European Commission
DG CONNECT

A European strategy on the data value chain

This paper outlines a European strategy on the data value chain. It looks at the underlying drivers and challenges and makes proposals on what the European Commission (DG CONNECT) can do in order to turn the potential of data for the economy and society into reality.

The current paper is the result of a prior detailed analysis and several rounds of internal and external consultations organised by the 'Data Value Chain' Unit in DG CONNECT. It aims at triggering a wider debate and action from the side of stakeholders to make data work for Europe.

The draft strategy makes the following value proposition:

1. nurturing **a coherent European data ecosystem** that will bring together large software firms, SMEs, data-intensive sectors (private and public), researchers, academic institutions and capital providers;
2. stimulating **research and innovation around data** as well as the uptake of data services and products that are cross-sector, cross-lingual and cross-border;
3. putting in place a specific set of actions to improve the **framework conditions** for extracting value out of data, including the competence base, infrastructure, standards as well as a favourable policy and legal environment.

In order to facilitate the implementation of the three general objectives of the proposed strategy, as a prerequisite, we need to considerably reinforce synergies between data-related European policy and funding activities (for example, data aspects of transport and of environmental policies, eGovernment, eScience, Smart Cities, Internet of Things, research and development activities in general, sector-based research and innovation activities addressing data, etc.).

The strategy on the data value chain will build on and complement the open data package[1] of the European Commission and the specific actions for the area of scientific data[2].

The data value chain is at the centre of the future knowledge economy, bringing the opportunities of the digital developments to the more traditional sectors (e.g. transport, financial services, health, manufacturing, retail). A better use of data, building on progress in data-analytics and processing, has the potential to transform Europe's service industries and significantly increase their efficiency. This unlocked value of data will enhance a wide range of new innovative information products and services thereby increasing the productivity of European companies, including SMEs. It is also a key area for web-entrepreneurship, given that there are no capital-intensive barriers preventing market-entry, in particular, in the light

[1] COM(2011) 882 final
[2] COM(2012) 401 final and C(2012) 4890 final

of developments such as cloud-computing, high performance computing and the availability of abundant bandwidth.

The strategy on the data value chain aims at extracting the maximum value from data by building on the intelligent use of data sources across the Member States. Given the fact that European data economy is fragmented to a great extent along sectors, languages and borders, it is essential that significant efforts are made towards a coherent European data ecosystem that will have a positive effect on the whole of the economy. Therefore a strategy on the value chain strategy is necessary to establish the pre-conditions for the cross-border, cross-sector and cross-language flow of large amounts of public and private data throughout Europe in order to create value again and again. This value will then be translated into growth and jobs which in the end bear benefits for the whole EU economy and all European citizens.

TABLE OF CONTENTS

1. WHAT IS AT STAKE?

Data is everywhere and has an impact on almost every part of human life.

A definition of data

According to ISO/IEC 2382-1, data are "a reinterpretable representation of information in a formalized manner, suitable for communication, interpretation or processing". Data can either be created/authored by people or generated by machines/sensors. Often, it is generated as a "by-product" of other processes.

A wide range of data falls under the scope of the above definition, including geographical information, statistics, weather data, research data, etc.

For example, a video is a huge amount of data as such, but the additional information therein, its interpretable content, is often more important than the stream of bits and bytes. Films and other cultural artefacts resulting from creativity are not as such addressed in the data value chain strategy, although techniques such as data-mining can also be relevant for the creative industries. Moreover, data can support the development of new creative products and services, for example apps on mobile phones, or the innovative use of metadata related to cultural objects (e.g. from Europeana).

The progress in the IT environment (availability of broadband and big data tools, cloud services, HPC) now makes it possible to better use data in order to generate a wide range of innovative information products and services; increase the productivity of all sectors of the economy through improved business intelligence; better address societal challenges; as well as improve research efficiency and speed up innovation. In the public sector, it will lead to cost reduction of operations, increase of efficiency and better and more personalised services for citizens. The intelligent use of data enables the creation of new products, the optimisation of the production or delivery processes, the improvement of the market, new organisation and management approaches, and the reinforcement of research and development.[3]

The data value chain is at the centre of the future knowledge economy, bringing the opportunities of the digital developments to the more traditional sectors (e.g. transport, financial services, health, manufacturing, retail). A better use of data, building on progress in data-analytics and processing, has the potential to transform Europe's service industries and significantly increase their efficiency. This unlocked value of data will enhance a wide range of new innovative information products and services thereby increasing the productivity of European companies, including SMEs. It is also a key area for web-entrepreneurship, given

[3] OECD (2012) report on 'Exploring data-driven innovation as a new source of growth', DSTI/ICCP(2012)9/REV1, p. 13

