Union*, C 102, COM/2012/010, 5 April 2012, http://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX:52012PC0010
[57] Gutwirth, Serge and Paul De Hert, “Regulating Profiling in a Democratic Constitutional State”, in Mireille Hildebrandt and Serge Gutwirth (eds.), *Profiling the European Citizen: Cross-Disciplinary Perspectives*, Springer Science + Business Media B.V. 2008, 271-293.

or identifiable natural person", with an identifiable person being defined as someone "who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity".[58] Recital 26 states that "to determine whether a person is identifiable, account should be taken of all the means likely reasonably to be used either by the controller or by any other person to identify the said person". In other words, the scope of the data protection framework can evolve when new techniques concerning identification in datasets become available, or when new data sources become public which makes such identification easier.
This notion of personal data is very wide, as it concerns 'any information' which can be linked to a natural person from the moment this person is identifiable. From the moment the datasets in use contain personal data the data protection framework is applicable and the big data policy has to be conform with it.
Also the range of activities to which the directive applies is very broad. The directive defines "processing" as "any operation or set of operations that is performed upon personal data, whether or not by automatic means, such as collection, recording, organization, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, blocking, erasure or destruction."
This means that any operation on information linked to an identifiable natural person has to comply with the data protection directive. Whenever the data in a big data-context contains information linked to an identifiable natural person, the processing has to be according to the data protection principles and the big data policy has to implement mechanisms to put the rights of data subject into practice.

A way to escape from the application of the data protection framework is to anonymize the data. Recital 26 states that "the principles of protection shall not apply to data rendered anonymous in such a way that the data subject is no longer identifiable". Adequate anonymization is not an easy task. The main criterion for adequate anonymization is the outcome: the data subject may no longer be identifiable by the controller or a third party. If such outcome cannot be guaranteed, the data protection framework remains applicable on the dataset. As said before, such identifiability is judged taking account of "all the means likely reasonably to be used". This implies that future developments can render anonymization techniques inadequate.
The Article 29 Data Protection Working Party (WP29)[59] identified 3 risks which have to be addressed through the anonymization techniques:
- singling out: the possibility to isolate records which identify an individual in the dataset
- linkability: the possibility to link 2 or more records concerning the same data subject (in one database or in different databases). When such linking would allow by combining several datasets to relate, with high probability, personal data to a specific person, the anonymization is broken. For instance, when health data from one dataset can be linked to specific individuals by using census data from another dataset.
- inference: the possibility to infer with high probability the value of an attribute based on a set of other attributes. For instance, when identification in a dataset is countered by

[58] European Parliament and the Council, Directive 95/46/EC on the protection of individuals with regard to the processing of personal data and on the free movement of such data, 24.10.1995, art. 2(a). A similar definition can be found in Regulation (EC) No. 45/2001 of the European Parliament and of the Council of 18 December 2000.
[59] The Article 29 Data Protection Working Party (WP29) is the advisory group, consisting of representatives of member state data protection authorities and the European Data Protection Supervisor (EDPS). It publishes regularly non-binding opinions.

generalization of birth dates to year of birth, but in certain age groups only one value for a specific attribute (e.g. a medical condition) is present or is dominant, it is possible to deduce this attribute with high probability for people with that age.
In her Opinion on Anonymisation Techniques, the WP29 gave an overview of the techniques currently in use and concluded that none of them could guarantee with certainty that all 3 risks of linking personal data to identifiable persons were prevented . It can only be decided on a case-by-case bases if certain anonymization techniques work adequately. The WP29 also warned for using pseudonymization as an anonymization technique. Pseudonymization replaces an attribute, which can be used as identifier, by another. Problem is that the granularity of the data remains the same. Pseudonymization can reduce linkability with other datasets, but the new attribute can as well be used as identifier.[60]
Especially in the big data-context (de-)anonymization is a vivid research topic and a range of failures are known[61]. This makes avoiding the application of the data protection framework through anonymization a question which needs careful technical analysis.

The Commission's GDPR proposal contains similar definitions of personal data as the old directive 95/46. On the other hand, the European Parliament added a new category of 'pseudonymous data', defined as "personal data that cannot be attributed to a specific data subject without the use of additional information, as long as such additional information is kept separately and subject to technical and organisational measures to ensure non-attribution"[62]. Pseudonymous data remains personal data and within the remit of the GDPR, but its use leads to some weakening of the obligations of the data controllers. This because of article 10, which stated in the Commission-proposal that "If the data processed by a controller do not permit the controller to identify a natural person, the controller shall not be obliged to acquire additional information in order to identify the data subject for the sole purpose of complying with any provision of this Regulation". The EP-proposal also included pseudonymous data in this article.[63] Access and other rights of data subject can be denied in case of anonymized data, but in the EP proposal also for pseudonymous data.
This addition of pseudonymous data is contested. As shown by the WP29 pseudonymization is not an adequate anonymization technique. On the other hand, the definition of pseudonymous data adds some extra limitations to the technique of pseudonymization and demands measures to ensure non-attribution. It can be questioned what the difference with anonymization still is. Main lack of clarity in the definition is what 'attributing' to a specific data subject means, compared to being 'identifiable' in the definition of personal data.
To get a more definitive view on the reach of the data protection framework, we will have to wait for the end of the legislative process. In general we can conclude that the use of personal data, that is any operation on information linked to an identifiable natural person, has to comply with the data protection directive. But adequate anonymization allows to avoid the application of this framework.

[60] WP29, Opinion 05/2014 on Anonymisation Techniques, 10 April 2014
[61] Narayanan, Arvind and Vitaly Shmatikov, *Robust de-anonymization of large sparse datasets.(How to Break Anonymity of the Netflix Prize Dataset)*, 5 Feb 2008, http://arxiv.org/pdf/cs/0610105v2.pdf; Arvind Narayanan, Elaine Shi, Benjamin I. P. Rubinstein, *Link Prediction by De-anonymization: How We Won the Kaggle Social Network Challenge*, 22 Feb 2011, http://arxiv.org/pdf/1102.4374v1.pdf
[62] European Parliament, Legislative resolution on the proposal for a regulation of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation), 12 March 2014, Amendement 98.
[63] Ibid, Amendement 104.

Data protection principles

The Directive 95/46 puts forward data protection principles to which any processing of personal data has to conform, like:
- legitimacy : all data processing has to have a legal base. The directive lists several criteria based on which the processing can happen legitimately. Main ground is the consent of the data subject. Much stricter criteria apply for a specific set of sensitive data, that is data concerning "racial or ethnic origin, political opinions, religious or philosophical beliefs, trade-union membership, and ... health or sex life"[64].
- finality : all data collection and processing has to be done for "specified, explicit and legitimate purposes". This links data collection and processing to specific purposes, specified at the moment of collection or earlier. Further processing for other purposes is not allowed, and needs a new ground of legitimation (e.g. consent that the data also can be used for this new purpose).
- proportionality and relevance: the personal data must be "adequate, relevant and not excessive in relation to the purposes for which they are collected and/or further processed". Only personal data that is useful and only the minimal amount needed may be processed. Also, when the data is not necessary anymore, it must be erased or kept in a form which does not allow identification.
- accuracy : the data must be correct and up-to-date. Inaccurate data has to be corrected or erased.
- transparency : What happens to his personal data has to be transparent to the data subject. This implies that the data subject has the right to get from the controller information on the identity of the controller, the purposes of and the logic behind the processing and who receives the data.
- data subject participation and control : The data subject has the right to access the data, to object to certain processing of personal data or to demand the rectification, erasure or blocking.
- data security: the data has to be kept secure to avoid unauthorized or illegitimate access, transfer or processing.

The directive further implements the rights of data subjects, like the right to information about the data processing, the right to access the data, the right to object and rectification. It also specifies the obligations of data controllers, like assuring the confidentiality and the security of the personal data and notifying or prior checking of automated processing to the supervisory authority. The directive foresees a control mechanism through the establishment of independent supervisory authorities. These data protection authorities have powers to investigate, to intervene and to start legal proceedings against violations of the data protection laws.

The draft versions of GDPR contains generally the same principles, but provides more detailed implementations. The right to be forgotten can at first sight looks new, but the right to obtain erasure of personal data in certain circumstances is already included in Directive 95/46. The European Court of Justice based its decision concerning the right to be forgotten in the Google-case on this right, in combination with a weighting of the interests involved as part of criteria for legitimate processing.[65] Another strengthening, in this case of the right to

[64] European Parliament and the Council, Directive 95/46/EC on the protection of individuals with regard to the processing of personal data and on the free movement of such data, 24.10.1995, art. 8(1).
[65] EUCJ, *Google Spain SL, Google Inc. v Agencia Española de Protección de Datos (AEPD), Mario Costeja González*, C-131/12, 13 May 2014.

access, is the right to data portability. This gives the data subject the right to obtain a copy of its personal data in a structured and commonly used format.

Also relevant in the big data-context is that Directive 95/46 contains a regulation of automated decisions. Such decision is defined as "a decision which produces legal effects concerning him or significantly affects him and which is based solely on automated processing of data intended to evaluate certain personal aspects relating to him, such as his performance at work, creditworthiness, reliability, conduct, etc."[66]. It grants the person the right to object to being subjected to such automated decision making. But it also allows that a person is subjected to such a decision if it is taken in the course of the entering into or performance of a contract or if it is authorized by a law, and if measures to safeguard the person's legitimate interests (like the opportunity to present his point of view) are present. Further the data subject has the right to know the logic behind the automated decision making concerning him.
The draft GDPR, in both the Commission as the EP version, contains a similar regulation of profiling, augmented with stricter conditions to avoid discrimination.

The implementation of these principles and the rights of data subjects in the big data context poses a lot of practical difficulties and can be a major hurdle for big data-projects. The Google-case shows how extensive the application of the data protection framework is. In this case the EUCJ made clear that this framework is also applicable on search engines. Collecting data on the internet, storing it and organizing it through indexing and information retrieval techniques are all operations which qualify as processing subjected to the data protection framework when it concerns personal data.[67]

The draft GDPR also contains major innovations concerning the obligations of the data controllers (like maintaining of documentation of all processing and the duty to notify data breaches) and the control and enforcement mechanisms. We cannot discuss all innovations in this text, but important for the big data-context are new obligations to assure that data protection is considered from the early development of data processing projects. Data protection by design and by default is one of new principles. A mechanism to assure this is the introduction of a data protection impact assessment, which is obliged when "the processing operations present specific risks to the rights and freedoms of data subjects by virtue of their nature, their scope or their purposes". The draft mentions operations similar to profiling, the monitoring of publicly accessible areas, or cases when sensitive information is involved. The EP draft augments this with an initial risk analysis for a wider range of operations, and a periodical data protection compliance review.

International transfer of personal data

The directive 95/46 aimed at enhancing the single market and lifting data protection as a barrier through a harmonization of data protection rules. Such barriers remain when personal data gets transferred outside of the EU.
A transfer of personal data to a third country outside of the EU is only allowed if that third country ensures an adequate level of protection. Such adequacy is determined by the Commission following a procedure established by the directive. Derogations from this

[66] European Parliament and the Council, Directive 95/46/EC on the protection of individuals with regard to the processing of personal data and on the free movement of such data, 24.10.1995, art. 15(1).
[67] EUCJ, *Google Spain SL, Google Inc. v Agencia Española de Protección de Datos (AEPD), Mario Costeja González*, C-131/12, 13 May 2014, §28.

principle is possible in a range of circumstances listed in the directive. Derogation is also possible when a member state considers that the data controller offers adequate safeguards. In this case adequacy is determined not on the data protection framework of a country, but on the data protection safeguards implemented by a specific data controller. The directive foresees that such adequacy can result from appropriate contractual clauses.[68]
The draft GDPR preserves this transfer regime. It includes a new option of 'Binding corporate rules' (BCR) developed in the practices of the data protection authorities. These BCR are internal rules adopted by a multinational group of companies with regard to the international transfers of personal data within the same corporate group to entities located in countries which do not provide an adequate level of protection. The approval of such global policy avoids the need for distinct adequacy investigations for each specific transfer of personal data within a this corporate group.

This regime of international transfer of personal data establishes a patchwork of rules to which big data processing involving personal data is subjected to, depending on the origins of the data and the destination where the processing takes place. The development of new instruments like the binding corporate rules give space for new private orderings of data flows, while remaining subjected to control by data protection authorities in diverse jurisdictions.

The Commission considers a limited range of countries as having adequate protection.[69] One of those regimes of adequate protection is the Safe Harbor regime established by the US Department of Commerce in negotiation with the EU. This allows transfer of personal data to the US by organizations which adhere to the Safe Harbor Privacy Principles. These principles require:[70]

- Notice: Organizations must notify individuals about the purposes for which they collect and use information about them.
- Choice: Organizations must give individuals the opportunity to choose if their personal information is used or disclosed through an opt out in general, or an opt in for sensitive data.
- Transfers to Third Parties: The notice and choice principles also have to be applied for transfers to third parties.
- Access: Data subjects have a right to access their personal data held by an organization and demand correction or deletion
- Security
- Data Integrity
- Enforcement: Adequate procedures have to be available for complaints, review of compliance and remediation.

US organizations notify the US Department of Commerce that they will adhere to these principles. The Federal Trade Commission (FTC) enforces compliance by the organizations which have registered.

The creation of international regimes for the transfer of personal data comes down to the creation of specific data ecosystems. It allows organizations to participate in a specific data

[68] European Parliament and the Council, Directive 95/46/EC on the protection of individuals with regard to the processing of personal data and on the free movement of such data, 24.10.1995, art. 25-26.
[69] European Commission, Commission decisions on the adequacy of the protection of personal data in third countries, http://ec.europa.eu/justice/data-protection/document/international-transfers/adequacy/index_en
[70] US Department of Commerce, U.S.-EU Safe Harbor Overview, http://export.gov/safeharbor/eu/eg_main_018476.asp

economy and to build data value chains using personal data within that environment. Or it locks them out. The specific regimes, like Safe Harbor, have a strong impact on how big data processing can be done.

Conclusion

The current European approach to data protection clearly poses challenges to big data processing and vice versa. Big data processing puts into question the sustainability of the basic concept of data protection: personal data, being distinguishable from other, non-personal data. Also the application of the data protection principles in actual big data processing proves difficult.
On the other hand, difficult is not the same as impossible. If impossible, the data protection regime can be considered 'broken'. In that case big data processing with personal data would lead to a contradiction with the data protection framework. However, the response of the data protection authorities shows that the data protection principles can be applied on big data processing without such contradiction.
This does not exclude that the combination data protection and big data processing can be difficult. The question if data protection is over-protective and seriously and unnecessarily hinders such processing, is again a political question demanding the balancing between interests. When the legislative process on the new GDPR has ended, we will see where the balance is struck. This question is not just a political tradeoff, as further technical developments and research can help to give body to more privacy-friendly methods of big data-processing.

Just like the other legal frameworks, data protection creates a patchwork of jurisdictions. This patchwork can be a barrier to combining diverse sources in big data processing and lead to distinct data ecosystems. How serious this risk is, depends on the variation in data protection rules. As we will see in the next chapter, the difference with the US system is quite substantial.
Methods exist to overcome these barriers. The Safe Harbor-regime is an important example thereof, as well as how through such brokerage elements of EU law get a wider territorial application. Binding corporate rules are another instrument to eliminate barriers with a corporate group. These methods lead to a re-ordering of the borders of the data ecosystem and are therefore very important for big data processing.

3.3 SECTORIAL LEGAL FRAMEWORKS AND DATA POLICIES

Specific legal frameworks and data policies have been developed on sectorial level. Here we focus on geospatial, environmental and scientific research data. These are sectors where big data processing plays an important role and where interoperability mechanisms and the data ecosystem is most developed. We will look into the state of affairs concerning the access and re-use policies and how far they reflect the principles found earlier in the policy documents.

3.3.1 INSPIRE

Article 17(1) of the INSPIRE Directive[71] requires each Member State to adopt measures for the sharing of spatial datasets and services between its public authorities. In particular, those measures shall enable those public authorities to gain access to spatial data sets and services, and to exchange and use those sets and services, for the purposes of public tasks that may have an impact on the environment.

Access to spatial data and services constitutes an important basis for environmental policies for all public authorities and is therefore a central aspect of INSPIRE. Since the Community institutions and bodies in most cases have to integrate and assess spatial information from all the Member States, INSPIRE recognizes the need to be able to gain access to and use spatial data and spatial data services in accordance with an agreed set of harmonised conditions.

Article 17(8) of the INSPIRE Directive requires the development of a Regulation on the provision of access to spatial datasets and services from Member States to the institutions and bodies of the Community. The main points of the Regulation[72] are the following:

- Metadata must include the conditions applying to access and use for Community institutions and bodies; this will facilitate their evaluation of the available specific conditions already at the discovery stage.
- Member States are requested to provide access to spatial data sets and services without delay and at the latest within 20 days after receipt of a written request; mutual agreements may allow an extension of this standard deadline.
- If data or services can be accessed under payment, Community institutions and bodies have the possibility to request Member States to provide information on how charges have been calculated.
- While fully safe-guarding the right of Member States to limit sharing when this would compromise the course of justice, public security, national defence or international relations Member States are encouraged to find the means to still give access to sensitive data under restricted conditions, (e.g. providing generalized datasets). Upon request, Member States should give reasons for these limitations to sharing.

In parallel with the definition of the formal Regulation, guidelines have been developed. They are meant to help Member States to implement the Regulation on INSPIRE Data and

[71] European Parliament and the Council, Directive 2007/2/EC of 14 March 2007 establishing and Infrastructure for Spatial Information in the European Community (INSPIRE), OJ L 108, 25 March 2007.
[72] European Commission, Commission Regulation (EU) No 268/2010 of 29 March 2010 implementing Directive 2007/2/EC of the European Parliament and of the Council as regards the access to spatial data sets and services of the Member States by Community institutions and bodies under harmonised conditions, OJ L 83, 30 March 2010.

Service Sharing, and include non-binding instruments, such as model contracts, and illustrate related concepts such as framework agreements.

As regards public access to data and services supplied under INSPIRE, the guideline[73] is that, if no provisions are contained in the agreement between Member States and the Community institutions and bodies, then access given by the institutions and bodies should be guided by whether public access is already, or could be, allowed in the Member State and under what conditions.

When this public access to spatial data sets or services cannot be allowed, due to an exemption provided for by law, data producers are encouraged to state the conditions under which such an access is possible, for example by removing sensitive information, downgrading the accuracy or restricting the size of the download possible. It is also suggested that any such measures are, as far as possible, harmonised within and between Member States, so that they can effectively be applied to aggregated data sets that potentially may come from a from large number of producers.

Public access to Member States' spatial datasets and services could be of interest to Member States and the public in general when Community institutions and bodies produce aggregated spatial data sets or services based on data coming from several Member States. In this case the resulting datasets offer added information value in that they provide access to spatial information collated at a European level. Public access should therefore be promoted as much as possible, while respecting any exemptions provided for by law.

INSPIRE provides examples of good practice related to sharing within and between Member States. Examples of identified good practices include: the Catalonia Spatial Data Infrastructure, the Spanish Cadastre, and the Hungarian Land Registry.

A list of topics considered particularly critical to a successful data- and service-sharing arrangement has been established, and for every topic criteria for good practice have been defined. Topics covered include public access, defined as: *"the ability of the public to discover, view and download information and data and to use available services and data. Public authorities have an obligation to provide INSPIRE data through online services to their citizens under national legislation based on the INSPIRE Directive.*

It is important for the citizens to obtain easy access to the information they are looking for. Citizens should be able to easily find information, to view the spatial data sets and to use the spatial data sets and services without too much difficulty. The public authorities should make their data and services available in a way that makes it easy for the citizen to obtain access. Use conditions and charges should be presented in an understandable way."[74]

The following criteria indicate that one has a good practice for public access:

- ***Awareness by public that data and services exist –*** The public knows where it can find data and services, i.e. there is a central portal with registries and search engines that allows the citizen to find out where to go to obtain access to data or services. Awareness raising activities are promoted also through other means (e.g. flyers...).

[73] European Commission, Guidance on the 'Regulation on access to spatial data sets and services of the Member States by Community institutions and bodies under harmonised conditions', 9 January 2013 (revision), p. 11.
[74] European Commission, *Good practice in data and service sharing*, 9 January 2013, p. 46.

Increasing awareness of the public usually will be reflected in increasing access to this website.

- ***Clear process for public to access data and services*** – The public authorities provide clear and user-friendly information on how the citizens can obtain access to data and services and under which conditions and charges they can do so. This information is also provided on-line, with contact details for obtaining more information.
- ***Online access wherever possible*** – Citizens can also obtain access to data online rather than via a paper copy, a digital copy on CD or a consultation on site.

All this shows that it is about more than bringing data together into a big data collection. The INSPIRE-directive sets up a framework concerning access, re-use and sharing of geospatial data. But it enhances this regulation of data availability with the other levels of interoperability identified in the EIF. It has a strong attention for semantic interoperability and usability of the data. Elements of organisational interoperability are present through the development of network services and the setting up of management structures with a wide range of public stakeholders. The directive further includes the public sector in a wider data ecosystem with the private sector and general public by defining the available network services.

3.3.2 Environmental data

We have considered two major Earth Observation initiatives, GEOSS and Copernicus.

GEOSS

The Global Earth Observation System of Systems (GEOSS) is intended as a global and flexible network of content providers allowing decision makers to access an extraordinary range of data and information at their desk. GEOSS is a system of systems that interconnects more than thirty autonomous infrastructures. It will be operational in 2015, however, it already allows discovering and accessing more than 70 million of extremely heterogeneous datasets. As such, GEOSS had and has to face several challenges related to Big Data.[75]

To realize a system of systems, GEOSS implements a brokering architecture. This innovative approach demonstrated successful in addressing those issues related to big data "Variety". This is actually perceived as the most important challenge for Earth Science and Observation infrastructures. More generally, "Variety" is the most relevant aspect in multidisciplinary systems of systems, and that will probably increase in future with the integration of crowdsourcing infrastructures. Leveraging cloud infrastructure capabilities, brokering platforms make also possible handling of big data "Volumes" issues, in terms of large amount of datasets accessible. GEOSS demonstrated the relevance of "Veracity", "Value", and "Validity" aspects for selection, ranking, evaluation, and eventually usage of datasets. In particular, ranking algorithms was fruitfully used to return meaningful and usable results of query to Users, avoiding raising interface complexity.

GEOSS explicitly acknowledges the importance of data sharing in achieving the GEOSS vision and anticipated societal benefits: *"The societal benefits of Earth observations cannot*

[75] Stefano Nativi, Paolo Mazzetti, Mattia Santoro, Max Craglia, Osamu Ochiai, *Big Data Challenges in building the Global Earth Observation System of Systems*, In publication.

be achieved without data sharing"[76]. To that extent, the Implementation Plan sets out a set of Data Sharing Principles for full and open exchange of data:

- There will be full and open exchange of data, metadata and products shared within GEOSS, recognizing relevant international instruments and national policies and legislation;
- All shared data, metadata and products will be made available with minimum time delay and at minimum cost;
- All shared data, metadata and products being free of charge or no more than cost of reproduction will be encouraged for research and education.

GEOSS introduces the notion of Data Collection of Open Resources for Everyone (Data-CORE), a distributed pool of documented datasets with full, open and unrestricted access at no more than the cost of reproduction and distribution. Data CORE has been a key mechanism to address the limitations identified in implementing the Sharing Principles and there has been a big push last year to increase the stock of the CORE, leveraging the voluntary nature of GEOSS. GEO Members are invited to encourage data providers to abide by the Data-CORE terms in publishing their datasets.

Copernicus

Copernicus is a European system for monitoring the Earth. Copernicus consists of a complex set of systems that collect data from multiple sources: earth observation satellites and in situ sensors such as ground stations, airborne and sea-borne sensors. It processes these data and provides users with reliable and up-to-date information through a set of services related to environmental and security issues.

The services address six thematic areas: land, marine, atmosphere, climate change, emergency management and security. They support a wide range of applications, including environment protection, management of urban areas, regional and local planning, agriculture, forestry, fisheries, health, transport, climate change, sustainable development, civil protection and tourism.

The main users of Copernicus services are policymakers and public authorities that need the information to develop environmental legislation and policies or to take critical decisions in the event of an emergency, such as a natural disaster or a humanitarian crisis.

Based on the Copernicus services and on the data collected through the Sentinels and the contributing missions, many value-added services can be tailored to specific public or commercial needs, resulting in new business opportunities. In fact, several economic studies have already demonstrated a huge potential for job creation, innovation and growth.

The Copernicus programme is coordinated and managed by the European Commission. The development of the observation infrastructure is performed under the aegis of the European Space Agency for the space component and of the European Environment Agency and the Member States for the in situ component.

[76] Group on Earth Observations, *10-Year Implementation Plan Reference Document*, ESA Publications Division, Noordwijk (The Netherlands), February 2005, p. 139, 205.

The Member States and the European Parliament have mandated the EC to define the overall Copernicus data and information policy. The Copernicus data policy Regulation takes full and open access to information produced by GMES services and data collected through GMES infrastructure as the basic principle. Security restrictions and licensing conditions, including registration, may limit the general principle. Free of charge or COFUR (Cost of Fulfilling User Requests) is envisaged as well. Purpose of use is not considered.

The Copernicus Regulation on the access to GMES dedicated data and GMES service information implies a commitment to follow the GEOSS Data Sharing.

3.3.3 Scientific data

We have considered CERN, JRC, ERC, RDA.

CERN

CERN produces extremely large volumes of data and is commonly mentioned in the context of Big Data, where data volume is a specific, prominent issue. Much of the practice consists of large-scale, collaborative efforts, with tens or even hundreds of international partners, in some instances over a long period of time, which generate massive data that is stored for further processing. Analysis of the vast quantities of data cannot be undertaken with a single desktop computer or a single large supercomputer, consequently Grid or Cloud technologies are used for analytical purposes.

As an example, the CERN Large Hadron Collider (LHC) Computing Grid, the world's largest computing grid, produces about 15 petabytes of data per annum. The LHC Compact Muon Solenoid[77] (CMS) experiment, one of the largest international scientific collaboration in history, involving 4.300 particle physicists, engineers, technicians, students and support staff from 179 universities and institutes in 41 countries, and one of the two general-purpose experiments at CERN's LHC that have been built to search for new physics, has collected so far around 64 petabytes of raw data from the collisions taking place every second, at peak performance, inside its detector.

In its policy[78], CMS upholds the principle that open access to the data will, in the long term, allow the maximum realization of their scientific potential. CMS data is classified into four levels in increasing order of complexity of information, which map on to different policies for re-use and preservation. In particular, CMS provides open access to its data at different points in time after a suitable embargo period, allowing CMS collaborators to fully exploit their scientific potential. For the widest possible re-use of the data, while protecting the Collaboration's liability and reputation, data are released under the emerging standard Creative Commons CC0 waiver. Data will also be identified with persistent data identifiers, and it is expected that the third parties cite the public CMS data through these identifiers, so that its re-use can be monitored and contribute to the assessment of the impact of the LHC program. The release of data could create a community of users that may be nurtured through regular events organized by CMS, allowing further monitoring of the data re-use.

[77] European Organization for Nuclear Research, "Compact Muon Solenoid experiment at CERN´s LHC", no date. http://cms.web.cern.ch/

[78] CMS experiment, "CMS data preservation, re-use and open access policy", no date. https://cms-docdb.cern.ch/cgi-bin/PublicDocDB/RetrieveFile?docid=6032&version=1&filename=CMSDataPolicy.pdf

JRC

The JRC coordinates the scientific and technical development of INSPIRE, investigating issues regarding technical and multidisciplinary interoperability of spatial datasets and services needed to support environmental policy and policies that affect the environment, and what are the challenges in sharing and providing access to data from a variety of sources, and in a variety of formats. As a consequence, the JRC has a specific interest in Public Sector Information, which is in the main scope of INSPIRE.

As regards Open Access, it must be born in mind that most of the data held by the JRC is not owned by the JRC, but by the Member States that provide access to the JRC for specific projects, or as part of legal requirements. Therefore the JRC would have to respect data ownership on the matter of Open Access, as constrained by third party IPR.

Being part of the European Commission, JRC refers to the official EC practice, so the uptake of Open Access to research data is in a very initial phase, and no direct experience is reported. Next year an open data project will be initiated. One aim of the project is to increase access to JRC data (including research data) following the commission decision of 12 December 2011 on the reuse of Commission documents.[79]

Currently, JRC has no common policy for data management. That would have to be defined in the context of the open data project. Likewise, there is no auditing in place for data utilization. A first approach to that is to establish a common data inventory.[80]

ERC

The ERC Scientific Council Guidelines for Open Access (17 December 2007) state: "free and efficient access to information, including scientific publications and original data, will be the key for sustained progress. It is indicated as essential that primary data, as well as data-related products such as computer codes, are deposited in the relevant databases as soon as possible, preferably immediately after publication and in any case not later than six months after the date of publication."

RDA

The Research Data Alliance (RDA) Big Data Analytics (BDA) Interest Group seeks to develop community-based recommendations on feasible data analytics approaches to address scientific community needs of utilizing large quantities of data. The objectives are complementary to the recent activities performed by NIST[81] Big Data working groups and supports formation of different working groups to tackle specifics problems.

The overall objectives of the group are:

[79] European Commission, Commission Decision of 12 December 2011 on the reuse of Commission documents (2011/833/EU), OJ L 330, 14 December 2011.
http://eurlex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2011:330:0039:0042:EN:PDF
[80] This part on JRC is based on the research done in the RECODE-project. Lorenzo Bigagli et al., *Infrastructure and technology challenges*, RECODE Deliverable D2.1, 31 March 2014, pp. 56-57.
[81] http://www.nist.gov/

- BDA will aim to clarify some foundational terminologies in the context of data analytics understanding differences/overlaps with terms like data analysis, data mining, etc.
- BDA will develop a recommendation documents with a systematic classification of feasible combinations of analysis algorithms, analytical tools, data and resource characteristics and scientific queries. These recommendation documents can serve as a best practice guide for scientific groups/communities interested in investing in Big Data technologies
- BDA will work towards a consensus amongst its members to achieve this desired goal
- BDA will collaborate with external bodies and initiatives - such as NIST, OGC, ISO, and others - for exchange; to this end, BDA findings will be made available for download publicly and freely. Specifically, our interaction with NIST Big Data working groups will entail:
    - BDA will encourage discussions linking NIST activities and reference architecture-related terminologies. BDA will focus on analysing selected use cases (with overlap with NIST [1]) targeting data analytics/analysis approaches (e.g. map-reduce, array databases, computational statistics, etc.)
    - BDA will support formation of working groups from NIST and BDA to work together on subtopics.

3.3.4 Geospatial data

We have considered ESA and EUMETSAT. The Copernicus policy is also relevant to this sector, as regards the space component of the Copernicus programme.

ESA

ESA data policy applicable to the various sensors is, in general, different and does impose limitations on re-use. However, as regards Sentinel, its last satellite constellation, ESA shares the ERC approach, stating that access to Sentinel data should be free, full and open.

The data volume of Sentinel-1, -2, -3 A-series production is roughly equivalent to 25 Envisat missions[82]. As the space component of Copernicus, the Sentinel data policy has been jointly decided by ESA and EC[83]. ESA Member States and EC have prepared joint principles[84] of a Sentinel data policy in Sep 2009. The Copernicus regulation[85] has been adopted in April 2014 and, among others, it contains the following provisions:

- Access to Sentinel data by anybody (European and non-European users) and for any use ("full and open")
- Free of charge data licenses ("free")
- Some restrictions may be required (e.g. security, technical constraints, etc.)

[82] European Space Agency, *Sentinel Data Policy and Access to Data. Workshop on GMES Data and Information Policy*, Brussels, 12-13 January 2012.
[83] ESA Unclassified, Sentinel-2 Preparatory Symposium, April 2012, slide 9; in Yves-Louis DESNOS, *The GMES/Copernicus Sentinels Missions and their Exploitation for Science and Applications*; https://earth.esa.int/documents/10174/642954/ESASentinels062013.pdf
[84] European Space Agency, *The Joint Principles for a Sentinel Data Policy*, ESA/PB-EO(2009)98, rev. 1
[85] European Parliament and the Council, Regulation 377/2014 establishing the Copernicus Programme and repealing Regulation (EU) No 911/2010, 3 April 2014.

EUMETSAT

The EUMETSAT Data Policy (January 2013) states:

"I. National Meteorological Services ("NMSs") of the Member States will receive all EUMETSAT data, products and services for their Official Duty use at no cost except for the cost of decryption key units."

"IV. A set of data, products and services to be determined by Council will be available on a free and unrestricted basis as "Essential" data and products in accordance with WMO Resolution 40 (Cg-XII)." ... as well as other sets of data available for research and educational use, Official Duty use by non- member states, commercial users, and "all others" who may have to pay for the data and be bound by conditions/restrictions on use.

3.3.5 Conclusions

The INSPIRE-directive, regulating the access, re-use and sharing of geospatial data, reflects the data ecosystem approach as it concerns more than data availability. Starting point is the sharing of data between public services, but it includes this in a wider data ecosystem with the private sector and general public by defining the available network services. The directive also reflects a broad approach to interoperability, with strong attention for semantic interoperability and usability of the data. The sector of geospatial data can therefore be considered as a testing bed for an approach to data which got reflected in the EIF.

Each of the reviewed institutions which produces large datasets (GMES/Copernicus, CERN, ESA, EUMETSAT) has adopted open access policies in general or partially, although sometimes limited for security concerns or IPR restrictions. All are data providers for a wide range of purposes and include a strong attention for semantic interoperability. Exception is JRC, which has no general data policy but is for most of its data dependent on the data policy defined by the Member State providing the data. Further we notice strong recommendations by ERC and RDA for open access. We can conclude that the general preference towards open data has been widely adopted.

4 DATA POLICIES IN EU MEMBER STATES

In this chapter we review data policies in a sample of EU Member States, in the context of the European legal frameworks reviewed earlier. We choose these countries as they represent a range from trendsetter UK to a mid-range follower Belgium in the Cap Gemini ranking of Open Data initiatives.

4.1 UNITED KINGDOM

The UK is one of the forerunners concerning the development of a policy on big data, including the adaptation of legal frameworks to enable the development of big data applications. Policy initiatives and debate is ongoing on all areas concerning big data.

IPR

The implications of big data for IPR were considered in the Hargreaves report of May 2011. IPR in the UK is based on the EU and international law discussed above, but the national implementation of these frameworks brings its own peculiarities. The UK has made very limited use of the possibility to include exceptions in IP law. For instance, the parody does not exist.
The Hargreaves report considered IP law outdated. Specifically on copyright it stated that normal daily uses involves now copying and that the existing framework turned ordinary customers into copyright violators when they copy some content from one carrier to another. It signalled a similar problem with text and data mining. Therefore it recommended to broaden the use of exceptions in UK law and to implement all opportunities for such exceptions provided by the EU framework. This included broadening the already existing exceptions for library archiving and for non-commercial research, where the formulation was too narrow to allow new methods like text and data mining. It also recommended that the UK government promotes at EU level a change in legislation providing for an exception to support text and data analytics, as well as a further copyright exception designed to build into the EU framework adaptability to new technologies. Further it recommended to make exceptions mandatory to avoid legal confusion.[86]

In its response the UK government endorsed these recommendations and planned to table proposals enlarging the UK's copyright exceptions by amendments to the Copyright, Designs and Patents Act 1988.[87] One set of new exceptions were approved by parliament on 14 May and came into force on 1 June 2014. These included exceptions for research, private study and text and data analysis for non-commercial research. Copyright is not violated when a "copy is made in order that a person who has lawful access to the work may carry out a computational analysis of anything recorded in the work for the sole purpose of research for a non-commercial purpose, and ... the copy is accompanied by a sufficient acknowledgement (unless this would be impossible for reasons of practicality or otherwise)"[88]. Further it also

[86] Ian Hargreaves, *Digital Opportunity: Review of Intellectual Property and Growth*, May 2011. http://www.ipo.gov.uk/ipreview-finalreport.pdf, pp. 41-52.
[87] UK Government, *Government Response to the Hargreaves Review of Intellectual Property and Growth*, 3 August 2011, http://www.ipo.gov.uk/ipresponse-full.pdf, p. 8.
[88] Copyright, Designs and Patents Act 1988, section 29A, as introduced by the Copyright and Rights in Performances (Research, Education, Libraries and Archives) Regulations 2014, Statutory Instruments 2014, n° 1372, http://www.legislation.gov.uk/uksi/2014/1372/contents/made

included exceptions allowing for digitization by libraries and archives. Another set of exceptions for quoting, parody and private copies are still under treatment.[89]
In its response the UK government also planned to raise enlarging the exceptions provided by EU-law at European level. In its international strategy on IPR the UK government included the aim "to secure further flexibilities at EU level that enable greater adaptability to new technologies".[90]

Data protection

Data protection is overseen by the Information Commissioner's Office (ICO), which oversees a wide range of information-related laws like the Data Protection Act 1998, the Privacy and Electronic Communications Regulations 2003, the Freedom of Information Act 2000, the Environmental Information Regulations 2004, and the INSPIRE Regulations 2009. All these acts are implementations of EU directives discussed earlier.
The ICO recently issued guidance concerning the application of data protection on big data. It contains advice on how to apply data protection principles like fairness, consent, purpose limitation and data minimization, etc in a big data context. It also discusses some tools for compliance, like Privacy Impact Assessments (PIA), privacy by design, transparency and privacy notices. The guidance also considers the application of these principles as foreseen by the draft GDPR. Further the ICO confronts the view that big data presents a fundamental challenge to the current data protection principles. ICO states its view that these principles "are still fit for purpose in the big data world"[91].
ICO has earlier given attention to and guidance concerning big data-related issues and data protection. In 2012 it published Guidance on the use of cloud computing[92], which focussed mostly on the responsibility of the data controller when using cloud services, and an Anonymisation code of practice[93]. ICO is also involved in the UK Anonymisation Network (UKAN), which establishes best practice in anonymisation and provides expert advice on techniques.[94] UKAN is funded by ICO and co-ordinated by a consortium of the Universities of Manchester and Southampton, the Open Data Institute and the Office for National Statistics.

Re-use of PSI and open data

The UK is also a forerunner in open data. Cap Gemini ranks the UK as most trend-setting nation concerning open data, with both the highest availability of datasets and usability of its data portal among EU countries.[95]

---

[89] Intellectual Property Office and Viscount Younger of Leckie„*Changes to copyright law*, 13 June 2014. https://www.gov.uk/government/news/changes-to-copyright-law
[90] Intellectual Property Office, *The UK's International Strategy for Intellectual Property*, 11 August 2011, p. 13. http://www.ipo.gov.uk/ipresponse-international.pdf
[91] Information Commissioner's Office, *Big data and data protection*, 2014, p. 41. http://ico.org.uk/for_organisations/data_protection/topic_guides/~/media/documents/library/Data_Protection/Practical_application/big-data-and-data-protection.pdf
[92] Information Commissioner's Office, *Guidance on the use of cloud computing*, 2012.
[93] Information Commissioner's Office, *Anonymisation: managing data protection risk code of practice*, 2012. http://ico.org.uk/for_organisations/data_protection/topic_guides/~/media/documents/library/Data_Protection/Practical_application/anonymisation-codev2.pdf
[94] http://ukanon.net
[95] Dinand Tinholt, *The Open Data Economy . Unlocking Economic Value by Opening Government and Public Data*, Cap Gemini, 2013, p. 5.

Opening up PSI was high on the government's agenda and it devoted new attention to open data, but also other options of re-using PSI. Objectives are first of all economic with the development of UK-based economic players in the data economy. Opening up PSI provides a home market with opportunities for such players. Other objectives are a more efficient government and stimulating scientific research as well as the development of the scientific base for a data economy. The economic aims have to be balanced with economic interests of certain governmental bodies in commercializing their data assets. Therefore the focus is not just on open data, but also on the wider PSI policy like charging policies.
Data.gov.uk is the main repository of open datasets and was launched in January 2010. At the start departments received directives to identify and release the most valuable datasets. This was later combined with the development of a user-driven and bottom-up process, in which datasets are prioritised for release based on user requests. After a recent review this policy is further augmented with the identification of datasets as part of the 'National Information Infrastructure' according to criteria concerning their economic and social value, their importance for the efficiency and accountability of the government and their use for connecting with other datasets.[96]
An institutional framework driving the data policy was created, first with the Public Sector Transparency Board to give advice on the general policy to open and publish data. Later also a Data Strategy Board (DSB) which has to give advice on which data to publish as open data, itself receiving advice from the Open Data User Group (ODUG). This is accompanied with planning at departmental level, where each government department publishes its Open Data Strategy providing its planning for the release of data over the next two years.[97]

The development of these open data strategies were guided by a set of Public Data principles:
"(1) Public data policy and practice will be clearly driven by the public and businesses that want and use the data, including what data is released when and in what form
(2) Public data will be published in re-usable, machine-readable form
(3) Public data will be released under the same open licence which enables free re-use, including commercial re-use
(4) Public data will be available and easy to find through a single, easy-to use, online access point (www.data.gov.uk)
(5) Public data will be published using open standards, and following relevant recommendations of the World Wide Web Consortium (W3C)
(6) Public data from different departments about the same subject will be published in the same, standard formats and with the same definitions
(7) Public data underlying the Government's own websites will be published in re-usable form
(8) Public data will be timely and fine-grained
(9) Release data quickly, and then work to make sure that it is available in open standard formats, including linked data forms
(10) Public data will be freely available to use in any lawful way
(11) Public data will be available without application or registration, and without requiring details of the user
(12) Public bodies should actively encourage the re-use of their public data
(13) Public bodies should maintain and publish inventories of their data holdings

[96] UK Government, *The Government Response to Shakespeare Review of Public Sector Information*, 28 June 2013, pp. 10-11.
[97] These departmental strategies can be found on http://data.gov.uk/open-data-strategies.

(14) Public bodies should publish relevant metadata about their datasets and this should be available through a single online access point; and they should publish supporting descriptions of the format provenance and meaning of the data"[98]

The government recognized that these principles were not immediately attainable for all datasets, but they reflect the aspiration. In its Open Data White Paper the UK government also put an extra focus on making the data as re-usable as possible. Therefore it also endorsed the 5-Star Scheme developed by Tim Berners-Lee to assess the usability of datasets.[99] This implies that the aim is to enrich the datasets into linked open data. A new cross-government Linked Data Working Group has been established.[100]

A system of licenses for PSI has been developed. The default license is the Open Government Licence (OGL), now in its second version, which allows free use and re-use for all purposes, both commercial and non-commercial. This license is compatible with the Creative Commons Attribution License 4.0 and the Open Data Commons Attribution License. When the licensed PSI is adapted and licensed under one of those open content licenses, this satisfies also the conditions of the OGL.[101]
If in specific circumstances a more restricted license is needed, there is also the Non-Commercial Government Licence, allowing free use and re-use but only for non-commercial purposes. When charges have been asked to re-use the PSI, a Charged Licence is available.[102]

Also the legislation has been updated to align it with the Public Data principles. The Freedom of Information Act 2000 was amended to enhance access to data. When a dataset gets released under FOIA by a public authority, it has to provide the data in a re-usable format and specify the license under which the data can be used and re-used. No additional request is necessary for the re-use.[103]
The UK government also plans to amend the Re-use of Public Sector Information Regulations 2005 to implement the recent changes to the PSI directive, but this has not been done yet.

In the Open Data white paper the UK government also stated its awareness of the privacy risks linked to open data, including as a consequence of the mosaic effect or the possibility of re-identification of people by combining datasets. Therefore it announced a stronger attention for the privacy aspects of open data. This by building expertise of the privacy aspects and of statistical disclosure control into the decision making process about the release of datasets. Privacy experts had to be appointed to the Public Sector Transparency Board and to sectorial boards. Further by harnessing the mandatory Privacy Impact Assessments by requiring to conduct disclosure testing when there were concerns about unintended re-identification.[104]

[98] Public Sector Transparency board, *Public Data Principles*, http://data.gov.uk/library/public-data-principles
[99] UK Government, *Open Data White Paper. Unleashing the Potential*, 28 June 2012, p. 24. http://data.gov.uk/library/open-data-white-paper
[100] More information on http://data.gov.uk/linked-data/UKGovLD.
[101] The National Archives, *Open Government Licence v2.0*, http://www.nationalarchives.gov.uk/doc/open-government-licence/version/2/
[102] The National Archives, *UK Government Licensing Framework for public sector information v.4,* 01/09/2013, pp. 14-15. http://www.nationalarchives.gov.uk/documents/information-management/uk-government-licensing-framework.pdf
[103] Protection of Freedoms Act 2012, section 102. http://www.legislation.gov.uk/ukpga/2012/9/section/102/enacted
[104] UK Government, *Open Data White Paper. Unleashing the Potential*, 28 June 2012, pp. 32-33.http://data.gov.uk/library/open-data-white-paper

That privacy is an issue when opening up datasets is clear from the problems the UK government has encountered recently with plans to open up medical records. Not as open data but as part of a restricted data sharing project, which also includes commercial actors like insurance companies. It was contested that the dataset was properly anonymised, as it would contain individual details like "NHS numbers, date of birth, postcode, ethnicity and gender."[105] This public discussion led to a delay of this release with 6 months, which is planned now for October 2014.[106] A similar discussion was raised when plans to sell access to data on taxpayers were announced.[107]

We can conclude that the UK is a trendsetter in all aspects of Big data policy. Where Big data poses a challenge to legal frameworks like IPR and data protection, in depth reviews or guidance is produced which also influences the debates at EU level. It has adapted it IPR framework to diminish the problems for data mining and has raised the issue at EU level. The ICO produces timely guidance on technical developments and problems they raise, like on anonymization.
Also on open data and re-use of PSI the UK proves to play an innovating role. Its policy has past the stage of initial work on enlarging the volume of available data and takes on concerns of semantic interoperability, while its user-driven approach resonates with the ecosystem approach proposed by the European Commission.
Drawback is that its ambitious approach also gets more easily confronted with problems, like public concerns on privacy.