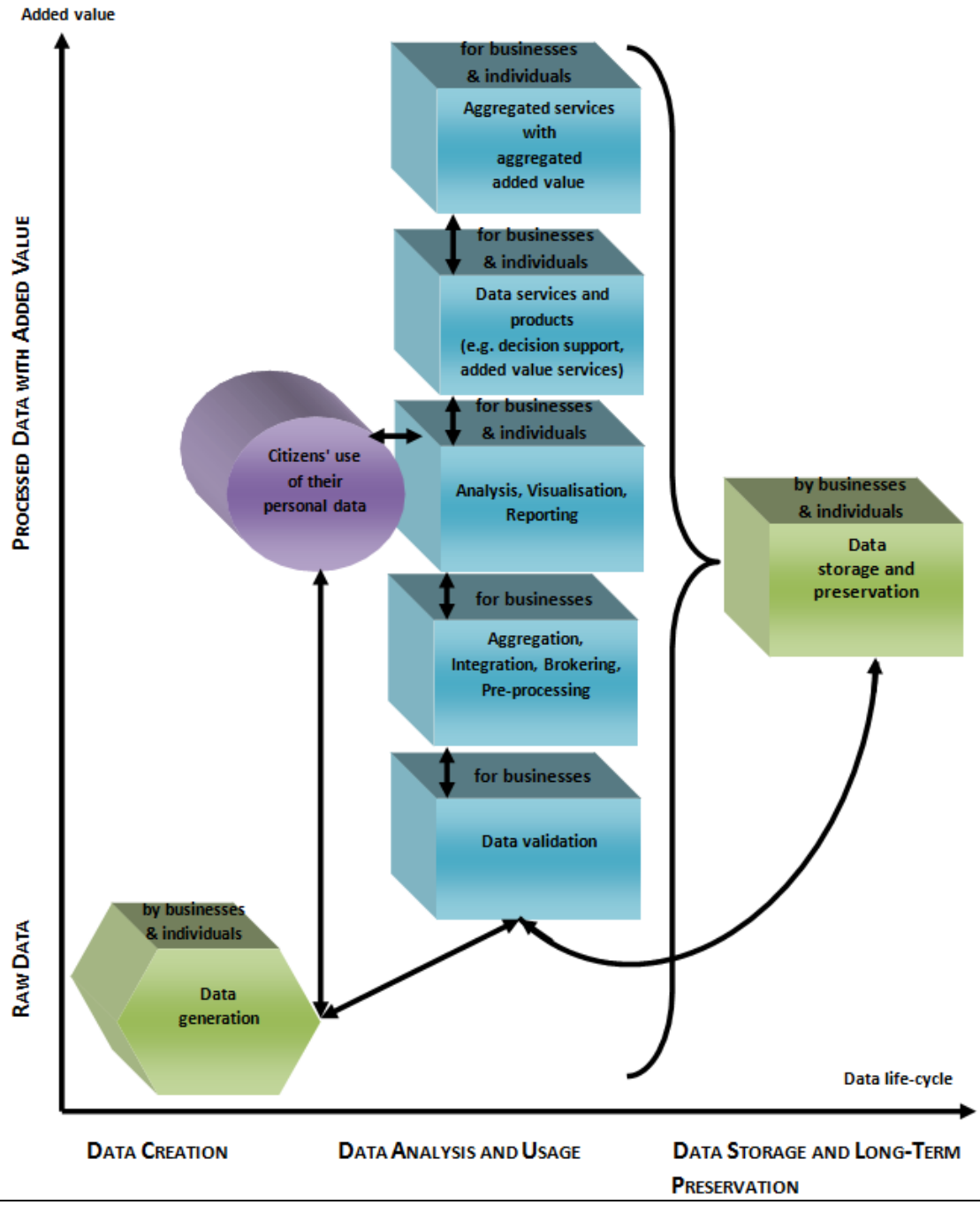
that there are no capital-intensive barriers preventing market-entry, in particular, in the light of developments such as cloud-computing, high performance computing and the availability of abundant bandwidth.

The intelligent processing and use of data are also essential for addressing societal challenges as well as for tackling environmental challenges. Data can, for example be used to enhance the sustainability of national health care systems in the Member States. Informed policy decisions in the areas of transport, land use and climate change also depend increasingly on analysis of the available data.[4]

A more effective use of data across the economy and society will contribute to achieving the aims of the Digital Agenda for Europe and to realising the potential of the digital single market. It will contribute to smart, sustainable and inclusive growth and the creation of jobs throughout the EU.

[4] COM(2011) 882 final

*A schematic overview of the **lifecycle of data** and added value in **the data value chain**[5] is presented below*[6].



The focus of the **data value chain strategy** will be on the **value generation** at the centre of the above graph.

[5] The data value chain is not a linear process but there is a large number of potential loops in it. For the sake of maintaining the clarity of the graph, these potential loops have only partly been visualised.

[6] The third dimension of the graph focuses on the socio-economic impact of value extraction at the different stages of the data value chain.

For some companies the production and sale of data is a business goal in itself, but more and more, valuable data is produced as a by-product of activities that serve a different primary purpose. Several layers of value can be added to the original raw data. Ideally data is generated by public sector bodies or private companies in a way that facilitates re-use and value generation based on the data over and over again.

Special attention is to be paid to a particular aspect of the data value chain, namely to citizens' use of their own personal data. This specific aspect also appears in the graph as it seems to be crucial to facilitate individuals' empowerment to better use their personal data for private or economic purposes.

When examining the European data market, some scattered examples of business intelligence systems can be found, like for example the services of Climpact (http://www.climpact.com) which help businesses understand, measure and anticipate the impact of weather conditions on their specific activities. Examples for data-based added-value services and products, as individual initiatives, can be found. However, further efforts seem to be necessary so that they function in a cross-sectoral context and at a larger scale:

- aggregated services with aggregated value: Thomson Reuters financial information service integrating a wide range of data-services.