4.2 FRANCE

France has only recently developed a policy for Big Data. In 2013, the "New Industrial France" report placed Big data as one of the 34 main items in France's industrial renovation.[108] The "Commission Innovation 2015" even states that Big data is part of the 7 main challenges France should address.[109]
Several conferences such as the BigData conference[110] or initiatives such the creation of a degree focusing on Big data by TelecomParitech[111] demonstrate a wide interest for Big data.
Big data attracts the interest of large French companies such as Cap Gemini[112] or Airbus[113].
Several companies also invest in infrastructures for Big Data and cloud computing. Bull, for instance collaborates with SFR in Numergy[114] when Orange and Thales aim, through Cloudwatt[115] at building a French cloud. Eventually, some companies such as Apicube[116] or Critéo[117] offer specialized services such as targeted marketing.

[105] The Guardian, "NHS patient data to be made available for sale to drug and insurance firms", 19 January 2014. http://www.theguardian.com/society/2014/jan/19/nhs-patient-data-available-companies-buy
[106] The Guardian, "NHS in England delays sharing of medical records", 18 February 2014. http://www.theguardian.com/society/2014/feb/18/nhs-delays-sharing-medical-records-care-data
[107] The Guardian, "HMRC to sell taxpayers' financial data", 18 April 2014. http://www.theguardian.com/politics/2014/apr/18/hmrc-to-sell-taxpayers-data
[108] Ministère du redressement productif, *La nouvelle France industrielle*, 12 September 2013. http://www.redressement-productif.gouv.fr/files/la-nouvelle-france-industrielle.pdf
[109] Hamel, Marie-Pierre and David Marguerit, *Note d'analyse - Analyse des big data. Quels usages, quels défis?*, 12 November 2013. http://www.strategie.gouv.fr/blog/2013/11/note-analyse-des-big-data/
[110] http://www.bigdataparis.com
[111] https://www.telecom-paristech.fr/formation-continue/masteres-specialises/big-data.html
[112] http://www.fr.capgemini.com/etudes/quest-ce-que-le-big-data
[113] http://www.sopra.com/page.php?menu_mnemo=EVENTS1&lang_code=FR&form_id=19109
[114] https://www.numergy.com/
[115] https://www.cloudwatt.com/en/

However, from a governmental perspective, data processing is not widely used. Furthermore, privacy protection remains a hot topic when it comes to data processing.

As part of the New Industrial France-strategy a Big Data plan was adopted on 2 July 2014.[118] This contains 3 lines of action:

- The development of a Big Data-ecosystem in France. One of the actions under this heading is improving the access of start-ups to data
- Sectorial initiatives on Big Data. This includes both projects in the public and private sector.
- Evaluation of regulation. This includes privacy regulations.

Actions are planned to start in the autumn of 2014.

Open Data

The French government and several French cities and French districts have launched Open Data portals since 2010. The French government releases datasets as open data on the www.data.gouv.fr-portal, which contains now more than 13000 datasets.[119] It organised sectorial consultations on which data to release.
The OKF ranks France at the 16th place.[120] Most open data is released under an open database license: data can be shared alike while citing the author of the database. However, some datasets are not machine readable and easily available.
The Open Data movement is also promoted by research institutes such as the Institut National de Recherche en Informatique et Automatique (INRIA) which organizes several conferences on the topic and opens some of its data.[121]
Eventually, companies such as the Société Nationale des Chemins de Fer, which is partially public, also open some of their data.[122] The company organized several "hackatons".

Personal data, definition and protection[123]

The Law No. 78-17 of 6 January 1978 on data processing, data files and individual liberties, as amended (DP Law) is the main law regulating data in France. This is one of the first legal initiatives on data regulation. There also exists different framework for several sectorial laws for instance in health or finance.

The DP Law applies to personal data. Personal data is any information relating to a natural person who is or can be identified, directly or indirectly, by reference to an identification number or to one or more factors specific to him (Article 2, DP Law).
In order to determine whether a person is identifiable, all means available to the data controller - the person or entity who determines the aim of data collection - should be taken into account.

---

[116] http://apicube.com/category/apicube/
[117] http://www.criteo.com/
[118] Ministère du redressement productif, Point d'étape sur les 34 plans de la Nouvelle France industrielle, pp.8-9.
[119] Valentin Jérémie, Portails Open Data en France, no date. http://timemapper.okfnlabs.org/jeremie34/portails-open-data-en-france; http://opendatafrance.net/
[120] Open Knowledge Foundation, "Open Data Index", https://index.okfn.org/country
[121] INRIA, http://lesfrancaisetlenumerique.inria.fr/
[122] SNCF, http://data.sncf.com/
[123] Sarah Pearce, Edwards Wildman Palmer UK LLP, "Data protection in France: overview", Practical Law, 1 July 2014. http://us.practicallaw.com/6-502-1481

The DP Law, Article 2 states that the law applies to any data processing (such as data collection, consultation use or deletion). The DP Law applies to personal data processing performed by data controllers established in France or using means of processing located in France.
The Commission Nationale Informatique et Liberté (CNIL) is the regulator in terms of data processing. Any personal data processing requires at least a notification with the CNIL (when data is processed on behalf of the state, a ministerial decree or an order issued by the Supreme Administrative Court (Conseil d'État) is necessary). Processing genetic data or data related to conviction and offences, and processing which may exclude persons from the benefice of a right, service or contract, necessitates an authorization from the CNIL.
The DP Law, Article 6, states that all data must be:

- Processed fairly and lawfully.
- Collected for specific, explicit and legitimate purposes, and subsequently processed in accordance with these purposes.
- Collected in an adequate, relevant, and non-excessive way, in view of the purposes for which it is collected.
- Accurate, comprehensive and, when necessary, kept up to date.
- Kept in a form that permits identification of data subjects for no longer than is necessary for the purposes for which the data was collected, or for which it is further processed. Personal data can only be stored beyond this necessary period for processing for historical, statistical or scientific purposes.[124]

Before processing, express consent of data subject is required for any processing of sensitive data or medical research involving the collection of biological sample identifiers. Otherwise, implied consent suffices. When data processing aims at protecting the data subject's life or performing a public service, no consent is necessary.
When collecting data, a data controller must provide information such as its identity, the purpose of the collection, the right of individual to oppose data processing and whether data is transferred outside of the EU. This also applies to electronic communication as data controller must notify data subjects of the use of cookies. The CNIL provides a guideline for cookies use.[125]
The data controller is responsible for safeguarding personal data security. Data controller must record data security breach and the CNIL decides if data subjects should be notified of security breaches. Data controllers are not allowed to transfer data outside of the EU except if they transfer data to a country with an adequate level of data protection policy. Otherwise, they must obtain the CNIL's authorization.
At any time, the data subject retains the right to ask the data controller to rectify complete, update, block or delete their personal data which is inaccurate, incomplete, equivocal, expired or prohibited from being collected, used, disclosed or stored.

We can conclude that the recent action plan of 2 July starts to approach Big Data in a wider context, where before attention for Big Data was limited to a technological innovation perspective. This will include evaluation of the regulatory framework of data access and use. Open data policies are well developed in consultation with end users.

[124] Legifrance, Loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés, version 19 March 2014, art. 6.
[125] www.cnil.fr/en-savoir-plus/fiches-pratiques/fiche/article/ce-que-le-paquet-telecom-change-pour-les-cookies

4.3 GERMANY

Big data policy

The German government published its digital policy plans in the 'Digitale Agenda 2014-2017'.[126] It is a broad agenda, focussing on infrastructural needs, questions concerning internet as a market place and so on. Also attention for a data economy and big data can be found. The agenda points to the growing automated data exchange and the advent of big data, having both attention for capturing the benefits as for dealing with risks for privacy, data security.

One important attention point is how data value chains impact on industrial processes in important industrial sectors like machine building, electrotechnics, the car industry, healthcare technology. This point of attention builds on the earlier initiatives around 'Industrie 4.0'. The term Industry 4.0, points to a fourth industrial revolution. After mechanisation, electrification and informatisation it concerns now the introduction of the Internet-of-Things in industrial processes. Machines and sensors built into production and business processes get networked into Cyber-physical systems, allowing optimization of the production process and turning the whole into Smart Factories. The political attention developed in the Forschungsunion Wirtschaft – Wissenschaft, a consultation process accompanying the government's High-Tech strategy between industry, research institutes and the federal government. In April 2011 a first industry-driven initiative 'Industrie 4.0' was launched[127] followed by a working group Industrie 4.0 to develop policy recommendations[128] These recommendations highlighted the need for developing standards, an industrial broadband network and attention for security-by-design. Several regulatory challenges were brought to attention. First the need to protect corporate data in an environment where data flows between companies. Further liability issues, mostly concerning data security, the protection of personal data and trade restrictions. In April 2013 the platform Industrie 4.0[129] was launched, where these recommendations are followed up and efforts are made to develop standards, although its results are not as fruitful as hoped.[130] The federal government included in November 2011 'Industrie 4.0' as a project in its action plan for its High-Tech strategy[131] and the 'Digitale Agenda 2014-2017' plans to continue this policy attention and further develop it in several sectors in the strategy on “Intelligente Vernetzung” or Smart Networking[132].

The Digital Agenda further points to the energy sector and smart grids as one of the areas where regulation had to establish the necessary framework to allow the deployment of such networked infrstructure.[133]

126 Bundesregierung, Digitale Agenda 2014 – 2017, August 2014, http://www.bundesregierung.de/Content/DE/_Anlagen/2014/08/2014-08-20-digitale-agenda.pdf?__blob=publicationFile&v=6

127 http://www.vdi-nachrichten.com/Technik-Gesellschaft/Industrie-40-Mit-Internet-Dinge-Weg-4-industriellen-Revolution

128 Forschungsunion and acatech, Umsetzungsempfehlungen für das Zukunftsprojekt Industrie 4.0, http://www.bmbf.de/pubRD/Umsetzungsempfehlungen_Industrie4_0.pdf

129 http://www.plattform-i40.de/

130 Karin Zühlke, Plattform Industrie 4.0 vor dem Aus: »Deutschland hat die erste Halbzeit verloren«, 10 February 2015, http://www.elektroniknet.de/elektronikfertigung/strategien-trends/artikel/116855/

131 Bundesministerium für Bildung und Forschung, Zukunftsprojekt Industrie 4.0 , http://www.bmbf.de/de/9072.php

132 Bundesregierung, Digitale Agenda 2014 – 2017, p. 14.

133 Bundesregierung, Digitale Agenda 2014 – 2017, p. 16

Big data further gets specific attention in the research area. Reference is made to the several specific strategies like 'Industrie 4.0'. Also the establishment of 2 competence centres in Berlin and Dresden are announced.[134]
A second important point of attention concerns e-government, including open data. We consider this specifically in the next paragraph on open data.

Re-use of PSI and open data

Germany has signed the G8 Open Data Charter, but is seen as lagging behind and a slow adopter of open data.[135] Openness has not a strong tradition in Germany and a general freedom of information act, establishing procedures for passive transparency at the federal level, was only established in 2005.[136] A year later Germany implemented the PSI-directive.[137] A legal base for open data can be found in the 2013 e-Government law.[138] This law regulated electronic communication with the government, but also includes the obligation for public bodies to use machine readable formats when publishing data and to include metadata. Machine readable means that the data can be read and processed in an automated manner by software.[139]

In its National Action Plan of November 2014[140] the federal government presents how it wants to implement this Charter. The federal government adheres to the principle of 'Open Data by Default', but states it is a long term goal. Since producing metadata is costly, the federal government states it can not afford to implement the principles for all data assets from the past. Therefore it decided to focus on data newly collected and published.
The previous federal government in the 17th legislative term (2010-2014) has already amended the 2008 Act on Access to Geodata (GeoZG), which established a right of access to geographic information held by the federal government, and issued in 2013 an Ordinance Setting the Terms of Use for the Provision of Federal Geodata (GeoNutzV), introducing free commercial and non-commercial usage rights to this data. A similar ordinance will be made applying to as many data held by the government as possible.
The open data portal of the federal government, GovData or www.govdata.de, started in 2013 and is coordinated by the Ministry of Interior. By the end of 2014 open data contact persons were designated at all ministries to coordinate the provision of datasets from their ministry. For 2015 a list is made of datasets to improve and extra datasets to publish, while all federal authorities commit to make at least 2 datasets available on GovData. GovData is also promoted as the central open data portal for state governments and local authorities. Still, these engagements are criticised for being weak and missing important datasets.[141] The old

134 Bundesministerium für Bildung und Forschung, Big Data - Management und Analyse großer Datenmengen, 16 March 2015, http://www.bmbf.de/de/23429.php; Bundesregierung, Digitale Agenda 2014 – 2017, p. 28.
135 Castro, Daniel and Travis Korte, Open Data in the G8: A Review of Progress on the Open Data Charter, Center for Data Innovation, March 2015, p. 4-5 and 15-17.
136 Germany, Gesetz zur Regelung des Zugangs zu Informationen des Bundes (Informationsfreiheitsgesetz – IFG), 5 September 2005, BGBl. I, 2005, p. 2722.
137 Germany, Gesetz über die Weiterverwendung von Informationen öffentlicher Stellen (lnformationsweitervenwendungsgesetz – IWG), 13 December 2006, BGBl. I, 2006, p. 2913.
138 Gesetz zur Förderung der elektronischen Verwaltung sowie zur Änderung weiterer Vorschriften (E-Government-Gesetz), 25 July 2013, BGBl. I, 2013, p. 2749.
139 Ibid., article 12.
140 Bundesministerium des Innern, Nationaler Aktionsplan der Bundesregierung zur Umsetzung der Open-Data-Charta der G8, 17 September 2014.
141 Castro, Daniel and Travis Korte, Open Data in the G8: A Review of Progress on the Open Data Charter, Center for Data Innovation, March 2015, p. 17.

license system was also criticised, but has been renewed and simplified. 2 licences are now used, one with obliged referencing to the author and one without any restrictions.[142]
Federal, state and local governments also work together on an agenda to standardize the metadata structure for open data. These commitments will get implemented in close dialogue with civil society and other stakeholders, e.g. to identify demand.

IPR

The Digital Agenda 2014-2017 announces it will adapt the IPR framework to the growing digitalisation under consideration of all interests, without further clarification of the options to follow. This adaptation is seen in an European context, as the intention is stated to engage strongly in the EU review of copyright.[143]
In Germany there is a broad movement of scientists and researchers, the Aktionsbündnis „Urheberrecht für Bildung und Wissenschaft" or Coalition for Action "Copyright for Education and Research", to limit copyright and to include exceptions for research in the copyright framework. Their attention is focussed on open access, but concerns about limitations for data mining and text mining due to copyright are raised as well. Licensing solutions are seen as not adequate.[144] Their demands on open access have already led to political discussion.[145] Big data-relevant issues like data mining have not shown up yet in the political debate.

Protection of personal data

Germany has a strong attention for the protection of personal data and its strict application of data protection. The German DPAs have a strong attention for big data and made a plea for a stricter regulation of (commercial) profiling.[146] They have been cautious about the use of proper anonymisation. For instance, the use of anonymised mobile phone data for movement profiles was limited by the federal DPA and could only be done in an aggregated manner, as the anonymized profiles gave too much possibilities for re-identification.[147]

Meanwhile the German government also has a lot of attention for the potential of big data as noticed in the discussion above. The tensions between both have therefore been subject of policy discussions in order to fine-tune the German government's positions on the GDPR. The minister of internal affairs De Maizière stated that the protection of data is not the objective as such, but rather the protection of privacy has to be the focus of the debate. Suggestions are made for additional data protection principles like risk minimization and discrimination prohibitions. De Maizière announced specific German proposals on pseudonymization and profiling.[148]

142 Govdata, Datenlizenz Deutschland, https://www.govdata.de/lizenzen
143 Bundesregierung, Digitale Agenda 2014 – 2017, p.15.
144 Aktionsbündnis „Urheberrecht für Bildung und Wissenschaft", Forderungen an die Bundesregierung, 11 Oktober 2013, http://www.urheberrechtsbuendnis.de/ge.html.de#forderungen
145 Bundesrat, Beschluss, Gesetz zur Änderung des Urheberrechtsgesetzes, 737/12(B), 14 December 2012, http://www.bundesrat.de/bv.html?id=0737-12
146 Decision of the Conference of the federal and state DPA's of 13 March 2013, in: BfDI, Tätigkeitsbericht 2013-2014, Annex 3 and 4, 246-250.
147 BfDI, Tätigkeitsbericht 2013-2014, 153.
148 Bundesministerium des Innern, Expertenrunde "Big Data – eine Herausforderung für den Datenschutz", 22 August 2014, http://www.bmi.bund.de/SharedDocs/Kurzmeldungen/DE/2014/08/expertenrunde-big-data.html

Conclusions

Germany has an extensive digital policy, with a lot of attention for big data. The government, encouraged by the industry, shows a lot of interest in the industrial applications of big data and IoT. It puts a lot of effort in developing standards and legal frameworks to allow its industry to capture the benefits of big data. The use of PSI and open data have received less attention, but Germany has started to catch up.
This positive policy towards big data conflicts with a strong attention for the protection of personal data and a strict enforcement by the German DPAs. This results in a lot of policy attention on how to make big data and data protection compatible and an active engagement in the legislative procedure of the GDPR.
Reform of IPR also has received attention, but the German government seems still to search for a balance between all interests at stake. The political discussion is mostly driven around open access and the interest of publishers versus scientists and researchers, but the public discussion includes the issue of data mining and text mining. The German government links internal reforms to the review of the European copyright policy.

4.4 SWEDEN

No specific big data policy

Sweden has a developed ICT policy[149], but it contains no specific big data policy. The attention remains focussed on the internet as a market, where private and public services meet human customers, but not so much on the internet as the medium for a data economy where data flows between a whole range of actors, human and machines. Seen its geography with a sparse population in most of the country a lot of attention goes to infrastructure, and more specifically to assuring access to broadband. Further attention goes to enabling e-commerce and providing services, including e-health, e-learning and a strong focus is given to e-government. An attention point within this e-government focus concerns the re-use of PSI and open data. Big data -related applications show up in the attention for smart grids and intelligent transport systems (ITS).[150] The ICT policy also contained attention for copyright, which we discuss further.

PSI and open data

Sweden has a long tradition on governmental transparency with its first access to documents act in 1766. But there was much more reluctance to share PSI with the private sector. The PSI-directive was only implemented in 2010 in the Act on the Re-use of Documents from Public Administration[151], 5 years after the deadline in 2005 and only after threats of an infringement procedure by the EU. Even this law did not fulfil the commercialisation aims of the PSI-directive and the Swedish government has promised to review the law in 2015.[152]

149 Swedish government, ICT for Everyone - A Digital Agenda for Sweden, November 2011, http://www.government.se/reports/2011/12/ict-for-everyone---a-digital-agenda-for-sweden/
150 Swedish government, ICT for Everyone - A Digital Agenda for Sweden, p. 51-52.
151 https://ec.europa.eu/digital-agenda/en/news/implementation-psi-directive-sweden
152 Kallberg, Jan, and Erik Lakomaa. "Institutional maximization and path dependency–the delay of implementation of the EU public sector information directive in Sweden.", 2015, p. 4, http://works.bepress.com/jan_kallberg/30

The 2010 Act was interpreted by the administration to be about documents on paper, and not concerning digital information or data.[153] The Swedish administrations were often dependent on the commercialization of their PSI themselves, creating strong resistance to change according to the PSI-directive.
In its 2011 ICT policy the Swedish government included the objective to promote new services by actors other than government agencies. It foresaw an evaluation and monitoring of the compliance by authorities with the Act on PSI.[154] The open data policy is further developed in the e-government strategy of 2012[155].
Governmental transparency remained an important motivation for open data, as can be seen in Openaid.se platform, providing information on Swedish aid disbursements in an open format and launched in April 2011. A general open data portal followed in December 2012 on http://oppnadata.se, as part of the new e-government strategy. Licensing is sometimes unclear, as some key datasets were not published with an open license[156] and this is in some case still the case. No specific government license exists, but several existing open licenses are used.

IPR

The Swedish policy debate on copyright has been largely driven by the Pirate Bay-case. This resulted in a focus on digital content in the discussion. Big data-related issues like data mining were not raised.
In April 2008 a special commissioner was tasked with reviewing the copyright law, who presented in 2010 an interim report on 'Contractual copyright' and in 2011 a final report 'A new Copyright Act'. Based on these reports the Swedish Copyright Act, which also includes database protection, was revised, as foreseen in the 2011 ICT policy. The report and the 2013 law changes included a new general collective license for situations where a large amount of copyrighted material is needed by a user, without being able to determine in advance which material he will use, and where it is practically not possible to obtain individual licenses from rights holders.[157] Again, this was a result from the discussion on Pirate Bay[158], who defended themselves as being merely a search engine referencing sources and in the practical impossibility of obtaining licences for the wide range of sources. This license is now available for several specific situations, like internet music providers offering a range of material for streaming. The law included also a general license to cover cases not mentioned. As similar situations can exist for big data applications, this extended collective license can provide a solution. These licenses can be obtained from a national collecting society.

Protection of personal data

Sweden was also a forerunner in the protection of personal data with its 1973 Data Act. This act got replaced when Sweden implemented the directive 95/46 in the Personal Data Act of 1998. General principle is that all automated processing of personal data has to be notified for a prior check to the Swedish Data Inspection Board or Datainspektionen, except when a data processor has appointed a data representative and given notice about this representative.

153 Ibid., p. 10.
154 Swedish government, ICT for Everyone - A Digital Agenda for Sweden, p. 23.
155 http://www.regeringen.se/informationsmaterial/2012/12/n2012.37/
156 Alina Östling, Independent Reporting Mechanism Sweden: Progress Report 2012–13, p. 33.
157 Erik Ullberg and Michael Plogell, Amendments to the Swedish Copyright Act, IRIS Merlin, http://merlin.obs.coe.int/iris/2013/7/article25.en.html
158 Case law overview on http://cyberlaw.stanford.edu/page/wilmap-sweden

Privacy-sensitive processing must always be notified, even when a data representative is appointed.[159] The 2011 ICT policy foresees a strengthening of the Data Inspection Board in order to deal with the rising challenges of the digitalisation.
The Data Inspection Board made clear that data protection has to be ensured when using cloud services. It prohibited municipalities to use Google's cloud services to store personal data. It considered that Google had too much discretion over the data and that authorities could not adequately ensure the data protection rights of its subjects.[160]

Conclusions

Sweden has no big data policy and attention for big data seems limited, although the DPA is attentive to the large big data companies. After initial difficulties an open data policy has started to developed and the hurdles towards private re-use of PSI seems to be lifted. In the discussion on IPR data and data mining did not receive attention, but the extended collective license for general purposes can provide an answer for IPR-related problems for data mining and big data processing.

4.5 ITALY

Italy is pioneering the issue of data policies, in particular as regards the governance of the Internet, with the activity of the Study Committee on Internet Rights and Duties[161], promoted in July 2014 by the President of the Chamber of Deputies.[162]

The initiative, which follows some recent developments (e.g. the European Court of Justice ruling introducing the so-called "right to be forgotten"[163]; the discussions in the UN context and other countries[164]), resulted in the Draft Declaration Of Internet Rights[165], released in October 2014, after a phase of public consultation.

The document is inspired by similar declarations, such as the Marco Civil in Brazil[166], the French report on Rights and Liberties in the Digital Age[167], the work of the German

159 Datainspektionen, Regulation amending Data Inspection Board Regulation (DIFS 1998:2) with regard to the obligation to notify the processing of personal data to the Data Inspection Board, DIFS 2001:1, 3 October 2001, http://www.datainspektionen.se/Documents/datainspektionen-foreskrifter-2001-1-english.pdf

160 Simon Davies, Sweden's data protection Authority bans Google cloud services over privacy concerns, The Privacy Surgeon, http://www.privacysurgeon.org/blog/incision/swedens-data-protection-authority-bans-google-apps/

161 http://www.camera.it/leg17/1174

162 http://www.zdnet.com/article/regulating-the-web-does-the-internet-need-its-own-bill-of-rights/

163 http://curia.europa.eu/jcms/upload/docs/application/pdf/2014-05/cp140070en.pdf

164 http://tg24.sky.it/tg24/mondo/2014/10/23/internet_bill_of_rights_italia_mondo_brasile_leggi_web.html

165 http://www.camera.it/application/xmanager/projects/leg17/attachments/upload_file/upload_files/000/000/189/dichiarazione_dei_diritti_internet_inglese.pdf

166 https://www.publicknowledge.org/documents/marco-civil-english-version

167 http://www2.assemblee-nationale.fr/14/commissions/numerique

Bundestag's committee on the Digital Agenda[168]. It particularly stresses the protection of the individual from widespread monitoring, in consideration of the growing collection of data by Internet companies, which are largely free to set their own policies in this respect.[169]

Concerning specific Big Data-related policies, the document tries to address the so-called "Consent Dilemma"[170] whereby, given the complexities of privacy policies and the advances in the Big Data analysis, it is impossible for users to properly evaluate the costs and benefits of disclosing their personal data. It maintains, therefore, that "consent shall be revocable" (article 4) and that it "does not constitute a legal basis for the processing of data when there is a significant imbalance of power between the data subject and the data processor". However, the document does not specify who should determine when the supposed imbalance has become substantial.

In general, the issues of Big Data are often discussed along data mining issues[171], mainly in the context of the EU debate on the General Data Protection Regulation ("Personal data protection: processing and free movement of data")[172]. However, the Italian uptake of Big Data can be considered still at an early stage. According to some[173], there would be no organization or application in Italy that make use of actual Big Data (i.e. in the order of the Zettabytes).

Among the initiatives relevant to Big Data in Italy, we may include: the Italian Grid Infrastructure[174], mainly aiming at research purposes; NextData[175], a national system for the retrieval, storage, access and diffusion of environmental and climate data from mountain and marine areas; SoBigData (European Laboratory on Big Data Analytics and Social Mining)[176], a collaborative project promoted by the National Research Council of Italy and the University of Pisa; the yearly (now at its second edition) Telecom Italia TIM Big Data Challenge[177], an evidence of the growing attention of the industrial sector to Big Data.

Concerning data protection and privacy, the Italian Data Protection Authority (DPA) has intervened several times on topics such as the profiling of physical persons behavior[178]. According to the Italian norm, profiling is allowed, also in the absence of previous consent, as far as the responsible party ensures to protect the anonymity of the interested subject, and

[168] http://www.bundestag.de/htdocs_e/bundestag/committees/a23
[169] http://www.zdnet.com/article/first-draft-of-internet-bill-of-rights-revealed-in-italy/
[170] http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2171018
[171] Stefano Neri. La disciplina del data mining alla luce della proposta di regolamento comunitario in materia di trattamento di dati personali: criticità, limiti e prospettive de jure condendo. CIRSFID. Filo Diritto. April 2015. http://www.filodiritto.com/documenti/2015/stefano-neri_la-disciplina-del-data-mining.pdf
[172] http://www.europarl.europa.eu/oeil/popups/ficheprocedure.do?reference=2012/0011(COD)&l=EN
[173] http://blog.debiase.com/2013/05/22/che-cosa-pensereste-se-vi-dicessero-che-in-italia-i-big-data-non-esistono/
[174] http://www.italiangrid.it/
[175] http://www.nextdataproject.it/?q=en
[176] http://www.sobigdata.eu/
[177] http://www.telecomitalia.com/tit/it/bigdatachallenge.html
[178] Among others, about interactive television, on 3 February 2005.

provides a clear definition of the use and purpose of the data collected. In line with the EU proposal, the internal legal debate has often viewed data management as potentially harmful, and hence under the scope of article 2050 of the Civil Law ("Responsibility for conducting dangerous activity").

In a recent interview, the DPA has underlined how "billions of information are collected all over the world, analyzed through a network of servers distributed everywhere, whose control is substantially in the hands of an oligopoly with an enormous power."[179] In particular, there is concern for the dominant position of big US companies, defined as "lords of data", as "the main IT resources concentrate in the hands of a few actors and are even physically aggregated in enormous data centers"[180], giving them the exclusive opportunity to fully exploit the potential of data analysis.

In the same interview, the DPA elaborates on the most relevant issues in the international debate, from Big Data to the notorious Prism[181] case, maintaining that "big American providers have a level of personal data protection absolutely lower than the European one, what creates a competitive disadvantage for European companies." [182]

For Italian jurists, this concern echoes the considerations of the Constitution, Article 41, on the concept of private initiative (and the social function of private property in general), which is free, but can not proceed against social utility, or so as to procure damage to human safety, freedom, and dignity.

Actually, the President of the Study Committee on Internet Rights and Duties[183], Stefano Rodotà, advocates the introduction of a principle of maximum tolerance limiting the use of personal data with respect to the purpose of their treatment. In analogy to the "Habeas Corpus", protecting the physical integrity, such "Habeas Data" would establish that an individual could not decide to give up his/her own personal data. In the future, personal data may be considered a constitutional right, such as the right to health (constitutionalization of the right to privacy).

An important related issue is the controversial "right to be forgotten"[184] ("diritto all'oblio"), allowing a European citizen to be excluded from Internet search results. This has already

[179] Antonello Soro (DPA President), Huffington Post interview, 29 August 2013. http://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/2606830
[180] Alessandro Mantelero, "Big data: i rischi della concentrazione del potere informatico digitale e gli strumenti di controllo", Diritto dell'informatica, pp. 125-134.
[181] An international scandal involving the US National Security Agency, and, indirectly, the US government itself, following the revelations by Edward Snowden, a former data analyst at the Central Intelligence Agency, and causing the US President to admit that the Agency violated several procedures in the acquisition of data and metadata about American and foreign citizens.
[182] Antonello Soro. Op. Cit., 2013.
[183] http://www.camera.it/leg17/1174
[184] http://www.zdnet.com/article/do-we-really-have-a-right-to-be-forgotten/

found application in the internal jurisdiction[185] and is fully supported in the Draft Declaration Of Internet Right[186], except when it conflicts with the right of the public to be informed.

According to article 11, users of digital platforms (Facebook, Google and the like), should have better control over their data, including "the right to terminate the relationship, to receive a copy of the data concerning them in interoperable form and to have the data concerning them removed from the platform".

Concerning intellectual property and data openness, the Article 68 of the Digital Public Administration Act establishes the core rules for all aspects related to openness in the Italian public sector: free and open source software (par. 1 and 2), open formats and open data (par. 3). A change introduced to the wording of this article in 2012 resulted in an unprecedented opening of government-held or produced data.

According to the Italian norms, open data are defined as:

1) Available under the terms of a license permitting their use by anyone, even for commercial purposes, in disaggregated format;
2) Accessible through the information and communication technologies, including public and private telecommunication networks, in open formats; are suitable for automatic processing by computer programs and equipped with relative metadata;
3) Available for free through the information and communication technologies, including public and private computer networks, or are available to the marginal costs incurred for their reproduction and dissemination.

The Agency for a Digital Italy[187] shall establish exceptional cases, identified according to objective, transparent and verifiable conditions, in which data are made available at higher rates to marginal costs. In any case, the Agency will follow the guidance provided by Directive 2003/98/EC of the European Parliament and of the Council of 17 November 2003 on the re-use of public sector information, implemented by legislative Decree 24 January 2006, n. 36.

4.6 BELGIUM

In general Belgium is a follower in policies concerning data and information. The legal framework and policies in Belgium concerning access and use to data are mostly driven by European or international initiatives. The legal framework concerning IPR, access and re-use of PSI, INSPIRE, and data protection are implementations of EU law and to a smaller extent international treaties like the Aarhus Convention. These implementations take place without the development of strong policies or substantial political debate.
Exception to that were the inclusion of a right to access of PSI into the Constitution in 1993 and the adoption of a law implementing that right in 1994 at the federal level, as well as the early data protection law in 1992 (which later has been profoundly changed by the

[185] See sentence n. 5525 (diritto all'oblio e diritto di cronaca), Corte di Cassazione, Section III, 5 April 2012.
[186] http://www.camera.it/application/xmanager/projects/leg17/attachments/upload_file/upload_files/000/000/189/dichiarazione_dei_diritti_internet_inglese.pdf
[187] http://www.agid.gov.it/

implementation of EU directive 95/46). But the introduction of these laws was also relatively late compared to other European countries.
Recently the international positive mood concerning open data has been picked up in Belgium. Driving forces are the regional governments, which develop policies on open data, with the federal government lagging behind.

IPR

The Belgian law concerning copyright and database protection in Belgium does not differ much from that in other EU countries, as it is mainly regulated by international treaties and European directives. As such the difficulties signalled earlier with the European framework can also be found in the Belgian law. The Belgian law on author rights contains a wide range of the exceptions, provided as options in the WIPO and TRIPS treaties, but some are included in a narrow formulation (like the right to quote exists only for specific use cases). Database protection by copyright is included in the Belgian law on author rights and neighbouring rights of 1994[188], while the sui generis right has been implemented through a separate law[189].

Notwithstanding the European framework, Belgian courts seem to apply the copyright law more restrictively than the ECJ, as can be concluded from the Copiepresse vs Google-decision of 5 May 2011 by the Court of Appeal in Brussels. Copiepresse, an organisation representing the French language newspapers, had sued Google for reproduction and communicating to the public of a significant part of news articles through Google News and of the whole article by giving the public access to the copy stored in cache. The court agreed with the position of Copiepresse and considered none of the exceptions in the law on authors rights applicable on Google's practice. The court case concerned mainly the communication by Google of search results on Google News and the public access it gave to its cache, and not the caching itself. But the argumentation of the court makes clear that it considers this caching itself already a violation of copyright. This interpretation of Belgian copyright law is problematic for text and data mining in general, as it needs a local copy in order to process it with certain data mining algorithms. The Google cache is such a copy, which gets processed to build the indexing on which the search function is based. The court stated that there was reproduction from the moment of storage of a copy. It refused the application of the exception for temporary and transient acts of reproduction as part of a technical process, due to the fact that the cache remained present for 30 days. The court stated that Google could not explain why this was necessary for the technical process. This still gives some space for data mining, but data mining and big data processing using copyright-protected material clearly has to proceed in a narrow grey zone.
Currently no political discussion or policy initiatives is taking place concerning the impact of copyright law on data mining and big data processing.

Another limitation is that the Belgian copyright law contains some obstacles to give wide authorizations through licenses or to waive rights as in the CC0 license. Contracts with the original author transferring rights need to be written and are interpreted restrictively, which prevents licensing for uses which are not known yet with new technologies. Also a global transfer of moral rights is not possible, only contracts concerning the application of these rights are possible (like publishing anonymously or under the name of an institution). When

[188] Law concerning author rights and neighbouring rights, 30 June 1994 (BS 27.07.1994)
[189] Law implementing in Belgian law the European directive of 11 March 1996 concerning the legal protection of databases, 31 August 1998 (BS 14.11.1998).

the license is not given by the original author, like with governmental re-use licenses, these constraints can be avoided.[190]

Re-use of PSI and open data

More activity at governmental level is taking place concerning re-use of PSI and open data. The EU PSI-directive has been implemented by both federal and regional governments in very similar texts.[191] The revised PSI-directive has not been implemented yet.
Recently the regional governments took the lead in developing open data policies.

Flanders

In Flanders an open data-policy first got shape around geographical information, through the development of a geographical information infrastructure around which access and re-use policies had to be developed. In 2009 the legal framework concerning geo-information was updated with the implementation of the INSPIRE directive.[192] After consultations with stakeholders a more general open data policy was developed with a concept note[193] in 2011 and a plan of action in 2013. 3 objectives were listed: transparency of government for the public, augmenting efficiency of government and stimulating innovation through new uses of PSI. The concept note put forward 6 leading principles:
1. Open data becomes the default, closed data needs an explicit legitimation.
2. Re-use of open data is allowed, also for commercial purposes. Standardized licenses would be developed.
3. Open data uses open standards.
4. Open data is released from the original sources. Such original source is managed by the public authority competent for its generation. It is best able to ensure data quality, to provide appropriate metadata and to take measures to avoid the release of confidential or privacy-sensitive data.
5. Releasing open data demands an integrated approach with other levels of government, including local authorities.
6. A central register is developed with business information about the Flemish government, which later can be released as open data.[194]
One important limitation is that the Flemish government has decided to exclude all personal information from the open data policy.[195]
A beta-version of Flemish open data platform[196] was launched on 14 June 2013, while open data from the Flemish government is also released on the federal open data platform.

A set of standard licenses has been developed for use to release PSI:[197]
1) the use of the Creative Commons Zero (CC0) license, through which intellectual property rights are waived.

[190] Flemish government, *Juridische nota bij de modellicenties Open Data*, p. 6-7.
[191] An overview can be found on https://ec.europa.eu/digital-agenda/en/news/implementation-psi-directive-belgium
[192] Flemish government, Decree on Geographical Data Infrastructure, 20 February 2009.
[193] Flemish government, *Concept note "Een concept van beleid met betrekking tot open data"*, 23 September 2011, http://www.bestuurszaken.be/conceptnota
[194] This is made available at http://www.bestuurszaken.be/overzicht-beschikbare-data
[195] Flemish government, *Open Data Handleiding 2.0*, p. 10.
[196] Flemish government, "Open Data Platform", http://ckan-001.corve.openminds.be/
[197] Flemish government, *Open Data Licence Models v1.1*, http://www.opendataforum.info/files/Modellicenties_ENG.pdf

2) Free Open Data Licence: Under this licence the authority holding intellectual rights authorizes the re-use of the data for both commercial and non-commercial purposes, free of charge and under minimal restrictions. This license is designed to be compatible with other open licences conditional on attribution, like CC-BY 3.0.
3) Open Data Licence at a Fair Cost: Similar to 2 re-use is authorizes for all purposes, but now for a fair compensation.
When the public authority wants to differentiate its charging policy between commercial and non-commercial uses it can combine the following two licenses.
4) Free Open Data Licence for Non-Commercial Re-Use
5) Open Data Licence at a Fair Cost for Commercial Re-Use
These 2 licenses have always to be used together. The first authorizes the re-use for non-commercial purposes without a charge, the second authorizes commercial re-use for a fair compensation.

Walloon Region

The Walloon regional government addressed open data in its Master Plan ICT in 2011, which is part of its Creative Wallonia strategy to stimulate innovation. Objectives focussed on starting up an open data platform. This is now online at http://opendata.awt.be, but still in a starting up phase. Licenses used are indicated with each dataset.

Concerning geographical information an open data practice is more developed and is made available through the portal website Géoportal de la Wallonie. Licensing practices are at the moment still varying and sometimes contradictory. An open data license[198] has been developed based on the French license by Etalab[199], and which is interoperable with other open content licenses like CC-BY 2.0. Although at the moment the page offering open data[200] which links to this license, also offers it directly under the more restrictive CC BY-SA 2.0 BE license (which only allows to re-distribute under the same license). Some geographical data is licensed for free but under more restrictive terms and needs approval as a contract.[201]

On 8 may 2014 a 'Plan Stratégique Géomatique pour la Wallonie' was approved. It contains a much more coherent and detailed approach on how to deal with geographical data. It has 4 main axes:

- promote the use of the geographical data
- creating a common framework to produce geographical data to ensure that the data is interoperable, well documented and of high quality and can be re-used
- organise the sharing of geographical data and setting up an infrastructure for such sharing
- develop a governance structure for this geographical data that is both coherent and participatory

The use of an open data license is foreseen in this Strategic Plan.[202]

[198] SPW-DGO4 Open Data, "Licence ouverte - Open licence", http://dgo4.spw.wallonie.be/dgatlp/dgatlp/Pages/DGATLP/Dwnld/OpenData/licence_ouverte_open_licence.pdf
[199] Etalab, "Etalab publie la « Licence Ouverte / Open Licence »", 17 October 2011, http://www.etalab.gouv.fr/article-etalab-publie-la-licence-ouverte-open-licence-86708897.html
[200] Direction générale opérationnelle - Aménagement du territoire, Logement, Patrimoine et Energie, "Open Data", http://dgo4.spw.wallonie.be/DGATLP/DGATLP/Pages/DGATLP/PagesDG/OpenData.asp
[201] Géoportail de la Wallonie, "Provision of data", http://geoportail.wallonie.be/en/home/ressources/mise-a-disposition-de-donnees.html
[202] SPW, *Plan stratégique géomatique pour la Wallonie*, 8 May 2014, p. 51.

It is planned that this Strategic Plan will be followed with an operational plan early 2015.

Brussels Capital Region

In Brussels the open data policy is also driven by geographical information. The Brussels region provides high-quality geographical information through UrbIS at http://www.bric.irisnet.be/en/our-solutions/urbis-solutions/urbis-data, which are available as open data since 1 April 2013. As part of its implementation of the Inspire directive a new central portal for geographical information, GeoBru, is established on http://geonode.geobru.irisnet.be.
The Brussels region established its smart city strategy in the White Paper Smart.brussels 2014-2019. Open data is one of the key goals and benchmarks in this strategy. The Brussels region wants to enlarge the amount of Brussels institutions providing open data and the amount of datasets. It plans to promote that data gets released in open formats and under a standard open license, as well as a central data portal for access. The use of this data is promoted through other actions under this strategy, like the promotion of more mobile apps using this open data.
The Brussels Regional Informatics Centre (BRIC) gets a key role as digital service integrator setting up all digital data communication between public institutions. In this role it also has to set up more adequate data management, ensuring access policies, data quality and semantic interoperability, data security. Setting up the open data platforms is part of this role. Now it already manages UrbIS and GeoBru.
An open license has been developed by BRIC and Brussels Mobility[203], based on the French license by Etalab[204], and which is interoperable with other open content licenses like CC-BY 2.0.

Federal government

The Belgian government has been lagging behind the regions concerning open data. The INSPIRE-directive was implemented in December 2011, where the regions have implemented it in 2009 and 2010. A data portal has been set up on http://data.gov.be/, with datasets provided by federal public administrations as well as regional ones. The ministry of economy has released a dataset containing the company register. However, the more detailed version is only available on an expensive commercial license.
Aside of the implementation of INSPIRE, no clear policy on interoperability of the data or cooperation structure with the regions exist yet.

We can conclude that the awareness of Big Data on policy level and in the political debate is still limited. The legal framework contains all tensions raised earlier when reviewing the EU framework and its application by courts runs behind the EUCJ case law. Changes to the legal framework follow narrowly the EU agenda.
Adoption of open data policies varies. The regions gather experience through the INSPIRE-implementation, including in the wider aspects of interoperability. Flanders is also working on these aspects in its general open data policy, while the other governmental levels are still in the start-up phase of their open data policies.

[203] CIRB, “Licence ouverte - Open licence”, http://www.cirb.irisnet.be/fr/nos-solutions/urbis-solutions/licence-urbis-open-data
[204] Etalab, “Etalab publie la « Licence Ouverte / Open Licence »”, 17 October 2011, http://www.etalab.gouv.fr/article-etalab-publie-la-licence-ouverte-open-licence-86708897.html

4.7 CONCLUSION

Big data policies vary a lot across Member States. The UK is a trendsetter in all aspects of Big data policy. Although it has no distinct Big Data policy document, we find that all elements of the EU policy discussion are raised or were influenced by the UK discussion. This is surely valid for the IPR and privacy discussions. Also Germany is a trendsetter. It has a lot of attention to the industrial use of big data and is a trendsetter on IoT. Also big data-related issues in data protection are strongly considered. IPR receives limited attention, while also the open data policy is lagging behind.

In Italy and France we did not find such in-depth policy, although these countries are have a lot of attention to big data in specific areas. In France Big Data figures as part of an industrial innovation strategy. As a result it has recently developed a Big Data plan, which includes an evaluation of its regulatory framework concerning big data. Italy has a lot of attention to big data in the context of data protection and on the level of research infrastructure.
The smaller countries Belgium and Sweden do lack attention to big data in their policies.

Open data policies shows similar variation. Open data policy in the UK has passed its start-up phase and attention widens to include data quality. The open data policy is embedded in a clear vision on how it has to support the development of a data economy. This resulted in the user-driven approach to create the data ecosystem. France also has a well-developed open data policy with user consultations.
Such clear vision is in general lacking in the open data policies in the other countries. Interesting is the pioneering role the INSPIRE-framework has in developing the experience to develop such vision.
Legal interoperability has been addressed in all countries except Sweden, as the governmental licenses are all made compatible with the Creative Commons license.

5 DATA POLICIES IN THE US

5.1 INTRODUCTION

The role of the US in the development of the digital economy and the dominant position of US companies makes its data policies and legal frameworks impacting on data access and use very relevant, even when they are not applying in the EU. In this chapter we look again at how private and public law has an impact on the space for big data operations. We focus on the private law framework of intellectual property rights and licensing, public law limitations provided by privacy laws and the policy concerning PSI.
In the next chapter we will go deeper on how private actors use the private law framework to shape themselves the possibilities and barriers to access and use data. Here we focus on the legal frameworks and policy discussions concerning them.

5.2 IPR AND CONTRACT LAW

The basic framework for copyright law in the US is provided by the Copyright Act of 1976, contained in Tıtle 17 of the United States Code, and which has been regularly amended since. The US does not have a separate act providing for database protection. Database protection has been developed by the courts based on the regular copyright framework.
This Act gives copyright protection to "original works of authorship"[205]. This definition contains clearly the originality requirement. The Act further makes clear the difference between the protected expression and factual matters outside that protection: "In no case does copyright protection for an original work of authorship extend to any idea, procedure, process, system, method of operation, concept, principle, or discovery, regardless of the form in which it is described, explained, illustrated, or embodied in such work." Basic idea is that facts are discovered and not the result of creativity. In other words, data as such is only protected if it can fulfill the criteria to be an original work of authorship.
Database protection is derived from the copyright protection of a compilation. A compilation is defined as "a work formed by the collection and assembling of preexisting materials or of data that are selected, coordinated, or arranged in such a way that the resulting work as a whole constitutes an original work of authorship".[206] Also here a certain creativity is required, this time for the selection, coordination or arrangement.
This creativity and the ensuing copyright is distinct from any copyright or lack thereof linked to the data. Copyright in a compilation, or more generally in any collective or derivative work, is only given to the actual contribution of the author, and does not extend to preexisting material, factual data lacking protection, etc. In other words, the scope of protection is commensurate with the creative contribution.