- data services and products: Decision Support System (DSS) designed for the oil sector; apps, based on traffic information, maps and meteorological information;

- analysis, visualization, reporting: activities of companies like Augify (http://www.augify.com/), Quadrigram (http://www.quadrigram.com/) and InstantAtlas (http://www.instantatlas.com/)

- aggregation, integration, brokering and pre-processing: activities of companies like Mendeley (http://www.mendeley.com/), DataMarket (http://datamarket.com/) and Talend (http://www.talend.com)

- data validation: activities of companies like Data Publica (http://www.data-publica.com/), Timetric (http://timetric.com/) and Duedil (https://www.duedil.com)

The data market growth is reflected in **job growth in analytics and data science** that has been steep over the last years. Between 1990 and 2010 the number of related jobs worldwide increased by some 1.200 %.[7] A recent study predicts that in the UK alone 58.000 jobs will be created as a result of new business start-ups and increased demand for data-specific roles in the coming five years.[8]

[7] 'Building Data Science Teams', D.J. Patil, O'Reilly Media, 2011; the analysis was done on the basis of Linked-in data

[8] Data equity, Unlocking the value of big data, report for SAS by Cebr, April 2012

The industries dealing with data have grown rapidly over the last few years. According to IDC, worldwide **Big Data[9] technology and services** are expected to grow from USD 3.2 billion in 2010 to USD 16.9 billion in 2015.[10] This represents a compound annual growth rate of 40% or about seven times that of the overall information and communications technology (ICT) market.

Even more important from an economic point of view is **the effect that a good use of data can have on specific sectors**. For instance, it is estimated that effective use of data in the US health sector could generate USD 300 billion in value per year[11]. The importance of data in the health sector is confirmed by the recent report of the eHealth task force.[12]

> The high-level group of experts on eHealth recognised the importance of data for safer, better and more efficient healthcare throughout Europe. Their key data-related recommendations are also relevant for areas beyond the health sector[13]. They aim at a better use of the data, also across borders, and at the empowerment of individuals to use and make decisions about their personal data. This could, for example, boost the integration of user-generated data with official medical data, leading to healthcare that is more integrated and personalised.

Major effects of a better use of data are also expected in sectors such as goods producing industries, transport, energy, finance, retails and government services.[14] Research shows that companies that use 'data-driven decision-making' enjoy a 5-6% **increase of productivity**.[15] In fact, access to data is more and more important for business success across the board. Access to and use of data can now make the difference between business success and failure.

The eScience initiative shows that a better use of data also has the **potential to revolutionise the way in which research is carried out**.[16] E-Science stands for research enabled by e-infrastructure/ICT. It is essential for meeting the challenges of the 21st century in scientific discovery and learning, given that scientific activities are increasingly undertaken through online global collaborations, using very large research data collections, huge computing resources and high performance visualisation.

---

9 McKinsey defines Big Data as ' large pools of data that can be captured, communicated, aggregated, stored, and analyzed' (Mc Kinsey report on 'Big Data, The next frontier for innovation, competition and productivity', May 2011). Another definition is 'A collection of large and complex data sets which can be processed only with difficulty by using on-hand database management tools' , http://mike2.openmethodology.org/wiki/Big_Data_Definition

10 Worldwide Big Data Technology and Services 2012-2015 Forecast, IDC, March 2012

11 Mc Kinsey report on 'Big Data, The next frontier for innovation, competition and productivity', May 2011

12 http://ec.europa.eu/information_society/activities/health/policy/ehtask_force/index_en.htm

13 http://europa.eu/rapid/press-release_IP-12-453_en.htm?locale=en

14 Data equity, Unlocking the value of big data, report for SAS by Cebr, April 2012 and Mc Kinsey report on 'Big Data, The next frontier for innovation, competition and productivity', May 2011

15 MIT-research, referenced in 'Big Data for All: Privacy and User Control in the Age of Analytics', O. Teme and J. Polonetsky, Northwestern Journal of Technology and Intellectual Property, 20 September 2012

16 The report 'Riding the Wave: How Europe can gain from the rising tide of scientific data, which resulted from the work of the High Level Group on Scientific Data Infrastructures, specifically looked into the multiple challenges of access to and preservation of the large amounts of research data produced by modern science "Riding the wave": http://ec.europa.eu/information_society/newsroom/cf/itemlongdetail.cfm?item_id=6204

Large players from outside the EU are **investing in data-related technologies and business**. Examples are Microsoft, IBM and Google. The US government recently announced a USD 200 million initiative for R&D around Big Data.[17] If Europe wants to have a place in the data-economy and benefit from the potential of data, it will have to move now. In our data value chain policy we will need to look for the areas and segments where we can make a difference.

In order to fully exploit the potential of data, in addition to the economic dimension of data, a whole range of other issues need to be dealt with and properly addressed. These issues include the social and environmental dimensions just as much as the considerations with regard to addressing privacy and security concerns through a range of policy and legal measures aiming at building trust in data providers and users, let them be individuals or business actors.

There is clearly one aspect that is specific to the European context and needs to be taken into account, namely **multilingualism**. This makes the cross-border, cross-sector exploitation of data particularly challenging. On the one hand, building intelligence on databases as well as on unstructured data and streaming data that are using different languages is technically more complex. On the other hand, developing data-products that are adapted to the different European markets entails supplementary costs and skills that many SMEs do not have. This specific aspect will need to be addressed when defining the various elements of a coordinated set of actions around data at the European level.