A lot of legal battles have been fought on what can be protected by copyright and what not, including cases concerning several sorts of data and compilations of data.[207] Main legal precedent is the Feist-case, in which the Supreme Court made clear that effort or investment is as such not protected by copyright. It took distance from court decisions which also granted protection to 'sweat of the brow' or 'industrious collection', through which courts had earlier developed a protection for factual collections. Instead it made clear that originality was an

[205] US, *Copyright Act*, 17 U.S.C., §102(a).
[206] US, *Copyright Act*, 17 U.S.C., §101.
[207] An overview of case law can be found in Leslie C. Ruiter and Gerald van Belle, "Data Extraction: Beyond the Sweat of the Brow", 2014. http://www.stokeslaw.com/uploads/pdf/data_and_the_law-gerald_van_belle_and_leslie_ruiter.pdf

essential requirement and that facts or factual compilations could therefore not receive copyright protection. The Court grounded that requirement on the objectives of copyright protection listed in the Constitution "to promote the Progress of Science and useful Arts". Copyright also needs to allow others to build upon the ideas and information contained in a work, which is the rationale for only granting protection to the expression but not to facts.
The case concerned the white pages of a telephone directory, consisting of an alphabetically ordered lists of names with their town and telephone number. The Court considered that such lists of facts lacked any originality and was not protected by copyright.

This means that the US law does not contain anything similar to the sui generis-protection of databases in the EU. Factual data in databases or other works are available for reproduction or extraction, even when this extraction is substantial.

Important is the consequence of the refusal to grant protection to 'sweat of the brow' or investment in a database for databases originating from the US in the EU. As there is no reciprocity for the sui generis protection in the US, the European sui generis protection does not extend to US databases. Databases on the internet from US companies can in the EU only claim the database protection based on copyright.

Result is that US companies often attempt to deviate from the default copyright regime through restrictive licensing practices. But the reception of shrinkwrap and clickwrap licenses by US courts has been quite divided. The validity of the consent to the license or to specific terms of the license has been doubted, resulting in a lack of enforceability of the license. An attempt has been made to clarify and harmonize the law on such transactions through the Uniform Computer Information Transactions Act (UCITA), but this attempt received strong criticism and has only been passed in two states (Virginia and Maryland).[208]
US contract law is on common law or laws at state level. In theory this can pose problems concerning uniform application, but in general contract law has been largely harmonized through the adoption of the Uniform Commercial Code (UCC). On the other hand, it results in difficulties giving a legislative answer to legal ambiguity or divided court decisions.

5.3 PUBLIC SECTOR INFORMATION IN THE US: OPEN GOVERNMENT POLICY

The Obama administration has from the start in 2009 given a strong impulse for enlarging the availability and access to public sector information. It builds on pre-existing legislation for passive and active transparency, but developed a strong policy to strengthen active transparency.
Passive transparency, the giving access to information on request, is provided by the Freedom of Information Act. Active transparency, the providing of information on the initiative of the government, is regulated by the E-Government Act of 2002 and the Paperwork Reduction Act. The Paperwork Reduction Act dates from 1980, but was strongly revised in 1995. It has its roots in earlier efforts to lower the burden of administrative requests for information to citizens, going back to the Roosevelt administration during World War 2.[209] But it also established the institutional framework for an information management policy across the executive branch, with a supervisory and directing role at the Office of Management and Budget (OMB). The 1995 revision added the purposes: "(a) to improve the quality and use of

[208] De Filippi, op. cit., 2006, pp. 58-63; Bix, Brian, and Jane K. Winn, "Diverging perspectives on electronic contracting in the US and the EU", *Cleveland State Law Review*, vol. 54, no. 175, 2006, 176-181.
[209] Relyea, Harold, *Paperwork Reduction Act. Reauthorization and Government Information Management Issues*, Congressional Research Service, 4-1-2007.

federal information to strengthen decision-making, accountability, and openness in government and society; (b) to provide for the dissemination of public information in a manner that promotes the utility of the information to the public; and (c) to ensure the integrity, quality, and utility of the federal statistical system".[210] The E-government Act augmented this framework to adapt it to the digital environment. The actual information management policy is laid down in Circular A-130. It states that agencies have to plan for managing information throughout its life cycle. It also makes clear that agencies have a responsibility to provide information to the public, both on request as on own initiative. In this dissemination tasks agencies have to avoid restrictive practices, like exclusive or restricted licenses interfering with the availability of information. User charges may be no higher than the actual dissemination charges, with the costs associated with the collection and processing of the data excluded.[211]

The Obama administration added to this legislative framework of the government information policy a policy initiative by the executive branch to give a stronger implementation of open government policy.
On his first day in office president Obama in his Memorandum on Transparency and Open government instructed his administration to develop recommendations for a "system of transparency, public participation, and collaboration".[212] Rationale behind this policy is on the one hand strengthening democracy by enhancing accountability towards the public and participation of the public. On the other hand the objective is to make the government more effective by strengthening cooperation within the government and with private actors.
A first result was the Open Government Directive of 8 December 2009 by the OMB, presenting a policy road map for the implementation of open government by executive departments and agencies.[213] This has been followed by 2 National Action Plans on open government, in 2011[214] and 2013[215], each listing the policy initiatives during the coming 2 years, as well a more recent Open Data Action Plan[216]. The National Action Plans contained a wide range of initiative concerning open government, like on open data but also on FOIA, whistleblowers, public participation, etc. Here we will focus on the aspects concerning access and use of data.

The Open Government Directive instructed agencies to make more government information available online in open formats. When deciding about publishing information, the presumption should be in favor of openness, that is "to the extent permitted by law and subject to privacy, confidentiality, security, or other valid restrictions". Attention was not only to enlarging access, but also to re-use and interoperability by providing that the publication of information should be preferably in open formats. An open format is defined

[210] McDermott, Patrice, "Building open government", *Government Information Quarterly,* vol. 27, 2010, 405.
[211] Office of Management and Budget, Circular A-130 Revised, Transmittal Memorandum No. 4, 28 November 2000, http://www.whitehouse.gov/omb/circulars_a130_a130trans4
[212] White House, Memorandum on Transparency and Open Government, 21 January 2009. http://www.gpo.gov/fdsys/pkg/FR-2009-01-26/pdf/E9-1777.pdf
[213] Office of Management and Budget, Open Government Directive, M-10-06, 8 December 2009. http://www.whitehouse.gov/sites/default/files/omb/assets/memoranda_2010/m10-06.pdf
[214] White House, *The Open Government Partnership, National Action Plan for the United States of America*, 20 September 2011. http://www.whitehouse.gov/sites/default/files/us_national_action_plan_final_2.pdf
[215] White House, *The Open Government Partnership, Second Open Government National Action Plan for the United States of America*, 5 December 2013. http://www.whitehouse.gov/sites/default/files/docs/us_national_action_plan_6p.pdf
[216] White House, *US Open Data Action Plan*, 9 May 2014. http://www.whitehouse.gov/sites/default/files/microsites/ostp/us_open_data_action_plan.pdf

as "one that is platform independent, machine readable, and made available to the public without restrictions that would impede the re-use of that information." This definition contains elements of legal (no restrictions of re-use) and of semantic interoperability (platform independent, machine readable). Objective of the open format is that the information can be "retrieved, downloaded, indexed, and searched by commonly used web search applications".
Several practical steps were provided. Each agency had as a first step to identify 3 high-value data sets and make them available through Data.gov, which would be launched as the open data portal of the US government in May 2009. Each agency had to establish an Open Government webpage to document its activities related to Open Government and allow the public to give input and feedback. Also each agency had to develop an Open Government Plan, describing how it will improve transparency, public participation and collaboration and of which the minimum contents were prescribed in the Open Government Directive. On www.whitehouse.gov/open an Open Government Dashboard makes available each agency's Open government Plan and monitors the progress of implementation. Further steps were announced to improve the information on government spending, documented on websites like USAspending.gov. Existing OMB policies would be reviewed to identify impediments to open government.
The process of Open government plans is clearly driving projects to open data across the federal government (as part of the wider open government approach, which includes also public participation initiatives). A wide range of informative websites were established at agency level, as well as providing the information in machine readable formats.[217] The open data-portal website Data.gov contains after five years about 110000 datasets[218] from 447000 data resources[219].
The open government policy has been continued in 2 National Action Plans on open government, in 2011[220] and 2013[221], each listing the policy initiatives during the coming 2 years. Specific on data it furthers stresses to enlarge the data available through support of the implementation by agencies of their open government plans. This has been followed recently by a US Open Data Action plan, which integrates the open data policy developed as part of the digital strategy and lists the major planned releases of datasets.[222]

The Obama administration also played a leading role in lifting this open government approach to the international level. First through the Open Government Partnership (OGP)[223], launched in 2011 and with a similar broad agenda including accountability and public participation. Further with the G8 Open Data Charter[224] in 2013, which focuses on open data. Both initiatives work with country action plans detailing the actions countries want to take

[217] Examples can be found in the reports on the open government policy at www.whitehouse.gov/open.
[218] Number of datasets provided at the search function of www.data.gov at the moment of writing is 110844 (accessed on 1 July 2014). Main contributors were the National Oceanic and Atmospheric Administration (32529), US Fish and Wildlife Service (10594) and U.S. Geological Survey (9952).
[219] Jeanne Holm, "Five Years of Open Data—Making a Difference", 20 May 2014. www.data.gov/agencies/five-years-open-data-making-difference/
[220] White House, *The Open Government Partnership, National Action Plan for the United States of America*, 20 September 2011. http://www.whitehouse.gov/sites/default/files/us_national_action_plan_final_2.pdf
[221] White House, *The Open Government Partnership, Second Open Government National Action Plan for the United States of America*, 5 December 2013. http://www.whitehouse.gov/sites/default/files/docs/us_national_action_plan_6p.pdf
[222] White House, *US Open Data Action Plan*, 9 May 2014. http://www.whitehouse.gov/sites/default/files/microsites/ostp/us_open_data_action_plan.pdf
[223] http://www.opengovpartnership.org
[224] UK Cabinet Office, *Open Data Charter*, UK Presidency of G8 2013. https://www.gov.uk/government/publications/open-data-charter

based on common principles. The National Action Plans mentioned above are part of the US commitment to the OGP.

A second policy initiative, distinct from but strengthening the open government initiative's attention for open data, focused on digital government. Rationale is to make government more effective and adapted to the new digital environment. The Federal Chief Information Officer (CIO) published on 23 May 2012 a digital government strategy entitled "Digital Government: Building a 21st Century Platform to Better Serve the American People".[225] This strategy lists 4 main principles: an information-centric approach, a shared platform approach, a customer-centric approach and a ensuring security and privacy.
Most important from our viewpoint of data policies is the information-centric approach, as it introduces an attention for semantic interoperability. It promotes a shift in thinking about digital information, away from the old approach focused primarily on presentation. An information-centric approach should focus on making data and content accurate, available and secure. It needs to turn unstructured content into structured data and to associate this structured data with valid metadata. Providing this data through web APIs enhances interoperability and makes the data assets widely available. It also supports device-agnostic security and privacy controls, shifting the focus from securing devices to securing data.
Interoperability is also addressed by the other principles. Making more use of shared platforms instead of each agency developing its own systems, is more cost effective but also enhances technical interoperability. Similarly, the focus on customers' needs wherever he is pushes towards a program- and device-agnostic service, which again puts the focus on technical interoperability.
To put the information-centric approach into practice the development is foreseen by OMB of an open data, content, and web API policy for the federal government. This adds to making data available attention for improving semantic interoperability by adding metadata. Further agencies are instructed to make high-value data in at least 2 existing major customer-facing services available through web APIs.

This strategy was followed up by the publication of a Memorandum on Open Data Policy[226] by OMB and an Executive Order of the president on 'Making Open and Machine Readable the New Default for Government Information'[227], instructing the implementation of this Open Data Policy.
Open data is defined in this Open Data Policy as “publicly available data structured in a way that enables the data to be fully discoverable and usable by end users”. A list of principles is given with which open data must be consistent with, like:

- Public: agencies must adopt a presumption in favor of openness.
- Accessible: data is made available in “open formats that can be retrieved, downloaded, indexed, and searched” and which are machine-readable (that is “reasonably structured to allow automated processing”). Such open data structures have to avoid discrimination and allow a wide range of users and purposes. This can mean that data is provided in multiple formats. Re-use is promoted by stating that “To

[225] White House, *Digital Government: Building a 21st Century Platform to Better Serve the American People*, 23 May 2012. http://www.whitehouse.gov/sites/default/files/omb/egov/digital-government/digital-government.html
[226] Office of Management and Budget, Memorandum on Open Data Policy-Managing Information as an Asset, M-13-13, 9 May 2013. http://www.whitehouse.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf
[227] White House, Executive Order - Making Open and Machine Readable the New Default for Government Information, 9 May 2013. http://www.whitehouse.gov/the-press-office/2013/05/09/executive-order-making-open-and-machine-readable-new-default-government-

the extent permitted by law, these formats should be non-proprietary, publicly available, and no restrictions should be placed upon their use".

- Described: Open data are described fully, through the use of "robust, granular metadata", documentation of data elements, data dictionaries, and other additional descriptions.
- Reusable: "Open data are made available under an open license that places no restrictions on their use."
- Complete: "Open data are published in primary forms (i.e., as collected at the source), with the finest possible level of granularity that is practicable and permitted by law and other requirements."
- Timely
- Managed Post-Release

This definition and principles of open data clearly promote legal and semantic interoperability, above promoting availability in full detail and granularity.
The Open data policy further describes the management of information according to these principles throughout the information life cycle. It prescribes that:

- information has to be collected or created supporting downstream information processing (including interoperability of information systems) and dissemination activities. This includes the use of machine-readable and open formats from the moment of collecting or creating, the use of data standards and common core and extensible metadata, as well as the use of open licenses.
- information systems have to be designed and build to support interoperability and information accessibility.
- data management and release practices have to be strengthened by creating and maintaining an enterprise data inventory, as well as public data listing. Agencies must engage customers to prioritize data release and make sure that roles and responsibilities are clear.
- at the same time privacy and confidentiality has to be protected and data has to be secured. The policy points to the 'mosaic effect', where information, which poses no risk for identification in an individual dataset, can lead to identification when combined with other available information. Therefore it prescribes a risk-based analysis before deciding to release data.
- these new interoperability and openness requirements have to be incorporated into core agency processes.

Further implementation of the Digital Strategy[228] include the availability of APIs (Application Programming Interface) to allow automated access and processing of the available data. These APIs are listed on http://www.data.gov/developers/apis. Also a new data catalog was released on http://catalog.data.gov.

5.4 PASSIVE TRANSPARENCY THROUGH THE FREEDOM OF INFORMATION ACT

Also concerning passive transparency the Obama administration took action. At his first day in office president Obama also published a memorandum on the Freedom of Information Act.[229] In this memorandum he stated the objective of encouraging accountability through transparency and instructed that all agencies should use the presumption of openness when

[228] Further implementation can be tracked on http://www.whitehouse.gov/digitalgov/deliverables
[229] White House, Memorandum - Freedom of Information Act, 21 January 2009. http://www.whitehouse.gov/the_press_office/Freedom_of_Information_Act

dealing with FOIA requests. "The Freedom of Information Act should be administered with a clear presumption: In the face of doubt, openness prevails." The practical step was the instruction to the Attorney General to issue new guidelines governing the FOIA, which were published in March 2009.[230]

The Freedom of Information Act[231] provides, aside of some basic active transparency rules, the framework for passive transparency, or the providing of information on request. Agencies have to make records available upon any request which "reasonably describes such records"[232]. This implies that the agency has to do some effort within reasonable limits to search for the information. Each agency establishes its procedures for such request and can ask fees. These fees are limited to the duplication cost when the records are not for commercial use and the request is made by an educational or noncommercial scientific institution or by news media. For commercial use the fees can include the cost for document search, duplication and review (to determine if the document can be disclosed and which parts have to be withheld), for all other cases only the cost for search and duplication. But the charges can be dropped or reduced if the disclosure is in public interest and not primarily in the commercial interest of the requester.
Exceptions are provided and include:

- classified information or authorized to be kept secret in the interest of national defense or foreign policy
- information relating to internal personnel rules and practices of an agency
- information specifically exempted from disclosure by statute
- trade secrets and commercial or financial information
- personnel and medical files and similar files related to personal privacy
- records or information compiled for law enforcement purposes, when this would have negative effects like disclosing sources or investigative techniques, being harmful to privacy or the right to a fair trial, dangerous for people or interfering with enforcement proceedings
- certain information related to regulation or supervision of financial institutions
- inter-agency or intra-agency memorandums or letters which would only be available by litigation
- geological and geophysical information and data, including maps, concerning wells.[233]

Records can be partially released after removing the parts exempted from disclosure by these exceptions.
These limitations also circumscribe the limitations for active transparency.

The policy initiative of president Obama did not change this legislative framework, but instructed the agencies of the executive branch on the application of the FOIA. His memorandum provided a decision rule to let openness prevail. The memorandum of the Attorney General clarified this presumption. First it meant that an agency has to make discretionary disclosures, instead of withholding information because it can legally do so by showing that it technically fell under an exemption. Secondly, that an agency must consider partial disclosure when full disclosure is not possible, instead of refusing disclosure immediately.

[230] Attorney General, Memorandum - Freedom of Information Act, 19 March 2009,.http://www.justice.gov/ag/foia-memo-march2009.pdf
[231] US, Freedom of Information Act, 5 U.S.C. § 552.
[232] US, Freedom of Information Act, 5 U.S.C. § 552 (a)(3)(A)
[233] US, Freedom of Information Act, 5 U.S.C. § 552 (b)

In other words, the Obama policy re-interpreted the grounds for refusal as being relative grounds instead of absolute grounds. This discretionary rule also applies in the decision making process about publishing datasets. By consequence access to PSI is made easier.

5.5 PRIVACY LAWS

A major limitation in the use of data is privacy law. The US does not know a comprehensive data protection framework like the EU. Therefore no general limitations exists concerning personal data, nor the duty to grant certain rights like access to the data subject. But data collection and use by government and private actors is limited and regulated in several specific area's through piecemeal privacy laws or other protections.

The 4th Amendment to the US Constitution protects people "in their persons, houses, papers, and effects" against the government and prevents the government from conducting "unreasonable searches and seizures". Such searches are only allowed with a warrant and upon probable cause. In the Katz decision concerning wiretapping of a public telephone the US Supreme Court has widened the interpretation of this protection, by requiring a warrant when the "reasonable expectation of privacy" would be violated. But in general this protection is limited by the 'third party doctrine', which holds that no reasonable expectation of privacy remains whenever the behavior of a person of information concerning a person is exposed in whatever way to public.[234] Similarly, no 4th Amendment protection applies whenever a person exposes information to another entity. This implies that no constitutional limitation apply to camera's in public places. Neither to obtaining records concerning a person at private companies. Also, this 4th Amendment protection only applies towards the government and not towards private actors. Specific laws have limited the 'third party doctrine' and widened the 4th Amendment protection to certain categories of information at other private actors, like communication data (Electronic Communications Privacy Act of 1986) or bank records (Right to Financial Privacy Act of 1978), but these laws have been weakened again after 9/11 by the Patriot Act.
The growing use of computers and the surveillance scandals from the Nixon and FBI director Hoover-era led to the formulation of Fair Information Practice Principles by a 1973 Advisory Committee.[235] The Code of Fair Information Practices is based on five principles:

- There must be no personal data record-keeping systems whose very existence is secret.
- There must be a way for a person to find out what information about the person is in a record and how it is used.
- There must be a way for a person to prevent information about the person that was obtained for one purpose from being used or made available for other purposes without the person's consent.
- There must be a way for a person to correct or amend a record of identifiable information about the person.
- Any organization creating, maintaining, using, or disseminating records of identifiable personal data must assure the reliability of the data for their intended use and must take precautions to prevent misuses of the data.

[234] Solove, Daniel, *Nothing to hide : the false tradeoff between privacy and security*, Yale University Press, New Haven, 2011.
[235] US Department of Health, Education & Welfare, *Records, Computers and the Rights of Citizens, Report of the Secretary's Advisory Committee on Automated Personal Data Systems*, July 1973. epic.org/privacy/hew1973report/Summary.htm

The FIPP are quite similar to the principles underlying data protection in Europe, but have only be put into law in specific areas. The Privacy Act of 1974 implements the FIPP in the government and is the main legal framework concerning the treatment of personal information by the federal government (not the states). This Act regulates and restricts the collection and retention of personal data, and the disclosure thereof by government agencies to a specific range of officials or for certain uses, or when requested or with consent by the person concerned. Further does it grant individuals a right of information, access and amendment or correction. Important loophole is that this Act does not apply to private companies selling personal data to the government.[236]
Important in the context of big data is the amendment added by the Computer Matching and Privacy Protection Act (CMPPA) of 1988. This states that no record may be disclosed to another agency for use in a computer matching program, except pursuant an agreement detailing purpose, legal authority, a description of the records including of each data element that will be used, as well as a range of procedures concerning the treatment of the data. These matching operations are under supervision of Data Integrity Boards within each agency. The agreements are available on request to the public. Further before a negative decision based on computer matching concerning payments to the individual may be taken, the agency has to verify the information and notify the individual and give him the opportunity to contest the findings. This application of this strict regulation of computer matching has been excluded for a lot of uses, like research, law enforcement, tax, foreign counterintelligence, …

Privacy law between private actors was first established through tort law. Four privacy tort actions are recognized in the Second Restatement of Torts, but these have no practical relevance for big data.
Since the 1970's a range of laws containing privacy protection for specific sectors have been established, like the Fair Credit Reporting Act (FCRA) of 1970, the Cable Communications Policy Act of 1984, the Videotape Privacy Protection Act of 1988, the Telephone Consumer Protection Act of 1991 and the Telecommunications Act of 1996, the Health Insurance Portability and Accountability Act (HIPAA) of 1996, the Children's Online Privacy Protection Act (COPPA) of 1998, the Financial Modernization Act of 1999, ...
We cannot go in detail on all these laws, but in general they implement the FIPP or part of it in the specific sectors and by consequence have an effect on big data practices with data regulated by these laws.
Further is the role of the Federal Trade Commission (FTC) very important in regulating privacy in the private sector. The FTC regulates and supervises market practices and has the authority to enforce trade law through investigatory and litigation powers. Basic consumer protection is provided by the FTC Act, which forbids “unfair or deceptive acts or practices in or affecting commerce”, while the FTC also has the authority to enforce other specific consumer protection laws, like the FCRA or COPPA, and the EU-US Safe Harbor Framework. The FTC has taken up the role of the de facto data protection authority by enforcing privacy policies of companies. Similarly to licenses, the legal status of privacy policies has been ambiguous and enforcement under contract law failed in practice. The FTC has treated violations by a company of its published privacy policy as such a deceptive and in several occasions unfair act. It has slowly developed through settlements a common law-like jurisprudence establishing norms concerning transparency, data collection and use, and data security. This jurisprudence evolved towards also treating the disrespect of industry standards

[236] Chris Jay Hoofnagle, “Big Brother's Little Helpers: How ChoicePoint and Other Commercial Data Brokers Collect, Process, and Package Your Data for Law Enforcement”, *N.C.J. Int'l L. & Com. Reg*., Vol. 29, No. 595, Summer 2004.

on these issues as a form of deceptive act. FTC settlements and opinions are therefore also an important source of law.[237]

The Obama administration has also taken the initiative to remedy the piecemeal privacy law by an overall consumer privacy regulation, called the Consumer Privacy Bill of Rights. This Consumer Privacy Bill of Rights gives a wider implementation of the FIPPs in the digital economy. It provides for:

- "Individual Control: Consumers have a right to exercise control over what personal data companies collect from them and how they use it.
- Transparency: Consumers have a right to easily understandable and accessible information about privacy and security practices.
- Respect for Context: Consumers have a right to expect that companies will collect, use, and disclose personal data in ways that are consistent with the context in which consumers provide the data.
- Security: Consumers have a right to secure and responsible handling of personal data.
- Access and Accuracy: Consumers have a right to access and correct personal data in usable formats, in a manner that is appropriate to the sensitivity of the data and the risk of adverse consequences to consumers if the data is inaccurate.
- Focused Collection: Consumers have a right to reasonable limits on the personal data that companies collect and retain.
- Accountability: Consumers have a right to have personal data handled by companies with appropriate measures in place to assure they adhere to the Consumer Privacy Bill of Rights."[238]

These principles will be further developed through multistakeholder processes in order to develop enforceable Codes of Conduct. The FTC would enforce this Bill. This can happen through a new authority provided by law or through its authority to prohibit deceptive and unfair practices. The Obama administration takes a double approach towards the further development. It prefers to enact the Consumer Privacy Bill of Rights through legislation in order to increase legal certainty, but if Congress does not want to vote this proposal into law, the implementation can anyway go on through the development of codes of conduct.
The law proposal itself has not seen a lot of action the last 2 years in Congress. But privacy multistakeholder processes are already convened by National Telecommunications & Information Administration (NTIA) under the Department of Commerce. This has resulted in a Code of Conduct for transparency in mobile apps[239], while such a process is ongoing concerning the commercial use of facial recognition technology[240].

The privacy regulations discussed till now are not specific to big data, but have an important impact on big data practices. The last years and especially the last months, a lot of high quality policy debate has been taken place concerning big data and privacy in the US, reflected in several important reports. One focus were data brokers, the other focus was specifically on big data and privacy.

[237] Solove, Daniel J. and Hartzog, Woodrow, "The FTC and the New Common Law of Privacy", *Columbia Law Review*, Vol. 114, No. 583, 2014.
[238] White House, *Consumer Data Privacy in a Networked World: A Framework for Protecting Privacy and Promoting Innovation in the Global Digital Economy*, 23 February 2012.
www.whitehouse.gov/sites/default/files/privacy-final.pdf
[239] NTIA, "Privacy Multistakeholder Process: Mobile Application Transparency", 12 Nov 2013.
www.ntia.doc.gov/other-publication/2013/privacy-multistakeholder-process-mobile-application-transparency
[240] Ibid.

The Government Accountability Office (GAO)[241], the Committee on Commerce, Science and Transportation in the Senate[242] and the FTC[243] have investigated the data broker industry and the problems it poses concerning privacy. The reports all conclude that consumers can not exercise rights foreseen in FIPPs towards this industry. The FTC recommends to subject the different branches of this industry to legislation similar to FCRA and to assure transparency, access and amendment for consumers.

As a reaction to the upheaval caused by the revelations by John Snowden about the NSA surveillance, president Obama launched a Big data review. This review was focused on big data and privacy, and mostly outside the NSA surveillance context. It resulted in 2 reports which are relevant for our topic. The first report[244] was made by a working group of senior Administration officials led by John Podesta and resulted from a broad process with stakeholder consultations and academic workshops. This report of the Big Data and privacy working group gives an overview of big data practices in the public and private sector, and points to both the positive gains as the dangers involved. It notes 5 areas where big data presents challenges: privacy in the marketplace, schools where big data can enhance learning opportunities but also presents privacy risks, the danger of new forms of discrimination and using data as a public resource. The report gives the following recommendations:

- Advance the Consumer Privacy Bill of Rights
- Pass national data breach legislation (like notification duties in case of data breach)
- Extend privacy protections to non-US persons (e.g. by applying the Privacy Act of 1974 to non-US persons)
- Ensure data collected on students in school is used for educational purposes (by modernizing the relevant privacy laws)
- Expand technical expertise to stop discrimination (at the lead civil rights and consumer protection agencies,, like the FTC or the Equal Employment Opportunity Commission)
- Amend the Electronic Communications Privacy Act. (to ensure that the protection offered online is consistent with the protection in the physical world)

Further, the President's Council of Advisors for Science and Technology (PCAST) conducted a parallel study of the technological trends underpinning big data, in order to assess the technical feasibilities of different policy approaches.[245] Also this report start with a broad sketch of uses of big data and the possible tradeoffs between privacy, security and convenience. PCAST states that a policy focusing on limiting data collection is not a broadly applicable or scalable strategy. Also because a lot of privacy problems arise after the collection with the fusion of data sources. It argues that the use of data is the place where

[241] United States Government Accountability Office, *Information Resellers. Consumer Privacy Framework Needs to Reflect Changes in Technology and the Marketplace*, GAO-13-663, September 2013. www.gao.gov/assets/660/658151.pdf

[242] Senate Committee on Commerce, Science, and Transportation, *A Review of the Data Broker Industry: Collection, Use, and Sale of Consumer Data for Marketing Purposes*, 18 December 2013. www.commerce.senate.gov/public/?a=Files.Serve&File_id=0d2b3642-6221-4888-a631-08f2f255b577

[243] Federal Trade Commission, *Data Brokers. A Call for Transparency and Accountability*, May 2014. www.ftc.gov/system/files/documents/reports/data-brokers-call-transparency-accountability-report-federal-trade-commission-may-2014/140527databrokerreport.pdf

[244] White House, *Big Data: Seizing Opportunities, Preserving Values*, 1 May 2014. www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_5.1.14_final_print.pdf

[245] President's Council of Advisors on Science and Technology, *Big Data: A Technological Perspective*, White House, 1 May 2014. http://www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_big_data_and_privacy_-_may_2014.pdf

consequences are produced and the technically most feasible place for protections. Further, some techniques for privacy protection used in the past do not seem robust anymore in the context of big data, like anonymization, data deletion (as old data sources can later prove useful in combination with others, while it often becomes impossible to surface all the data known about an individual) or distinguishing the treatment of data from metadata (as metadata can be as much a risk to privacy as the data itself). Also the notice and consent framework is considered unworkable. This framework places the burden of privacy at the individual, while this individual is placed in an unequal position in relation with the provider. The responsibility for using the personal data in accordance with the preferences of the data subject should better be shifted to the provider.
Interesting is that this assessment puts doubt to the robustness of FIPPs, which still are underlying the Consumer Privacy Bill of Rights and other privacy regulations. PCAST still endorses these principles as sound, but states that big data puts effective operationalization at risk. It suggests several adaptations in line with its recommendation to focus on use of data instead of data collection. Concerning rights meant as consumer empowerments, PCAST recommends to recast these empowerments as obligations of the entity using the data whenever such empowerment has become practically impossible to exercise in a meaningful way.

5.6 CONCLUSIONS

In general we can see clear differences in the IPR framework and the privacy regulations, which make big data operations easier in the US compared to the EU. Legal frameworks concerning access and re-use of PSI are relatively similar in the US and US. If not in their details and origins, they have at least a similar impact on big data operations.

The copyright framework in the US has less problems with new technologies, including big data. It has a weaker protection of databases compared to the EU, which makes re-use of data easier. And its fair use-system of exceptions to copyright protection is more compatible with data mining.
Licensing has similar problems as in the EU when it becomes more than a mere authorization and turns into a contract.

We can conclude our review on access and use of PSI with the observation that both the US and the EU have access regimes to PSI. Open data policies have become a policy topic around the same period.
In the US the rationale was more focused on transparency and public control of the government, where the EU policy takes clearly an economic view. The data is seen as an enabler of democratic control, not of business processes. Such differences have perhaps more to do with on which grounds such policy can be legitimated or linked to competences. But an important consequence is that open data functions less as part of a market policy or a vision concerning value creation. In the policy documents on open data we find strong attention for semantic interoperability, but much less for involving stakeholders.

Privacy law consists in the US of sectorial regulations. A general data protection framework does not exist. Important also is that such privacy regulations are in general a part of consumer protection. The constitutional protection of privacy only applies to governmental intrusions of privacy and not to those of private actors.
Result is that data (re-)use gets much less limited by privacy law compared to the EU. Both public authorities and companies can much easier commercialize personal data, like census

data. And public authorities publish in general much more personal data compared to those in EU Member States. This gives much more freedom in collecting, combining and using data from a wide range of sources. As such it was an enabling factor in the advent of data-intensive companies like Google.
Of course this wider freedom comes with a price. Privacy concerns have led to recent high-level attention for privacy risks linked with big data. Although the EU approach is generally seen as too strict in its approach to privacy, the Consumer Privacy Bill of Rights would introduce a much more general privacy protection. The Fair Information Practices included in this Bill are quite similar to those in the Safe Harbor-framework. This Bill would bring the US privacy regulation much nearer to that in Europe. In the academic and political debate we see the same doubts raised about the compatibility of these FIPPs with big data processing.

6 GLOBAL DATA POLICIES BY PRIVATE ACTORS

6.1 US COMMERCIAL COMPANIES

Large companies work on a global scale, so also the large big data pioneers like Google, Facebook, Twitter or Amazon. They shape flows of data and regulate access to and interoperability between the large datasets they have under their custody and other datasets or services. This shaping is done on a technical level through the user interface of their services and APIs (application programming interfaces), and on a legal level with terms of use and privacy policies.
An API enables the access to the service and the data in the cloud controlled by the service provider. This API also allows for access control through authentication mechanisms like API-keys (as we will see for Twitter and Google) and to subject this access to constraints like request limits or other limits on downloading or availability of data. The API ensures or restricts the technical interoperability with the dataset. The API, together with the dataset it controls, also plays a role in semantic interoperability through the form and method with which data is provided. The license defines the legal interoperability (as far as possible within the legal frameworks it has to function).
Through these instruments the global big data service providers try to develop a distinct and partially open, partially closed data ecosystem favorable to their business model, in the literature often designated as platform politics. As a whole these platform politics or business models define the organizational interoperability, or how this dataset can be linked up to provide other services.
In this part we will look at a couple of those data enclosures. We will focus on the legal level, but it is important to keep in mind that the API and the operations it allows also are part of what can be seen as a big data policy.

The word license has again several meanings. It gives permission to do something which would otherwise be forbidden, but this otherwise forbidden behavior can have different legal bases and qualifications. The license can be a copyright license, giving permission to use the copyrighted data or database in certain ways. But it is often also a service license, giving permission to access and use the internet service in certain ways. And this license can be a bare license or part of a contract.

These companies get confronted with fragmented legal regulations, in the first place on IPR and contracts but also other legal frameworks like privacy. This means that the qualification and the enforceability of licenses can be different depending on the jurisdiction in which it has to function. Together with presenting the way how these big data actors shape their data ecosystem, we look at the legal problems and obstacles they get confronted with.

Several stakeholders are present in such a big data ecosystem. First the big data companies assemble their dataset. This can be through:

- buying or licensing copyrighted material (e.g. satellite pictures in Google Maps) or creating own content (e.g. Street View).
- user-generated content through offering services (e.g. Twitter, Facebook, YouTube or Google Maps). In this case the license will define the rights on the content.
- data collection on the internet through web crawlers (e.g. Google search and other search engines). In this case the data is public domain or the use of copyrighted material is supposed to be allowed by the exceptions in the copyright law (e.g. fair use).

- data on user behavior when using the services. This overlaps partly with user-generated content, but important is that it often is or contains personal data and that this data collection is not done as a service to the users but for targeted advertising towards these users.

Later access to this big data is offered. This can be as part of:

- the service to the users (e.g. search engines, Google Maps)
- a different service to another category of users, like targeted advertising towards users of the main service (e.g. Google, Facebook, Twitter) or analysis for marketing or other purposes (e.g. Twitter)

Both stages can present very different forms of access and attitudes towards re-use and sharing, and will be covered by different licenses.

Twitter

A first global big data company we look at is Twitter. Big data at Twitter means a user-generated dataset of communications. Twitter offers a communication platform to users, on which can send tweets or short text messages (max. 140 characters). They can develop a public that follows them or follow themselves the tweets of people they are interested in. Interaction is also possible with an easy method for referring and reacting to each other. Started with the first tweet in March 21, 2006, Twitter developed in a major communication platform with 255 million monthly active users and 500 million tweets sent per day[246] and figures at the 8th position in the Alexa-list of most visited websites[247].
All this tweets together form a gigantic dataset of communications and links between people. Analysis of this data, as a historical archive or through real-time monitoring, can give insights in what is hot in the public debate, in the popularity of topics and products, in who talks with whom and who is influential, etc. All the data collected from the users of the communication platform forms the real treasure on which Twitter builds its business model. This business model is built on using this data and the access to the users for targeted advertising, but it includes also selling access to this data. In this context it is the access to this data which interests us.
Twitter does give free access to this data through its API, but only to a limited sample. To get access to a larger subset or the whole you have to buy this access. Twitter has limited such access to a small set of companies, the certified partners of Twitter. And access to the 'Twitter Firehose', which delivers all the tweets, is commercialized through only 4 companies (Gnip, Topsy, Datasift and NTT Data), which pay heavily for this access. Over time Twitter did limit the access to the Firehose and locked out some of the companies which developed around it like PeopleBrowsr. Gnip has been bought up by Twitter itself to keep one of the resellers under its control, after an attempt by Apple to get access to the Twitter ecology through buying Topsy.[248] As a whole, this leads to a tightly controlled ecology of companies with access to the data.

We look in more detail to the free uses provided by Twitter, as user of the communication platform and as user of the Twitter data through the API. Both uses are regulated through terms of use.
A user signing up to use the Twitter communication platform is greeted upon with a clickwrap license as the form states: “By clicking the button, you agree to the terms

[246] Twitter, https://about.twitter.com/company
[247] Alexa, “The top 500 sites on the web”, no date. http://www.alexa.com/topsites
[248] The Guardian, “Twitter buys Gnip, one of only four companies with 'firehose' access”, 16 April 2014. http://www.theguardian.com/technology/2014/apr/16/twitter-buys-gnip-firehose-analytics-apple-topsy

below"[249]. In other words, by signing up a user enters in a contractual agreement, which contains the Terms of Service[250] (which includes also another text: the Twitter Rules[251]), the Privacy Policy[252] and a text on Twitter's use of cookies and similar technologies[253]. The contractual nature is emphasized in the Terms of Service: "Your access to and use of the Services are conditioned on your acceptance of and compliance with these Terms. By accessing or using the Services you agree to be bound by these Terms. " and "You may use the Services only if you can form a binding contract with Twitter". In the terms Twitter gives the user a license to use its services, with a range of conditions and restrictions.

In these terms the legal status of the data is also regulated. A user retains his rights on the content he posts, displays or submits, but grants a world-wide, non-exclusive and royalty-free license to Twitter, including the rights to sublicense.[254] This license to Twitter is for a wide range of uses ("use, copy, reproduce, process, adapt, modify, publish, transmit, display and distribute") and includes the transfer to other companies.

As a tweet is a short message it is doubtful it has the level of originality to be protected by copyright, but length is as such not a criterion for copyright and it is possible that at least some tweets are protected by copyright.[255] With this license a user allows Twitter to publish these tweets anywhere and in whatever form and to transfer them to other companies.

A tweet also has a sender and his name and the twitter handle or username are personal data, while the handle links to a personal profile containing other personal data. Users give through accepting the Privacy Policy as part of the terms, their consent to "the collection, transfer, manipulation, storage, disclosure and other uses of your information". This also includes an authorization for the transfer to other countries: "Irrespective of which country you reside in or supply information from, you authorize Twitter to use your information in the United States and any other country where Twitter operates". The Privacy Policy list the types of personal information that are collected and states how they are used. Concerning "Tweets, Following, Lists and other Public Information" it states that the Twitter service is designed to share information and that the default is that the information provided is public, although users can change this default to a more restrictive use. This information includes the tweets but also " the lists you create, the people you follow, the Tweets you mark as favorites or Retweet, and many other bits of information that result from your use of the Services". It is stated that the information considered public is instantly disseminated and that all this information is searchable by search engines. Concerning information sharing and disclosure the policy states that Twitter "may share or disclose your non-private, aggregated or otherwise non-personal information", but non-private and non-personal information is all the information the user allows to be public. This includes "your public user profile information, public Tweets, the people you follow or that follow you".

To conclude, a user of the Twitter service allows Twitter to gather, publish and transfer to other companies a whole range of information, including personal information. Thanks to these permissions Twitter can build up a huge dataset of communications and communicating users.

[249] https://twitter.com/signup
[250] Twitter, *Terms of Service*, version June 25, 2012, https://twitter.com/tos
[251] Twitter, *Twitter Rules*, https://support.twitter.com/articles/18311-the-twitter-rules#
[252] Twitter, *Privacy Policy*, version October 21, 2013, https://twitter.com/privacy
[253] Twitter, "Twitter's use of cookies and similar technologies", https://support.twitter.com/articles/20170514#
[254] Twitter, *Terms of Service*, version June 25, 2012, art. 5, https://twitter.com/tos
[255] Beurskens, Michael, "Legal Questions of Twitter Research", in Katrin Weller et al., *Twitter and Society*, Peter Lang Publishing, Inc., New York, 2014, 125-127.

The access to and the data flow out of this dataset proves to be much more restrictive. Access to the data is obtained by using one of the Twitter APIs. Through one of these APIs application can request tweets from an account, perform searches, etc. Using an API involves also an authentication mechanism with an API-key.
When developers register an application they can obtain an API-key. Such registration includes clicking to agree with the 'Developer Rules of the Road', or the terms of use of the API.[256] This implies a contractual agreement with Twitter, which also includes the earlier mentioned Terms of Use and connected documents.
In some cases, like searches on the twitter website or through displaying tweets with 'web intents', it is possible to get some data without clicking to agree. Although Twitter mentions that also for such uses the Terms of Use apply, in these cases there is no contractual agreement and Twitter can only enforce the Terms by disallowing access in a technical way or sue for copyright violations if any. But these cases involve only a marginal use of data. When more robust access to the Twitter data is wished, registration and use of the API involving the authentication mechanisms is needed.
We will not go deeper into the technical details, but look at the API terms.

First of all the general terms of use make clear that only these APIs may be used to obtain data, or that a separate agreement with Twitter has to be made (like the commercial agreements it has with its certified partners):
"You may not do any of the following while accessing or using the Services: … (iii) access or search or attempt to access or search the Services by any means (automated or otherwise) other than through our currently available, published interfaces that are provided by Twitter (and only pursuant to those terms and conditions), unless you have been specifically allowed to do so in a separate agreement with Twitter (NOTE: crawling the Services is permissible if done in accordance with the provisions of the robots.txt file, however, scraping the Services without the prior consent of Twitter is expressly prohibited)".[257]

The API terms[258] make clear that access is limited: "You will not attempt or encourage others to:
A. sell, rent, lease, sublicense, redistribute, or syndicate access to the Twitter API or Twitter Content to any third party without prior written approval from Twitter.

- If you provide downloadable datasets of Twitter Content or an API that returns Twitter Content, you may only return IDs (including tweet IDs and user IDs).
- You may provide spreadsheet or PDF files or other export functionality via non--programmatic means, such as using a "save as" button, for up to 100,000 public Tweets and/or User Objects per user per day. Exporting Twitter Content to a datastore as a service or other cloud based service, however, is not permitted."

These limitations do not allow large scale access and analysis of the Twitter dataset. Who wants to do such analysis has to buy access from Twitter or one of its certified resellers.
Twitter does not transfer copyright over its data: "You expressly acknowledge that Twitter and its end users retain all worldwide right, title and interest in and to the Twitter Content, including all intellectual property rights therein. You also acknowledge that as between you and Twitter, Twitter owns all right, title and interest in and to the Twitter API, Twitter Marks, and the Twitter service (and any derivative works or enhancements thereof), including but not limited to all intellectual property rights therein." Instead the API terms do give a permission

[256] Twitter, *Developer Rules of the Road*, version July 2, 2013. https://dev.twitter.com/terms/api-terms
[257] Twitter, *Terms of Service*, version June 25, 2012, art. 8, https://twitter.com/tos
[258] Twitter, *Developer Rules of the Road*, version July 2, 2013. https://dev.twitter.com/terms/api-terms

to use the dataset as provided through the API and subject to the limitations build in this API: "You may use the Twitter API and Twitter Content in connection with the products or services you provide (your "Service") to search, display, analyze, retrieve, view, and submit information to or on Twitter. … Your use of the Twitter API and Twitter Content are subject to certain limitations on access, calls, and use as set forth in the Rules, on dev.twitter.com, or as otherwise provided to you by Twitter. If Twitter believes that you have attempted to exceed or circumvent these limitations, your ability to use the Twitter API and Twitter Content may be temporarily or permanently blocked."

Through this licensing practice for small-scale access and separate contracts with certified companies for larger scale access, Twitter controls and shapes the ecology of applications and analysis practices build on top of its dataset of users communication.
It would be difficult to control access of this dataset purely based on copyright. It is doubtful copyright exist on most of the tweets, and fair use exemptions of the database copyright would probably allow to use a much larger part without any agreement. But through its licensing practices Twitter can substitute the default copyright regime with a much more restrictive access regime and commercially exploit this access control.

Google

We will further look at some big data practices by Google. Google can be considered as the big data pioneer and also developed several techniques now in common use for big data, like the map-reduce method. It is the most visited website globally and offers a wide range of services through which it collects data for targeted advertising towards users of its services. Starting from the Google search-service it has widened its services: specialized search engines like Google News, Google Scholar or Google Books, geographical information with Google Earth and Maps, website statistics with Google Analytics, E-mail service with Gmail, video with YouTube … Here we will look at Google Search and Google Maps.