2. WHY DO WE NEED A DATA VALUE CHAIN POLICY AT EU LEVEL?

Extracting value from data presupposes that we can process efficiently huge amounts of different types of data from a high number of various types of sources. The data challenge stems from these 'V's (volume, velocity, variety[18]). There is a fourth challenge that is often considered, namely veracity referring to the trustworthiness of data.[19]

The internet is rapidly expanding. By 2020, it will be 44 times as big as it was in 2009.[20] At the same time, data is generated constantly in an ever growing number of places and by an ever growing number of actors while a large proportion of potentially re-usable data resides within silos within institutions or companies. The **volume** sub-challenge requires novel approaches, often referred to as Big Data technologies and methodologies. These are needed when conventional database technologies cannot be applied to storage and computing issues. The issue of big data has been referred to as "the next frontier in computing".[21]

---

17 http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release.pdf

18 As defined by Doug Laney's (META Group, now Gartner, 2001) http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

19 The major difference between addressing the first three challenges (volume, velocity and variety) and the fourth one (veracity) seems to be that the first group of challenges can be tackled by using technological solutions which is not necessarily the case for the fourth challenge. (http://www-01.ibm.com/software/data/bigdata/)

20 'The Digital Universe Decade – Are you ready?' IDC, May 2010

21 http://www.zdnet.com/blog/btl/us-government-commits-big-r-d-money-to-big-data/72760.

The **velocity** sub-challenge applies to situations where data needs to be collected, processed and interpreted without delay because fast (real time) decisions are necessary, for example in manufacturing process controlling, transport and energy systems, financial (e.g. live trading) applications. When the decision-makers or end-users are humans, concise, expressive and accurate data visualisation is important to convey the most important trends and characteristics underlying the data to the person that needs to take appropriate action.

The **variety** sub-challenge applies when data come from diverse sources and in diverse forms, for example social media content created by hundreds of millions of online users; traffic flow data or environmental parameter data from sensor networks; position data from GPS systems and devices; official documents from public administrations etc. This sub-challenge requires the ability to interpret and analyse data, especially when the data is unstructured human language (spoken or written) or pictures and videos. It also requires a certain degree of interoperability, so data coming from different sources can be combined.

In order to benefit from the potential of data and to create **Value**, the different 'V's cannot be considered individually, but ought to be addressed in a holistic way. As prerequisites for the value extraction, data needs to be of good quality and follow a pre-defined set of international standards.

3. WHAT IS OUR VISION?

The data value chain strategy aims at extracting the maximum value from data to provide value for the economy and citizens, by building on the intelligent use of data sources across the Member States.

Extracting value from data presupposes that we can process efficiently large amounts of different types of data from a high number of various types of sources (business operations, product development, personal data, etc.). As prerequisites for the value extraction, data needs to be of good quality (i.e. complete, accurate, consistent, timely and compliant with a pre-defined set of international standards). A major leap is to be made in Europe from manual data-handling on an ad hoc basis towards a systematic, fully automated data-generation, data-processing and value extraction practice.

This can be compared to the industrial revolution, which has profoundly changed our society. Also there, small scale, individual production was transformed into large scale automated production processes. However, the transformation around data will not take several centuries, but will materialise in the coming decade.

The following three guiding principles will filter through all the various segments and dimensions of the proposed data value chain strategy for Europe:

1. a wide availability of good quality data, including the free availability of publicly funded data;
2. free flow of data across the European Union, as part of the digital single market; and
3. finding the right balance between individuals' potential privacy concerns and the exploitation of the potential of the reuse of their data while also empowering citizens to use their data in any way they wish to.

The European data value chain strategy will focus on nurturing **a coherent European data ecosystem** that will **stimulate research and innovation around data** as well as the uptake of data services and products. For that purpose, it will put in place a specific set of actions to **improve the framework conditions** for extracting value out of data, including the competence base, infrastructure, standards as well as a favourable policy and legal environment.

In order to facilitate the implementation of the objectives of the proposed strategy, as a prerequisite, **synergies between data-related policy and funding activities** (for example the cloud strategy, HPC, eGovernment, eScience, research and development activities in general, sector-based research and innovation activities addressing data, data aspects of transport and of environmental policies, Smart Cities, Internet of Things, etc.) in the European Commission will be considerably reinforced by strengthened coordination activities.

A European strategy on the data value chain will have far-reaching impact on the European society and economy as a whole and, in particular, on the data-intensive sectors.[22]

4. KEY POLICY TARGETS

The key policy targets of the data value chain strategy are the following:

- increase the number of data-related jobs (at least 250.000 new data related jobs in Europe in 2017);[23]
- increase the number of data-related start-ups and fast-growing SMEs;
- increase the revenue generated based on data in the Member States;
- improved use of data for decision-making processes in the private sector and the public sector

[22] Source: OECD (2012) report on 'Exploring data-driven innovation as a new source of growth', DSTI/ICCP(2012)9/REV1

[23] Figure extrapolated from the UK study referenced above

- increase citizens' use of data for informed behavioural decisions

Progress will be monitored through a data market monitoring tool (see below under section 5.1.2).

5. MAIN STRANDS OF A EUROPEAN DATA VALUE CHAIN STRATEGY

While making sure that the three guiding principles set out in the vision statement are fully respected in every element of the data value chain strategy, a value proposition consists of the following strands:

1. nurturing **a coherent European data ecosystem** that will bring together large software firms, SMEs, data-intensive sectors (private and public), researchers, academic institutions and capital providers;
2. stimulating **research and innovation around data** as well as the uptake of data services and products that are cross-sector, cross-lingual and cross-border;
3. putting in place a specific set of actions to improve the **framework conditions** for extracting value out of data, including the competence base, infrastructure, standards as well as a favourable policy and legal environment.

As described under Point 3, the implementation of the above objectives will be supported by strengthened coordination activities aiming at **reinforcing synergies between data-related European policy and funding activities**.

5.1. Development of a coherent European data ecosystem facilitated by research and innovation

5.1.1. Towards an ecosystem for data - bringing the actors at the European level together

In Europe there is substantial but fragmented research around data in universities and research institutes. Larger software firms (SAP, ATOS, etc.) are working and growing in the data market. SMEs offer solutions for specific applications in many sectors (eHealth, energy, finance, government, etc.) or are active in specific innovative areas (e.g visualisation, analytics). Overall data-use is growing, but companies and public services are not fully benefiting from the intelligence that could be extracted from their own data and from data available outside their organisations.

There is, currently, no coherent data ecosystem at the EU level[24], where the total represents significantly more than the individual elements and where the presence of one type of player is beneficial for the whole value-chain.

[24] There are, however, some examples for local data ecosystems that are developing, for example, in the Netherlands (http://www.almeredatacapital.nl/) and Austria (www.networkeddata.at).

A **coherent data ecosystem** would have the following actors and features:

- Cooperation between universities/public research institutes and private partners on R&D for take-up in the different sectors;
- a sufficient number of qualified 'data-workers';
- symbiosis between larger firms and SMEs, where SMEs carry out specialised research and development tasks, and where the larger firms support starting and growing SMEs by giving them business opportunities;
- a constant investment-flow towards start-ups and growing firms active in data technologies and applications;
- public organisations that act as 'launching customers' for new data services;
- wide availability of re-usable data (preferably in machine-processable format) that can be a basis for new activity and testing, as well as an environment where developers share back their cleaned/integrated data for further use;
- a solid enabling infrastructure, based on fast internet and data storage services, including infrastructures to support data-driven R&D. An effective use of cloud computing will lead to EU-wide economies of scale.

The development of a well-working data ecosystem could significantly improve the impact of publicly and privately funded research, as well as the chances for start-ups to survive, for SMEs to grow and for larger firms to reap the benefits of the data-revolution.

A pre-condition for the emergence of a European data ecosystem is that the different actors – researchers, software firms and the interested sectors – talk to each other, define a shared vision and jointly identify gaps in the current data landscape. It is essential that owners of large amounts of data, in the public and private sector, are also involved in these discussions.

The Commission can contribute to this by **bringing the relevant players together** and by **steering the available financial resources** that facilitate collaboration among the various stakeholders in the European data economy. On-going activities under FP7 support the work in this area. However, this is only the first step on the way. A more structured method of getting input from industry on policy and technology issues relevant for the data value chain and for fostering the development of the data ecosystem is necessary. This could take the shape of a **European Data Value Chain Platform** that would address technology needs and policy issues around the data economy while bringing all the stakeholders together, including the ones with capital to invest in the development of a European data ecosystem. It would, amongst others, develop a technology roadmap for data that would translate the basic requirements of the sectors into a research agenda for data while defining and promoting the strengths of EU technology in the context of global technology challenges with respect to data. The Platform will be at the basis of **a European PPP around data**.

Bringing together the data stakeholders through a major yearly event will also help to create a stronger identity and more coherence in the data ecosystem. In 2012, CONNECT organised a **European Data Forum** in Copenhagen. Following the event, stakeholders have asked to establish the European Data Forum as a multidisciplinary and cross-sector meeting place. A second edition of the European Data Forum is currently in preparation and will take place

under the Irish Presidency. The event should not just be a conference, but gradually develop in an event that combines a more classical conference format with a space where entrepreneurs can present themselves and their research and services also to venture-capitalists.

Proposed actions:

- Setting up a European Data Value Chain Platform, bringing together all key stakeholders relevant for the European data ecosystem by 2014;
- Developing the European Data Forum into a major yearly event that is, by 2016, widely recognised as a focal point for the development and interconnection of stakeholders in the European data industry.

5.1.2. Towards a data market monitoring tool

Better understanding the market development, facilitated by a reinforced European data ecosystem, is a pre-condition for measuring progress on the proposed key policy targets. A **data market monitoring tool** will give us an overview of where Europe is at the moment and where it is supposed to be going both in absolute terms and as compared to its main competitors in the global data market. It will also provide valuable information for the regular reviews of the Digital Agenda Scoreboard.

The data market monitoring tool should measure job-generation based on data-related activities, the increase in start-ups and SMEs in the area, the revenue based on data products and services, and the efficiency gains generated by an improved use of data. It should also look at key issues such as the availability of data scientists, the impact of the presence of infrastructures and 'data gravity'[25]. In addition to the economic dimension, the monitoring tool will also attempt to address the societal and environmental aspects.

Baselines as well as progress along these indicators need to be measured by sector as well and both at the level of the individual Member States and at EU level.

The data market monitoring tool will be developed and kept up-to-date with the help of external expertise, including the expertise of the wide stakeholder base of the European data economy. The collected information will be made available as open data.