Google Search gathers its data through web crawling. It reads and stores webpages encountered through its web crawling programs, analyses them and build its search database based on that information.
This practice of data collection is not without legal hurdles. It hits upon copyright over webpages and privacy rights. Google has had a range of legal battles concerning copyright with services like Google Search, Google News and Google Books, while privacy issues were raised concerning Street View and Google Search and its sub-services.
In copyright discussions Google in general claimed its data collection stayed within the exceptions to the exclusive rights granted by copyright law. As said before in the US an exception exists for fair use, which leaves space for a wide range of activities, while European copyright laws contain exceptions for press clippings, quoting etc. In cases concerning Google News, it contested that titles of news articles, which are shown as search results, were original enough to be protected by copyright. Also, Google claims its search engine only automatically transmits information published elsewhere and falls within the remit of the safe harbor-clauses for communication services in the Digital Millennium Copyright Act (DMCA) in the US and the E-commerce directive in the EU or the temporary copy-exception in the InfoSoc-directive. Further Google claimed publishers can easily opt out

through the robots.txt-file. This file indicates to web crawlers to not include a webpage in search results and is a generally accepted standard respected by search engines.[259]
Most cases ended in settlements, but the views of Google were not accepted in the Belgian Copiepresse SCRL v. Google Inc.-case concerning Google News. On the other hand, the Infopaq- and Meltwater-decisions of the European Court of Justice, while acknowledging the possibility of copyright for a sentence, accepts that the temporary copy-exception is applicable for copying an article in the process of analysis for indexing.[260] We have discussed copyright in depth above and will not repeat that discussion, but it illustrates the difficulties which can arise with data collection by search engines.
Similar problems did arise concerning privacy. In the recent case before the ECJ Google also based its claim on the safe harbor-clause in the E-commerce directive in the EU. Therefore it is not a data controller but only a data processor according to the EU data protection directive, and it is the original publisher which is responsible. In this case the European Court of Justice did not accept Google's claim, considered Google as a data controller and the functioning of the search engine subjected to the data protection framework. Google had therefore to implement the right to be forgotten and to remove links where the privacy interests of the data subject outweighs the public interest in the publication[261].

For user-generated content Google also works with a license. In general users do not have to click to agree before using Google Search or Google Maps. The Terms state that "By using our Services, you are agreeing to these terms." and "Additional terms will be available with the relevant Services, and those additional terms become part of your agreement with us if you use those Services".[262] But it is doubtful that courts will accept the use of the service as acceptance of contractual obligations. In this case the Google Terms of Service are not a contract and no obligations can be pushed upon the user other than those contained in the legislation applicable by default.
This changes when users make an account and accept the terms. But still it is doubtful that clicking for acceptance during that process implies a valid acceptance for all the terms applying over the whole range of Google services. It will probably be considered as an adhesion contract, where only the main elements of the transaction are considered to be accepted.
The Terms of Service are a mixture of a service license and a copyright license. Both have different legal bases which protect different exclusive rights, and the license is in each case an authorization to perform otherwise forbidden actions. The service license allows the use of the services. If this would otherwise be forbidden is questionable, but this is a valid authorization in case it is by cybercrime legislation or other. The Google Maps/Earth Additional Terms of Service state: "Use of the Products. Google grants you a non-exclusive, non-transferable license to access the Google Maps service, to download and use the Google Earth software and service, and to access the Content (as defined below) within the Products and according to the Terms." This license text gives the permission to use the service and access the content. Access is not one of the rights protected by copyright, so this part does not concern copyright but only the use of the service. Similarly, the Google Maps/Earth

[259] Xalabarder, Raquel, "Google News and Copyright", and Miquel Peguera, "Copyright Issues Regarding Google Images and Google Cache" in Aurelio Lopez-Tarruella (ed.), *Google and the Law. Empirical Approaches to Legal Aspects of Knowledge-Economy Business Models*, T.M.C. Asser Press, The Hague, 2012.
[260] EUCJ, C-5/08, *Infopaq International A/S v Danske Dagblades Forening*, 16 July 2009; EUCJ, C-360/13, *Public Relations Consultants Association Ltd v. Newspaper Licensing Agency Ltd and Others*, 5 June 2014 (aka the Meltwater decision).
[261] EUCJ, *Google Spain SL, Google Inc. v Agencia Española de Protección de Datos (AEPD), Mario Costeja González*, C-131/12, 13 May 2014, §28.
[262] Google, *Terms of Service*, version April 30, 2014, https://www.google.be/intl/en/policies/terms/regional.html

Additional Terms of Service state further "Content in the Products. Google Maps and Google Earth allow you to access and view a variety of content, including but not limited to photographic imagery, map and terrain data, business listings, reviews, traffic, and other related information provided by Google, its licensors, and its users (the "Content")."[263]

Copyright is covered by other parts of the license texts. The Terms of Service state: "Using our Services does not give you ownership of any intellectual property rights in our Services or the content you access. You may not use content from our Services unless you obtain permission from its owner or are otherwise permitted by law". The meaning of use in this last sentence is very unclear, but clearly includes the uses covered by the exclusive rights provided by copyright.
The Google Maps/Earth Additional Terms of Service state: "Restrictions on Use. Unless you have received prior written authorization from Google (or, as applicable, from the provider of particular Content), you must not: (a) copy, translate, modify, or make derivative works of the Content or any part thereof; (b) redistribute, sublicense, rent, publish, sell, assign, lease, market, transfer, or otherwise make the Products or Content available to third parties; ... (d) use the Products in a manner that gives you or any other person access to mass downloads or bulk feeds of any Content, including but not limited to numerical latitude or longitude coordinates, imagery, and visible map data; … (g) use the Products to create a database of places or other local listings information."
Both texts make clear that for ordinary users no authorization is given for uses covered by the exclusive rights under copyright. As a lot of the material in Google Maps is provided by TomTom, a producer of GPS guiding systems this part also specifically forbids to use the service or content for real time navigation or route guidance.

Concerning user-generated content the Terms state: "Some of our Services allow you to upload, submit, store, send or receive content. You retain ownership of any intellectual property rights that you hold in that content. In short, what belongs to you stays yours.
When you upload, submit, store, send or receive content to or through our Services, you give Google (and those we work with) a worldwide license to use, host, store, reproduce, modify, create derivative works (such as those resulting from translations, adaptations or other changes we make so that your content works better with our Services), communicate, publish, publicly perform, publicly display and distribute such content. The rights you grant in this license are for the limited purpose of operating, promoting, and improving our Services, and to develop new ones."
Similarly to the Twitter terms a user retains the copyright over his contributions, but gives a wide authorization to Google to use the contribution.

Google also collects a wide range of personal data and states in its terms: "By using our Services, you agree that Google can use such data in accordance with our privacy policies". Again, it is doubtful that use of services by ordinary users constitutes acceptance. Who signs up for an account ticks a box for agreeing. In its Privacy Policy Google explains the wide range of personal data it collects during the use of its services and how it uses it. It also explains how this information can be accessed and corrected. The Privacy Policy states that personal information from several service is combined, but that DoubleClick cookie information (present through Google-delivered advertisements on other websites) is not

[263] Google, *Google Maps/Earth Additional Terms of Service*, version 1 March 2012. https://www.google.com/intl/en_ALL/help/terms_maps.html

combined with personally identifiable information unless the user opts in.[264] On the other hand, avoiding data collection through the DoubleClick cookie or Google Analytics, both of which are very widespread, needs opting out through browser add-ons.

So far we see a wide data collection, but the use of this data is limited to accessing it through the Google services. A wider access is provided through the APIs, which are in general free for limited use but are available on a larger scale commercially. Google has a wide range of APIs available for its services. Although some automated access is possible without registering, in general registering and obtaining an API key is necessary. This process involves ticking a box for agreeing to the Terms and implies by consequence a contractual relation. The specific Google Maps/Google Earth APIs Terms of Service mentions both clicking to accept or agree to terms and using the API as a method to agree to the Terms.[265]
The use of these APIs is subjected to the terms of the service to which it gives access (which includes the general terms and privacy policy), general API terms and service-specific API terms.

Important is also that access to the content is only allowed through the APIs and not through other methods.[266] Also Google does not allow to build other implementations of its services or its APIs above its own or on the data it manages. For instance, the Google Maps/Google Earth APIs Terms of Service clearly forbids wrapping its service or content into another: "No "Wrapping." You must not create or offer a "wrapper" for the Service, unless you obtain Google's written consent to do so. For example, you are not permitted to: (i) use or provide any part of the Service or Content (such as map imagery, geocoding, directions, places, or terrain data) in an API that you offer to others; or (ii) create a Maps API Implementation that reimplements or duplicates Google Maps/Google Earth. For clarity, you are not "re-implementing or duplicating" Google Maps/Google Earth if your Maps API Implementation provides substantial additional features or content beyond Google Maps/Google Earth, and those additional features or content constitute the primary defining characteristic of your Maps API Implementation."[267]
The general API terms also make clear that large-scale access is not allowed except with a separate permission: "Unless expressly permitted by the content owner or by applicable law, you agree that you will not, and will not permit your end users to, do the following with content returned from the APIs:
1. Scrape, build databases or otherwise create permanent copies of such content, or keep cached copies longer than permitted by the cache header;
2. Copy, translate, modify, create a derivative work of, sell, lease, lend, convey, distribute, publicly display or sublicense to any third party;".[268]

Restrictions to access do exist, but mostly to the free use. Entering into business agreement and paying for the service allows for wider use. The Google Maps/Google Earth APIs Terms of Service makes clear that the use of the service is subjected to a "limit on the number of transactions you may send or receive through the Service".[269] Content delivered through this

[264] Google, *Privacy Policy*, version 31 March 2014.
https://www.google.com/intl/en_ALL/policies/privacy/#access
[265] Google, *Google Maps/Google Earth APIs Terms of Service*, version 25 November 2013, art. 2.1.
https://developers.google.com/maps/terms
[266] Google, *Google Maps/Google Earth APIs Terms of Service*, version 25 November 2013, art. 10.1.1(a)
[267] Google, *Google Maps/Google Earth APIs Terms of Service*, version 25 November 2013, art. 10.2
[268] Google, *Google APIs Terms of Service*, version December 9, 2011, https://developers.google.com/terms/
[269] Google, *Google Maps/Google Earth APIs Terms of Service*, version 25 November 2013, art. 4.2,
https://developers.google.com/maps/terms

channel can contain advertisements. Who wants higher limits or to opt out from advertisement can enter into a commercial relation, for which the Google Maps API for Business - Purchase Agreement applies.[270]
The Google Maps/Google Earth APIs Terms of Service makes clear it is both a service license ("to use the Service") and a content license through which the authorization is given to exercise otherwise protected rights ("to access, use, publicly perform and publicly display the Content in your Maps API Implementation, as the Content is provided in the Service, and in the manner permitted by the Terms.")[271] Free use of the API is conditional on the requirement that the Maps API implementation is freely and publicly accessible, otherwise a commercial agreement is needed.

The use of other APIs is regulated in quite similar ways. In general we can conclude that Google offers a much wider access to its services and the big data assets it controls than we noticed with Twitter. Basic reason is the different business model from Twitter. Twitter offers only one service through which it gathers information and had difficulties finding a good business model. It oriented itself to targeted advertisements to its users and analysis of the communication for marketing and other purposes. Therefore it guards and constrains much more the access and uses of the data assets it manages.
Google on the contrary developed its business model of targeted advertising through attracting a wide range of unregistered users to its services. Starting with Gmail it also developed the Google account which is necessary to use some of its services, but it continues to track other unregistered users as well through cookies. Google's business model is dependent on a further spread of its services which allows it to track users along their internet use. Therefore its interest lays in developing services and allowing the use of it by and in as many other web services as possible. It will rather control that its services are not circumvented to access the data or in a way that its business model of user tracking and targeted advertising is circumvented. Result is a much more open ecology around its big data services, aimed at widening its tracking and data collection capacity.

6.2 Open content Licenses

We have just seen how large big data companies shape their data ecosystem across borders and on a global level. Other actors try to do something similar, but not to shape their enclosure but to open their data. They do this with open content licenses, an instrument based on intellectual property law but aimed at waiving the protection offered by this law.
Also these licenses get confronted with similar problems as the global corporate players concerning interoperability across different jurisdictions. The fragmentation due to the national character of intellectual property law, and also contract law, presents difficulties to make a standard license text which functions across all jurisdictions. Also, open content licenses are not always interoperable among each other. Mixing content released under one open content license with content released under another, leaves the question under which license the end result has to be released and if such combined derivative use is not always violating the terms of one of the licenses. The unintended result can be that open content licenses each create a data enclosure of their own. Efforts are made by developers of open content licenses to address these problems, but obstacles remain. Till now open content

[270] Google, *Google Maps API for Business - Purchase Agreement*, https://www.google.com/enterprise/earthmaps/legal/us/maps_purchase_agreement.html
[271] Google, *Google Maps/Google Earth APIs Terms of Service*, version November 25, 2013, art. 8.2 and 8.3; similar in Google Maps API for Business - Purchase Agreement, art. 1.1

licenses have been seldom tested in courts, but further practice will have to show if the current licenses are able to deal with the use cases emerging in the data economy.
Fundamentally open content licenses are patchwork, which can barely fix the problems created by an unadapted intellectual property framework. This framework provides too strong protections, leading to strict enclosures, and contains exceptions which were functional in the paper-era but cannot adequately deal with practices in the digital era. Creative Commons, one of the main developers of open content licenses, makes a plea for copyright reform instead: "CC licenses are a patch, not a fix, for the problems of the copyright system. … Our experience has reinforced our belief that to ensure the maximum benefits to both culture and the economy in this digital age, the scope and shape of copyright law need to be reviewed. However well-crafted a public licensing model may be, it can never fully achieve what a change in the law would do, which means that law reform remains a pressing topic. The public would benefit from more extensive rights to use the full body of human culture and knowledge for the public benefit. CC licenses are not a substitute for users' rights, and CC supports ongoing efforts to reform copyright law to strengthen users' rights and expand the public domain".[272]

A wide range of open licenses exists and here we will review the Creative Commons v.4 International Public licenses as one example very commonly used and applicable to data and databases. Other licenses exist, like the Open Database License (OdbL), which concerns only database rights and where the contents needs to be licenses separately.
First question is what are the rights held on the dataset. A license is an authorization to use data under certain conditions, where such authorization to that specific use is exclusively reserved for the right holder. Only when rights are held, there is something to authorize. Otherwise, the license does not apply and no conditions can be linked to that use.
When rights are held which can be licensed, the Creative Commons licenses provide a method to authorize use. Open content licenses try to remain copyright licenses, that is unilateral permissions, and avoid to become contractual obligations, which have to be accepted by the licensee and which are only valid between the parties. As this is not always possible and equally applicable across jurisdictions, the CC licenses use contractual law as fallback option: "By exercising the Licensed Rights (defined below), You accept and agree to be bound by the terms and conditions of this Creative Commons Attribution 4.0 International Public License ("Public License"). To the extent this Public License may be interpreted as a contract, You are granted the Licensed Rights in consideration of Your acceptance of these terms and conditions, and the Licensor grants You such rights in consideration of benefits the Licensor receives from making the Licensed Material available under these terms and conditions."[273]
In general open content licenses contain the following elements: the parties (Licensor-Licensee), the license issuer (e.g. Creative Commons) and license application, general license features, license grant, conditions (e.g. attribution), a disclaimer and limitation of liability, term and termination, and the governing law and competent courts.[274]
The general license features define the scope and key elements of the license, the license grant defines the types of rights granted and the subject matter of the license. The Creative Commons-licenses state in section 2:
"a. License grant.

[272] Creative Commons, "Creative Commons and Copyright Reform", 16 October 2013. http://creativecommons.org/about/reform
[273] Creative Commons, *Creative Commons Attribution 4.0 International Public License*, no date. http://creativecommons.org/licenses/by/4.0/legalcode
[274] Prodromos Tsiavos, *Licence Interoperability Report*, LAPSI 2.0, no date, pp. 18-19.

1. Subject to the terms and conditions of this Public License, the Licensor hereby grants You a worldwide, royalty-free, non-sublicensable, non-exclusive, irrevocable license to exercise the Licensed Rights in the Licensed Material to:

A. reproduce and Share the Licensed Material, in whole or in part; and

B. produce, reproduce, and Share Adapted Material."[275]

The license is worldwide in order to function in all jurisdictions, royalty-free so without charge, and non-exclusive so open to everyone who fulfills the requirements. It is irrevocable, which means that when the licensor withdraws the license and republishes under a much more restrictive license the old license remains applicable for the older re-use. Further the license is non-sublicensable, meaning that the licensee cannot re-license the same material under another license. This is further clarified with "A. Offer from the Licensor – Licensed Material. Every recipient of the Licensed Material automatically receives an offer from the Licensor to exercise the Licensed Rights under the terms and conditions of this Public License."[276] In other words, when further disseminated it is the original right holder which grants a license.

The license concerns copyright and similar rights. In a separate section the license also authorizes the protected uses under sui generis database rights. The license does not apply to patent and trademark rights. Moral rights, privacy rights and other personality rights are not licensed, but where possible waived to the extent to allow the exercise of the licensed rights.

The variation between the licenses can be found in the conditions. The Creative Commons set of license contains the following conditions:

- Attribution (BY): This implies that shared material (also derivative work) has to contain an attribution to the creator.
- ShareAlike (SA): sharing of material has to be done under the same license, also derivative work.
- NoDerivatives (ND): sharing is allowed but only of identical works (which implies also the same license). Adaptations are not allowed.
- NonCommercial (NC): Sharing is allowed only for non-commercial use, also of derivative works.

The CC licenses always contain the attribution condition, and can be combined with the other optional conditions. This leads to 6 versions: BY, BY-ND, BY-SA, BY-NC, BY-NC-SA, BY-NC-ND, which can be found on http://creativecommons.org/licenses/.

Creative Commons also offers a 7th option, the CC0 1.0 Universal-license or Public Domain Dedication.[277] In this license the licensor certifies that the work is in the public domain and freely available, or, when he is the copyright holder, waives all rights to the maximum extent possible, including attribution. This CC0-license functions to provide legal certainty about the public domain status, but can also be important to ensure that data is effectively open and not enclosed in an enclosure defined by the type of license.

Even attribution-only licenses can create such enclosures when they include specific rules for such attribution. The same can happen when licenses with similar performance include specific rules concerning the reference to the license (like publishing with the license text). Such conditions effectively prevent the interoperability of open content licenses, and the combined used of datasets covered by such different open content licenses. The rise of governmental open data policies is often accompanied with the publication of their own

[275] Creative Commons, *Creative Commons Attribution 4.0 International Public License*, section 2, §a.1

[276] Creative Commons, *Creative Commons Attribution 4.0 International Public License*, section 2, §a.5

[277] Creative Commons, *CC0 1.0 Universal*, http://creativecommons.org/publicdomain/zero/1.0/legalcode

governmental open data licenses. These licenses risk similar, unintended problems of legal interoperability.

6.3 CONCLUSIONS

Private actors can shape their own access and use policies concerning the data they control. On a legal level these actors use the private law tools of IPR and contract law. Access and use gets controlled with terms of use and privacy policies.
In this chapter we first compared the terms of 2 important global actors: Twitter and Google. Both are major big data companies but with very different business models. A detailed analysis of their terms reveals how both shape their own distinct and partially open, partially closed data ecosystem in a way favorable to their business model. These business models define their platform politics or the organizational interoperability they allow with their services. Through their terms they define how the data under their control can be linked up to provide other services.
In general we can conclude that Google offers a much wider access to its services and the big data assets it controls than we noticed with Twitter. Twitter offers only one service through which it gathers information and therefore constrains much more the access and uses of the data assets it manages. Google on the contrary developed its business model of targeted advertising through attracting a wide range of users to its services. Google's business model is dependent on a further spread of its services which allows it to track users along their internet use. This results is a much more open ecology around its big data services, aimed at widening its tracking and data collection capacity.

The same private law tools are used for making open content licenses. These licenses do not aim to shape data enclosures but to open their data. They also use intellectual property law but to waive the protection offered by this law. The most widely used is the Creative Commons set of licenses, which is becoming a standard and of which the most recent version also is usable for data. Several of the Open Government Licenses we saw earlier try to achieve compatibility with this set.
Open content licenses are instrumental in shaping a much more open data ecosystem. However, fundamentally they attempt to fix the problems created by an unadapted intellectual property framework, which through too strong protections leads to enclosures. These licenses get confronted with similar problems as the global corporate players concerning interoperability across different jurisdictions. This fragmentation presents difficulties to make a standard license text which functions across all jurisdictions. Also, open content licenses are not always interoperable among each other. The unintended result can be that open content licenses each create a data enclosure of their own.

These two opposite approaches clearly demonstrate how big data processing is dependent on legal interoperability of data sources and how the legal framework can be used to define the data ecosystem in which such big data processing can take place. It also demonstrates how this is linked with organizational interoperability. Big data companies use the legal tools to allow uses fitting their business model while locking other out. Open data policies use open content licenses to build a much wider ecosystem and allow a wide range of business processes, in order to support positive network effects from diverse data sources.

7 DATA POLICIES IN OTHER COUNTRIES

In this last part we also want to look outside the EU and US jurisdiction. Therefore we add a short review of other countries outside this realm: China, Japan and Australia. Australia as an example which remains near to the common law tradition, while China and Japan are important industrial countries with a different legal tradition.

7.1 AUSTRALIA

Australia recently devoted quite some attention to big data, as part of governmental ICT operations, and published a Big Data Strategy. Focus is on making public services more efficient and on improving the exploitation of the data assets hold by the government. Through the development of an open data policy also private sector uses of PSI came into focus. The private sector and big data were addressed in the National Digital Economy Strategy.
But in general the legal framework surrounding data processing has not been changed or reviewed in the context of big data, although the advent of the digital age has led to changes in privacy law.
First we will look into the legal framework concerning IPR and data protection. Then we will discuss the big data and open data policies of the Australian government.

IPR

The Australian IPR framework complies with the international framework in the WIPO and TRIPS treaties. Its Copyright Act 1968[278] has been updated to function in the digital economy where copyrighted works are in large part shifting to digital carriers and formats. But the data economy has received no specific attention. Databases do not receive a specific treatment in the Copyright Act, but are protected as compilations and considered with application of the general criteria in this Act. Although early court decisions seemed to recognize a 'sweat of the brow' doctrine granting protection to work and investment, more recent decisions made application of an originality requirement. To be original a data compilation must have an identifiable human author, it may not be copied and the compilation has to be the result of independent intellectual effort. This effort had to linked with the expression in the compilation itself and not with the content.[279] A sui generis protection of databases does not exist in Australian law.
The Copyright Act contains a series of exceptions to the exclusive use listed in the Act as fair dealings, but no specific exception for text and data mining. During a public consultation by the Australian Law Reform Commission (ALRC) some suggest that text and data mining are covered by the exception for 'fair dealing for purpose of research or study' and for 'temporary reproductions of works as part of a technical process of use'[280], but the ALRC considers this as unlikely[281]. The ALRC recommends adopting a general 'fair use'-exception like in the US

[278] Australian government, Copyright Act 1968, ComLaw, http://www.comlaw.gov.au/Details/C2014C00291
[279] Fitzgerald, Anne M. & Dwyer, Natasha, "Copyright in databases in Australia", 2012. http://eprints.qut.edu.au/50425/4/50425.pdf; Mark Vincent and Katrina Crooks, "Australia: Can a database be protected by copyright?", 7 February 2014. http://www.mondaq.com/australia/x/290668/Copyright/Can+a+database+be+protected+by+copyright
[280] Australian Law Reform Commission, *Copyright and the Digital Economy* (DP 79), 31 May 2013, §8.53, p. 166.
[281] Australian Law Reform Commission, *Copyright and the Digital Economy. Final Report*, ALRC Report 122, 30 November 2013, §11.65, p. 262.

as a flexible and technology-neutral solution.[282] This recent report has not led yet to legislative action.