Proposed action:

- Set up and run a data market monitoring tool to measure the achievement of the targets (tool set up in 2014, first measurement in 2015).

[25] The concept of data gravity refers to the following phenomenon: when data accumulates, the likelihood that it attracts data-related or data-based services and products is proportionate with the size of the data mass accumulated.

5.2. Stimulating research and innovation around data

5.2.1. Offering Big Data services for the whole of Horizon 2020

Data has become a key resource and asset for the European economy and society, with ever more products and services building on data sources. It is therefore not surprising that data is an inherent part of all segments of Horizon 2020. In particular, data technologies have the potential to transform and enrich each of the Societal Challenges and to have a major beneficial impact on the other pillars of Horizon 2020.

At the same time, the multiplicity of activities involving data technologies presents the possibility to create synergies around data handling across the programme. It would, however, require an active approach to avoid a duplication of effort, to stimulate technology transfer, to foster business opportunities and to spread good practices throughout Horizon 2020.

In order to harmonise and mutually reinforce data activities across Horizon 2020, the possibility of putting in place a dedicated Big Data services needs to be explored. This service package could fulfil co-ordination and service functions on data for all three pillars of the programme. Ideally, it will offer tools for a seamless use of different sorts of capacity across HPC, Cloud, Grid and Big Data computing for the benefit of all areas of Horizon 2020.

Proposed actions:

- Exploring possibilities for the co-ordination of transfer of knowledge and technologies for Big Data handling across the Commission under Horizon 2020;
- Setting up a range of Big Data services, fully operational by 2016.

5.2.2. Research driven by the needs of the different sectors

Under Horizon 2020 we intend to fund technologies for extracting value from data, tackling in a holistic way the issues of volume, velocity and variety. The technology development should be largely based on needs expressed by different data-intensive sectors (e.g. transport, health, government, retail, finance, public sector). Experiments should build on large and complex datasets from different sectors. The research will address data in different formats, various languages, arising from different sources and representing different content and media types.

To increase the relevance of research on data that the European Commission funds, a better mechanism needs to be put in place to identify the most relevant research problems including inter-disciplinary research problems and industry needs. The **European Data Value Chain Platform** (described under sub-chapter 5.1.1.) could play a major role in this process. It would bring together players from the different data-dependent sectors, from the research community on data and from the software industry. The European Data Value Chain Platform would have a central role in **defining a roadmap** for research and innovation around data. The roadmap should take into account Horizon 2020, national R&I stimulation programmes, and privately funded R&I.

The Platform can be at the basis of a European **public-private partnership** (**PPP) for data** addressing not only the technology and market aspects but also the societal dimension. A possible PPP preparation will need to be driven by industry, both large companies and SMEs working in close collaboration, involving all the relevant stakeholders.

Proposed actions:

- Define our research and innovation investment in data in Horizon 2020 on the basis of the needs of a range of data-intensive sectors;
- Preparation of a European technology roadmap for data by end 2014 as a basis for a possible PPP on data.

5.2.3. Promoting innovation – new data services and products

The data value chain strategy will **promote innovation around data services and products** (business intelligence, decision support systems, other added value services) that are cross-sector, cross-lingual and cross-border to ensure that these innovative data services and products actually make it to the market. The cross-sector, cross-lingual and cross-border aspects ensure that European data services and products support Europe's trade and political interactions on a global scene. Special attention will be given to SMEs as creators, aggregators, value adders or final users of data and related services. Large companies are equally important in providing large-scale industrial challenges and use cases, access to large and diverse data pools and user bases.

Market barriers that slow down the development and availability of innovative services should be tackled by **using a wide range of instruments**: grants, venture capital, procurement (public institutions as a launching customer), experiments, and data competitions.

More synergies will be sought between actions addressing data from the data industry angle and from the angle of the interest for a specific sector (e.g. eHealth, Smart Cities, eGovernment).

Proposed action:

- Promote initiatives which are cross-sector, cross-lingual and cross-border and support for SMEs by using the whole range of financing instruments in Horizon 2020 to stimulate innovation around data in the 2014-2020 period.

5.3. Improving the framework conditions for value extraction out of data

5.3.1. A favourable legal and policy environment for data

Different policy practices and legislation in the Member States have a large impact on what can be done with data, and lead to fragmentation within the digital internal market. Problems caused by these differences have been addressed by policy initiatives and legislation in different areas. The data value chain strategy will focus on the aspects that are directly affected by or have a direct relevance for the way the data market will develop in future.

5.3.1.1. Driving the open data policy

Open data policies, and in particular the Public Sector Information Directive, will increase the availability of raw/unprocessed data for companies to build on at no or low cost and create the conditions for the development of cross-border innovative products and services based on public sector data. The importance of open data policies was underlined by the G8 in an Open Data Charter[26] adopted in June 2013 and by the European Council in October 2013 Conclusions.[27]

The European Commission's open data policy got a boost through the open data package adopted by the European Commission on 12 December 2011.[28] The cornerstone of the package was a proposal to revise Directive 2003/98/EC on the re-use of public sector information. The revised Directive was adopted by the European Parliament and Council on 26 June 2013.[29]

A solid implementation of the package can make a difference in the way open data is used as a basis for innovation across Europe. At the same time the European Commission and the other institutions should set the example also at the practical level. Studies paid for by the European Commission and the related datasets should for example be systematically made available in a way that facilitates re-use (standardised, machine processable meta-data should be delivered by the contractor).

Proposed actions:

- Use the implementation of the revised PSI-Directive to drive open data policies in the Member States;
- Develop guidelines on licensing and charging within one year after the adoption of the PSI-Directive;
- Foster the process of data-harmonisation within the European Commission and extend the European Commission re-use Decision to other EU Institutions and Agencies from 2013 onwards.

5.3.1.2. Exploiting the potential of data while addressing privacy concerns

The trust of individuals in the way their personal data is handled may become a crucial factor for the development of a flourishing data-industry[30]. A European data ecosystem can therefore only function properly if this aspect is taken into account. In fact, companies

[26] https://www.gov.uk/government/publications/open-data-charter/g8-open-data-charter-and-technical-annex

[27] http://www.consilium.europa.eu/uedocs/cms_data/docs/pressdata/en/ec/139197.pdf

[28] For scientific data, the package was complemented by a Communication and Recommendation on scientific information, adopted on 17 July 2012.

[29] http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2013:175:0001:0008:EN:PDF

[30] A Eurobarometer survey found that 70% of Europeans are concerned that their personal data held by companies may be used for a purpose other than the one for which it was collected. http://ec.europa.eu/public_opinion/archives/ebs/ebs_359_en.pdf

offering data products and services that cater for the sensitivities of individuals in relation to their personal data may turn that feature into a competitive advantage.

Dealing with personal data across borders requires a clear and harmonised legal framework that provides the right balance between individuals' potential privacy concerns and the exploitation of the potential of the reuse of large amounts of data. The recent European Commission proposal for a Regulation on privacy therefore further harmonises privacy legislation across the EU[31].

When looking at the best ways to extract value out of data, one question is how individuals can be empowered to better use personal data for their own purposes, or even actively bring it into the economic process. The phenomenon of data exchanges where individuals trade their own personal data on their own terms already exists. A possible development is a switch from an all-or-nothing model of personal data privacy to a trade-off model in which people are reliably informed of the likely consequences of certain disclosures. Making business out of data while at the same time guaranteeing individuals' privacy needs to be at the focus of attention.

At the same time, individuals will want to get more control over their personal data held by public or private organisations and use this data in different contexts. Trust in data processors and data re-users clearly plays a key role in this respect. For instance, health data, held by doctors and hospitals could be combined by the individuals concerned with lifestyle data they collect.

Another area that needs further consideration is the potential power of new data technologies to combine information from different databases, unstructured data and data streams (e.g. through sensors) to find out information about individuals. Ways and means to deal with this issue need to be considered already now, in order to avoid that it becomes a barrier to a better and more efficient use of data in the medium term.

Technology can have a role in this. One possible avenue to be explored is 'Privacy by design', enhancing the technologies with basic features that ensure that privacy is respected, while promoting the use of data and releasing its potential value. More in general technological progress can play a major part in establishing the conditions for individuals to have increased control over their data and the use of it. It will open up new scenario's that empower individuals, and at the same time facilitate innovative ways to use data. The development of these technologies, together with large scale tests of what types of data-use individuals would feel comfortable with, will be addressed in the Horizon 2020 programme.

<u>Proposed actions:</u>

- Tackling new methods to integrate privacy features in IT-equipment ('privacy by design') and developing ways to empower individuals to handle their personal data in the context of Horizon 2020.

[31] 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), COM(2012) 11 final

- Sustained dialogue with stakeholders on the different aspects of Big Data and privacy from 2013 onwards.

5.3.1.3. Addressing issues related to intellectual property rights

For the development of a coherent European data ecosystem, legislation on the control of data based on **intellectual property rights** is also a relevant factor.

Databases may be covered by copyright or by the *sui generis* right of the 1996 Database Directive[32] (protecting non-original databases).

It is, however, not always clear what data (that are part of a database) can be used under which circumstances, even if the European Court of Justice has clarified the scope of the *sui generis* right of the Database Directive in a series of judgments.[33] Also, differences still exist between the Member States in handling the issue, which can be a barrier to the cross-border use of data.

One of the questions on the table is that of data mining: In what situations and under which conditions can large (proprietary) data collections be analysed in an automatic or semi-automatic way in order to extract new interesting patterns and dependencies. So far the debate has mostly focused on the use of data mining and text mining techniques for scientific purposes, including ways and means to lower the transaction- and opportunity costs involved.[34] The issue was addressed in the group on text- and data mining group in the context of the Licences for Europe s dialogue, [35] but no agreement was reached between the representatives of the different stakeholders.

In addition to considerations related to the legal framework, attention should be given to the development of a market for data protected by IPR. Currently, there are no well-working clearance systems that facilitate the use of data. Experiments that bring the offer and the demand of data together or that help sharing of data between owners of the data can contribute to a wider use of existing data within and across sectors. This type of experiment could be further supported under the Horizon 2020 programme.

<u>Proposed actions:</u>

- Dialogue with the stakeholders on the different aspects of Big Data and IPR from 2013 onwards;

---

[32] Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases, OJ L77.27-3-1996, p. 20. According to the provisions of the Directive, a database is "a collection of independent works, data or other materials arranged in a systematic or methodical way and individually accessible by electronic or other means".