Data Protection

The Privacy Act 1988 originally focused on the public sector but has been gradually broadened to the private sector. Exempted remain small business, defined as a business with an annual turnover of $3,000,000 or less. Most important exceptions for which this small business exemption does not apply are for credit reporting bodies or for businesses which provide a health service and hold health information, businesses which disclose personal information about another individual to anyone else for a benefit, service or advantage, or which provide a benefit, service or advantage to collect personal information about another individual from anyone else.[283]
The Privacy Act has recently been subject of a major review with changes coming into force on 12 March 2014. Personal information is defined very wide as information "about an identified individual, or an individual who is reasonably identifiable"[284], which is comparable to EU law. The Privacy Act contains a list of 13 Australian Privacy Principles (APP), which are very similar to the principles in EU data protection or the fair information principles. Principles like finality and purpose limitation, consent as a base for legitimate processing, transparency and data subject rights of access and correction are present in these APPs. Apart from minor differences the Australian Privacy Act will lead to similar constraints for big data processing as the European data protection framework.

Big data strategy and open data policy

The Australian government published in August 2013 a Big Data Strategy. This strategy focused on big data processing for governmental purposes. It was a follow-up from the Australian Public Service ICT Strategy 2012 – 2015[285], which identified 3 priority areas: the delivery of better government services, improving the efficiency of government operations and supporting open engagement. The open engagement-strand included action items on big data, such as the development of this big data strategy and of a governmental Centre of Excellence, and on open data, focusing on releasing more data on data.gov.au and encouraging its use. This strategy also figured as action point in the 2013 update of the National Digital Economy Strategy, together with a focus on open data.[286] So both strategies provided for similar action points in this area.

The Big Data Strategy puts forward some guiding principles for agencies in their approach to big data:

- Principle 1: Data is a National asset

[282] Australian Law Reform Commission, ibid., p. 13.
[283] Australian government, Privacy Act 1988, ComLaw, article 6D.
http://www.comlaw.gov.au/Details/C2014C00076
[284] Australian government, Privacy Act 1988, ComLaw, article 6.
http://www.comlaw.gov.au/Details/C2014C00076,
[285] Australian government, Department of Finance and Deregulation, *Australian Public Service Information and Communications Technology Strategy 2012-2015*, October 2012.
http://www.finance.gov.au/files/2013/01/APS_ICT_Strategy.pdf
[286] Australian Government, Department of Broadband, Communications and the Digital Economy, *Advancing Australia as a Digital Economy: An Update to the National Digital Economy Strategy*, 12 June 2013.
http://apo.org.au/files/Resource/Advancing-Australia-as-a-Digital-Economy-BOOK-WEB.pdf

This means that data is to be used for public good and is not just to be used for the department holding it but made available for the government as a whole and, if possible, the public.

- Principle 2: Privacy by design
- Principle 3: Data integrity and the transparency of processes

This includes that the development of big data projects have to use peer review, public consultation, Privacy Impact Assessments (PIA). All parties involved have to be aware of their responsibilities concerning providing source data, maintaining control of personal and sensitive data and the management of the project.

- Principle 4: Skills, resources and capabilities will be shared

This concerns both among government agencies, as well as with industry.

- Principle 5: Collaboration with industry and academia
- Principle 6: Enhancing open data

The strategy contained 6 actions, mostly concerning developing knowledge and guidance concerning big data analytics in practice, including the privacy and security aspects. Further information asset registers would be developed. In general we can conclude that this strategy shows that the adoption of big data is still in a very early stage and the focus is on the development of skills and knowledge.
Big data analytics capability will be further strengthened through the Data Analytics Centre of Excellence (DACoE), as well as setting up pilot projects.

Open data and the re-use of PSI was raised as part of an Open Government Agenda with the 2009 report of the Government 2.0 Taskforce, which contained a chapter on PSI[287], and turned into policy with the Declaration of Open Government[288] in 2010. This policy got more body with the publication of the Principles on open public sector information developed by the Office of the Australian Information Commissioner (OAIC), which provide the main guidelines for opening PSI:

- Principle 1: Open access to information - a default position

This implies a presumption of openness. When there is no legal need to keep the information protected, access should be granted.

- Principle 2: Engaging the community

This implies that public consultation is part of the decision making process about releasing data.

- Principle 3: Effective information governance

This means that agencies have to make information management a core responsibility and have to develop a clear policy.

- Principle 4: Robust information asset management

This implies maintaining a register of the agency's information and develop clear responsibilities and procedures for decision making about publication, as well as ensuring proper archival practices and security against inappropriate use or disclosure.

- Principle 5: Discoverable and useable information

This involves publishing an information asset register, publishing in open, standards-based and machine-readable formats and attaching of high quality metadata.

- Principle 6: Clear reuse rights

[287] Government 2.0 Taskforce, *Engage: Getting on with Government 2.0*, 22 December 2009, ch.5. http://www.finance.gov.au/publications/gov20taskforcereport/chapter5.htm
[288] Australian Government, Department of Finance, *Declaration of Open Government*, 16 July 2010. http://www.finance.gov.au/blog/2010/07/16/declaration-open-government

The Australian government decided to use the Creative Commons BY license as default license.[289]

- Principle 7: Appropriate charging for access
- Principle 8: Transparent enquiry and complaints processes[290]

Conclusion

Australia is well-advanced in big data-related issues. It has a well-developed big data-policy, mostly focusing on governmental uses but through an open data-policy it also engages the private sector and the general public. Open data policy is still in a start-up phase but a framework with stakeholder involvement is set up.
Legal frameworks concerning IPR and privacy are similar to those in the EU, especially with the latest update of the Privacy Act. Possible challenges that big data poses for these frameworks has been the subject of political debate, but has not yet led to changes.

7.2 CHINA

Introduction

Big data may seem like a rather new topic in China, but it is not. The first book on the topic, Big Data Revolution written by Tu Zipei appears in 2012. It mainly analyses the development of Big Data and Open Data in the USA and calls for a similar trend in China. However, Big data are of wide interest in both the public and the private sector, and China has the largest Web 2.0 industry after the US, but ahead of any other country.
As soon as 2012, Wang Yang (one of China's Vice-Premiers) quotes Tu Zipei's book and asserts that governmental data should be openly available. From the Chinese government's perspective, Information technologies and Big Data are important political tools. Several public institutions have recently organized conferences on the topic. In 2013, for instance, the Chinese ministry of industry and information technologies, MIIT, organized a national conference on Big data, while Fudan university hold an International symposium on the future of e-governance. Chinese cities and governmental institutions tend to open their data and foster their use. Fudan University, for instance, organized the China Computer Federation's (CCF) Young Computer Scientists & Engineers Forum in 2013. CCF also holds an annual conference on Big Data.[291]
In the private sector, companies such as Govmade[292] focus on Big data and Open data analysis. Major Chinese companies also invest in Big Data. Jack Ma (Alibaba's CEO) now holds 99.1% of Heng Sheng Electronics Technologies Co Ltd (a company which has access to a large amount of financial data). Tencent has developed applications such as QQ and

[289] Attorney-General's Department, Commonwealth of Australia, "Statement of Intellectual Property Principles for Australian Government Agencies", 1 October 2010. www.ag.gov.au/RightsAndProtections/IntellectualProperty/Documents/StatementofIPprinciplesforAusGovagencies.pdf; Attorney-General's Department, Commonwealth of Australia, *Guidelines for Licensing Public Sector Information (PSI) for Australian Government Agencies*, February 2012. www.ag.gov.au/RightsAndProtections/IntellectualProperty/Documents/GuidelinesforlicensingPSIforAusGovagencies.doc
[290] Office of the Australian Information Commissioner, *Principles on open public sector information: Report on review and development of principles*, May 2011. http://www.oaic.gov.au/information-policy/information-policy-resources/information-policy-reports/principles-on-open-public-sector-information-report-on-review-and-development-of-principles-may-2011
[291] http://bdtc2013.hadooper.cn/sponsors_en.html
[292] http://www.govmade.cn/english/index.html

wechat (instant messaging) or Weibo and Qzone (social networks).[293] It now also develops applications focusing on data analysis such as Mind 3.0 (a marketing tool for Tencent's platforms). Finally, the Li Ka Shing Foundation has given Stanford University US$ 3 million to analyse large medical datasets.[294]

Open Data (政府数据公开)[295]

Over the last years, China has taken steps towards Open Data. China has been a member of Open Knowledge Foundation (OKF), a non-profit international organization, which promotes Open Data since 2013.[296] Out of 70 countries, the OKF ranks China at the 33th place.[297] Datasets such as the government's budget and the elections results are yet to be open. As for the open datasets (legislation, national statistics, etc.), they are not openly licensed. Chinese cities are also involved in Open Data. Between 2012 and 2014, Beijing, Shanghai, Qingdao and Guangzhou for instance have launched Open Data portals.[298]
Even if the open datasets may look limited or difficult to process, they led to the development of several initiatives. The government takes part in sponsoring "Hackatons".[299] Private companies, such as Alibaba, support open data processing in order to build "danger maps" (maps which present polluted or dangerous areas in China).[300] Eventually, individuals encourage the development of Open Data and their use.[301]

Personal data, definition and protection[302]

There is no Chinese national law on personal data protection. However the Information Security Technology Guidelines for Personal Information Protection (信息安全技术 个人信息保护指南) within Public and Commercial Services Information Systems (Guideline) is a national standard dealing with personal information. The Guideline is not binding but remains the most precise description of data protection principles in China. The Guideline is enforced by the Ministry of Industry and Information Technology.
Besides the Guideline, several laws provide principles about personal data protection. For instance, the General Principles of the civil law of the PRC (Article 120) protects a citizen's personal name, portrait, reputation and honour. The Provisions on Protecting the Personal Information of Telecommunications and Internet Users regulates the collection and use of the personal information of telecommunications and internet users.

---

[293] http://5loom.com/what-we-think/tencents-big-data-plan/
[294] The China Post, "Business tycoons leading the way as big data firms make big impression in China", 22 April 2014. http://www.chinapost.com.tw/business/asia-china/2014/04/22/405905/Business-tycoons.htm
[295] Open Knowledge Foundation, Open Data China Timeline, no date. http://timemapper.okfnlabs.org/okfncn/open-data-china-timeline
[296] Open Knowledge Foundation, "Open Knowledge China", no date. http://okfn.org/open-knowledge-foundation-china
[297] Open Knowledge Foundation, "Open Data Index", no date. https://index.okfn.org/country
[298] The government in China, as well as many local governments have started to make data accessible. For instance Shanghai's government at http://www.datashanghai.gov.cn.
[299] Rebecca Chao, "The Hunt for Open Data in China", *Techpresident*, no date. http://techpresident.com/news/wegov/24332/hunt-open-data-china
[300] http://www.weixianditu.com
[301] Rebecca Chao, "In China, An Open Data Movement is Starting to Take Off", *Techpresident*, no date. http://techpresident.com/news/wegov/24940/China-Open-Data-Movement-Starting-Take-Off
[302] Marissa Xiao Dong, "Data protection in China: overview", *Practical Law*,no date. http://uk.practicallaw.com/4-519-9017; Ministry of Industry and Information Technology, http://www.miit.gov.cn/n11293472/n11293832/n11293907/n11368223/13590447.html

Personal information under the Guideline means computer data capable of being processed by an information system that is relevant to a certain natural person and that may be used solely or along with other information to identify that natural person. Personal information can be divided into sensitive personal information and general personal information. The Postal Law guarantees the protection of freedom and privacy of correspondence and the safety of email. The Provisions on Protecting the Personal Information of Telecommunications and Internet Users regulates the collection and use of the personal information of telecommunications and internet users. Each law may be enforced by a specific entity.

Several constraints apply on data collection and processing. Data subjects must consent to the collection and processing of their data. The sensitive personal information of minors under 16 years old and any persons with limited or no civil capability must not be directly collected. Collection of such data requires the consent of legal tutors. Data subjects can object to data collection or processing. They have a right to request the deletion of their data.
The data controller must notify the data subjects before processing data. The notification must contain the following items:

- The purpose of handling the personal information.
- The manners and means of personal information collection, the specific contents to be collected, and the time/duration of retention.
- The scope of use of the collected personal information, including the scope of disclosure or provision of personal information to other organizations and institutions.
- The measures for protecting personal information.
- The name, address, contact information and other relevant information of the personal information administrator.
- The risks that the data subjects may encounter after providing personal information.
- The consequences if the data subjects are not willing to provide personal information.
- The channel for the data subjects to file a complaint; and in circumstances where personal information needs to be transmitted or entrusted to another organisation, data subjects must be expressly notified with information that includes but is not limited to the:
    - purpose for transmission or entrustment;
    - specific contents and scope of use of the transmitted or entrusted personal information;
    - name, address, and contact information of the receiver of the entrusted personal information.

The data holder must adopt informed methods and means to process the personal data.
Eventually, the data controller is responsible for ensuring the security of data and guaranty that they are in a up to date state.

Conclusions

China has its own commercial big data players. Specific policy, including on open data, remains in its start-up phase. The relevant legal frameworks are rather weak, which can give space for big data projects.

7.3 JAPAN

Introduction

Big data is a major economic topic for the Japanese government. The 2013 White paper on information and communication states that developing Big data in Japan could create large benefits.[303] Japan has also developed strong industries in the Web 2.0 sector.
However, the debate around big data revolves as well around privacy protection. As private actors want to analyse big data, Japan plans on setting a means to certify big data users in order to leverage privacy concerns.[304]
Big data attracts the interest of public sector (the government has recently decided to launch an economic indicator based on big data)[305] and private sector. For instance, Toyota collects data from its cars and plans on building services on top of them. Such services could include helping drivers to handle traffic difficulties. Toyota plans to make some of these data freely available. The Nippon Telegraph and Telephone Data corporation (NTT) also offers big data services[306] and leads several projects in the field[307].

IPR

Japan has updated its copyright law with a new exception giving space for information analysis. It has excluded from this exception databases specially aimed for information analysis. This sounds contradictory, but is similar to the exception often made for educational purposes. Works specifically meant for use in schools cannot use this exception, in order to protect academic publishers. Similarly, the exception is not meant to discourage development of databases.[308]

Open Data

The OKF ranks Japan at the 27th place.[309] Most open data is not openly licensed.[310] The Open Government Data Strategy has been adopted in 2012 by Japan as a means to foster transparency.[311] The portal http://www.data.go.jp gathers all the available data. Before 2012, the Ministry of Trade and Industry launched its own open data portal (http://datameti.go.jp/).

---

[303] Ministry of Internal Affairs and Communications, Information and Communications in Japan 2013 (summary) Technology Policy, http://www.soumu.go.jp/main_sosiki/joho_tsusin/eng/whitepaper.html
[304] Nikkei Asia Review, "Japan to certify big data users to ease privacy concerns", 8 March 2014. http://asia.nikkei.com/Politics-Economy/Policy-Politics/Japan-to-certify-big-data-users-to-ease-privacy-concerns
[305] Mitsuru Obe, "Japan Looks to Big Data for Timely Economic Indicator", Wall Street Journal, 24 September 2013, http://blogs.wsj.com/japanrealtime/2013/09/24/japan-looks-to-big-data-for-timely-economic-indicator
[306] NTT Data, Big Data Solutions, http://www.nttdata.com/global/en/services/bds/index.html
[307] St. John, Jeff, "AutoGrid Lands NTT Data as Big Data Energy Partner", GreenTechGrid, 22 August 2013. http://www.greentechmedia.com/articles/read/autogrid-lands-ntt-data-as-big-data-energy-partner
[308] Triaille, Jean-Paul, Jérôme de Meeûs d'Argenteuil and Amélie de Francquen, *Study on the legal framework of text and data mining (TDM)*, March 2014, pp. 10-11.
[309] Open Knowledge Foundation, "Open Data Index", no date. https://index.okfn.org/country
[310] Open Knowledge Foundation, "Open Data Index – Countries/Japan", no date. https://index.okfn.org/country/overview/Japan
[311] Prime Minister of Japan and His Cabinet, "Summary of the Open Government Data Strategy ", 4 July 2012. http://japan.kantei.go.jp/policy/it/20120704/sum.pdf

The NTT data plays a major role in standardizing and fostering the interoperability of open data.
The use of open data is still quite new. For instance, it is centred around earthquakes analysis. The Sinsai portal encourages citizens to share data about earthquakes (http://www.sinsai.info/). Hack for Japan (https://sites.google.com/site/hackforjapan/) is a developer community which aims at using their technical skills to help recovery from natural disasters. Eventually, there exist several hackatons such as the Linked Open data challenge[312].

Personal data[313]

The Act on the Protection of Personal Information (Act No. 57 of 2003) (APPI) is the main applicable law on data protection. It applies to business operators (persons or entities which handle databases with personal information about more than 5000 people). There also exist sectorial guidelines.
The Consumer Affairs Agency is main regulator in terms of data protection.

The Act regulates data, which permits to identify individuals and are stored in databases. The Act draws a difference between data stored for less than six months and "retained personal data". The act applies to any data processing (access, storage or alteration). When data are used for news reports, literary, academic, political or religious activities performed by the relevant actors, it is not regulated.
Some guidelines define "sensitive information" (such as religious or health information).

A business operator handling personal information (BOPI) must notify the data subject of the purpose of data collection. This restriction also applies in the case of cookies, which contain personal information. The BOPI must obtain the data subject's consent when collection personal information and if data is used beyond the advertised purpose of data collection. It must also obtain such consent when sharing data with a third party. The Act does not specify any rule to collect this consent. There are restrictions to the necessity of consent (the BOPI can share data on legal grounds for instance).
Data subjects can ask for the editing of their data or refuse its sharing with a third party. They can also ask the BOPI to stop using their data if the BOPI uses data beyond the purpose it has advertised. In such cases, data subjects can ask the deletion of their data.
The BOPI is responsible for putting up an adequate data security mechanism (such as access control and employees training).

Conclusions

Japan is clearly advanced on big data issues and has a well-developed ICT strategy. It has updated its IPR framework to create space for information analysis and considers an update of its privacy framework. Open data is still in its start-up phase without a lot of uses yet.

Two of the three countries (Japan, Australia) clearly make work of the framework conditions for big data. Both have developed strategies on or involving big data. Similar legal discussions as in the EU pop up surrounding big data.

[312] LOD Challenge, http://lod.sfc.keio.ac.jp
[313] Carter, Lawrence G. and Mizuho Miyata, "Data protection in Japan: overview", *Practical Law*, no date. http://us.practicallaw.com/5-520-1289

China seems to run behind in the development of legal frameworks and privacy, although it has an important home market of data and homegrown industrial players.
Open data is a clear policy choice in all three countries but its implementation is in a start-up phase.

8 CONCLUSIONS

Our comparative review of data policies shows a diverse adoption and development by governments of big data policies. Notwithstanding this we also see similar problems arising concerning IPR and privacy law, and a general turn towards open data policies.

Big data is approached from different angles in the big data strategies we encountered. Comprehensive big data strategies are rare. Only the EU and Australia have a comprehensive big data strategy. Most often big data is raised from a specific angle or integrated in a broad ICT-strategy.
The EU approach of a data-driven economy proves to be rather unique. In the US and Australia is the focus on transparency and the creation of more efficient government, while privacy or other concerns are dealt with in specific policy reviews.
This concept of a data-driven economy puts data central and shifts the attention from an infrastructure or technology towards the business processes it enables. In this context the broader conceptualization of interoperability levels (technical, legal, semantical and organizational), developed in the EIF in the context of public services, proves useful as analytical tool.

A first important building block of data policies is the private law framework of intellectual property rights and contract law. We illustrated the use of these private laws tools by private actors through a comparison of the terms of use used by Twitter and Google. This revealed how both shape their own distinct and partially open, partially closed data ecosystem in a way favourable to their business model. These business models define their platform politics or the organizational interoperability they allow with their services. Through their terms, or the legal interoperability, they define how the data under their control can be linked up to provide other services.
The same private law tools are used for making open content licenses. These licenses do not aim to shape data enclosures but to open their data. They also use intellectual property law but to waive the protection offered by this law. Open content licenses are instrumental in shaping a much more open data ecosystem. However, fundamentally they attempt to fix the problems created by an unadapted intellectual property framework, which through too strong protections leads to enclosures.

This IPR framework proves too restrictive for big data processing. The exceptions in the IPR framework date from the pre-internet era and cannot cover data mining in a lot of jurisdictions. We found policy documents proposing changes or actual changes being made (UK, Japan). The US IPR system has less trouble with data mining. Its fair use-regime proves to be more flexible to adapt to technological changes.
Another problem is the patchwork of jurisdictions which can create legal obstacles for combining data from diverse sources. These obstacles are more related to contract law. Open content licenses are confronted with the challenge to develop a text which can function over a wide range of jurisdictions.

Also in privacy regulations proves the EU to be the most restrictive, compared to the US. Big data processing puts into question the sustainability of the basic concept of data protection: personal data, being distinguishable from other, non-personal data. Also the application of the data protection principles in actual big data processing proves difficult. Critics argue that data protection has to shift its attention to use instead of collection. However, the response of the data protection authorities shows that the data protection principles can be applied on big data

processing through an interpretation which incorporates some of these elements. A similar discussion can be found in the US. In answer to privacy concerns raised by big data, an attempt is made the enlarge the scope of privacy protection through a Consumer Privacy Bill of Rights. Also here the question is raised if the Fair Information Principles in this Bill of Rights do still work in the era of big data.

Big data and the data economy have a large impact on the regime of access and re-use of public sector information through the advent of open data policies. Governments have become aware that the data held by them is a valuable resource in the knowledge economy. Access regimes shift from a transparency tool mostly based on information requests into active providing of large amounts of data through open data policies. These open data policies entail more than making data available. Also the usability and interoperability of the data becomes a concern: making data machine-readable and enriching it with metadata improves the semantic interoperability. This building of data ecosystems with open data also has its twin in the EU effort to make public services interoperable across borders.

The comparative research reveals a mixed adoption of big data and open data policies. Both within the EU and within third countries we find trendsetters and followers. The UK prove to be a trendsetter in dealing with the legal frameworks impacting big data and with open data, while Germany is pioneering with establishing frameworks for the industrial use of big data. Within the EU is the INSPIRE-project important in gathering first experience in the wider aspects of the data economy like semantic and organizational interoperability.

In general, this review finds that big data policies are very much in a developmental phase, both in Europe and around the world. While some countries and sectors are more advanced than others, no specific country has a comprehensive policy on big data that addresses all of the important aspects identified here. Furthermore, this report also finds that big data policy is a complex arena that includes elements associated with intellectual property, privacy and data protection, open data and economic development. As such, it raises a question as to whether such a comprehensive policy is possible at the moment. Considering big data developments across a range of different policy areas may be the most effective way to ensure that the possibilities and pitfalls of big data are considered across the European policy landscape. But it remains a necessity to ensure that localized approaches do not lead to limited tunnel visions, and are brought in contact with each other. The fact that formulating a comprehensive policy is difficult points to the need to set up areas for more comprehensive policy learning.