[33] See for example the judgment of the European Court of Justice of 1 March 2012 in Case C-604/10, Football Dataco Limited and others v Yahoo UK Limited and others, which contains an overview of the earlier relevant cases.

[34] See for example 'The Value and Benefits of Text Mining', JISC, 2012

[35] http://ec.europa.eu/licences-for-europe-dialogue/

- Initiatives leading to IPR-exchanges for data(bases) protected by IPR (thus stimulating the market for data), and to bring more transparency in the offer and demand of data.

5.3.1.4. Addressing issues related to security

The handling, transfer, storage and analysis of huge datasets inevitably raise security concerns, especially when carried out in an international context.

The Data Value Chain Strategy will seek synergies with the on-going initiatives stemming from the EU's Cyber-security and ePrivacy strategies to ensure that data technologies develop in full alignment with the highest standards of network security and data protection.

When involving leading industry stakeholders and experts in the domain, special focus is to be dedicated to how to efficiently coordinate action against current and future security threats and facilitate innovation and investment in security-related technologies for data-related environments.

Proposed actions:

- Dialogue with the stakeholders on the different aspects of Big Data and security from 2013 onwards.

5.3.2. Enhancing interoperability

The lack of interoperability between datasets is relevant across the board and represents a major problem when combining data from different sources and turning them into new data products and services or using them for business intelligence purposes. It makes it also more difficult for machines to analyse, combine and extract value out of large amounts of greatly varied data.

Better standardisation and the creation of data in harmonised machine-processable format can help to overcome this barrier. In fact interoperability issues need to be taken into account at the stage of the production of the data. Data-harmonisation will contribute to a 'Semantic Web' allowing data to be shared and reused across borders, languages and sectors. (This point is further elaborated in Annex 2.)

Where possible, interoperability issues need to be addressed through collaboration beyond the European Union.

Proposed action:

- Ensure that all standardisation aspects relevant for the data value chain are sufficiently addressed and tackled in the ICT Standardisation Strategy of the European Commission.

5.3.3. Reinforcing the infrastructure for open data

A pool of data that can be easily found and used by everyone is an important base for innovation. Therefore the European Commission will support the development of a **pan-European portal for open data** as one of the **digital service infrastructures** under the Connecting Europe Facility.

The portal will make it easier to find and use data from different European sources and across the different languages for innovative services and products. At the same time it will drive the harmonisation of data-formats and licensing conditions across Europe, and will offer services such as an apps store.

In addition, a wide **availability of research data** in line with the 2012 Communication and Recommendation on scientific information[36], will lead to better and more interdisciplinary oriented research, faster innovation and shorter time to market. Therefore the actions undertaken in FP7 to develop a **pan-European infrastructure for accessing publicly funded research publications and data** should be taken further.

Proposed actions:

- Launch, develop and populate the EU open data portal, as well as the pan-European open data portal, by end 2014, building on existing projects, and develop it under the CEF framework (2014-2020);
- Continue the work to link at EU level the infrastructures for accessing publicly funded research data (eInfrastructures).

5.3.4. Supporting competence around data in Europe

Already now European companies are facing a shortage of data experts. It is likely that the demand for qualified data scientists will continue to outstrip the supply in the near future. The real challenge seems to be that the European data economy will need more and more people who are highly trained in data management, who have data analysis competences while having an understanding of business issues from a series of domains.

Over the next decade, the number of servers (virtual and physical) worldwide will grow by a factor of 10, the amount of information managed by enterprise data centres is expected to grow by a factor of 50, and the number of files the data centres will have to deal with growth by a factor of 75, at least. Meanwhile, an increase of less than a factor of 1.5 is expected in the number of IT professionals in the world[37]. Even though no such estimates are available for the specific EU context, it is not hard to predict that the trends in the broad term will be very similar in Europe as well.

Some individual research institutes and universities in Europe (e.g. the Karlsruhe Institute of Technology, the Southampton Open Data Institute) have recognised the exponentially

[36] C(2012) 4890 final

[37] 'Extracting value from chaos', IDC, June 2011.

increasing demand for European data experts and therefore give special attention to data. For instance, DG CONNECT co-finances a project that aims at developing a new university curriculum around data. However, more will be needed to bring these actions to the necessary scale and to make a real contribution to bridging the competence gap.

The shortage of data-experts needs to be addressed in the context of a **broader initiative on skills and ICT** that is currently under preparation by the different European Commission services. In the framework of this initiative ('Grand Coalition on Digital Skills and Jobs'[38]), the European Commission will join forces with companies, industry associations and the Member States in order to jointly tackle the double challenge of the lack of ICT skills and the large number of unfilled ICT-related vacancies in Europe.[39]

At the same time, a **data-specific initiative to improve the competence base** could be developed under Horizon 2020. The attention given by a network of some prestigious European academic institutions to data could in turn attract good students, both at BSc and MSc levels, and boost the number of people specialising in this area.

The data value chain strategy will facilitate work towards **a European network of centres of competence for data**, combining research around data with education activities (including online learning material), and which could also serve as an incubator for data-value related start-ups. The initiative would build on existing structures/institutions willing to give data-related issues a stronger presence in their profile. A close and continuous cooperation between the universities or research institutes and the market actors of the data industry is a pre-requisite for the success of this initiative.

Proposed action:

- Strengthening the skills base through the creation of a European network of centres of competence in collaboration with prestigious European academic institutions from 2014 onwards.

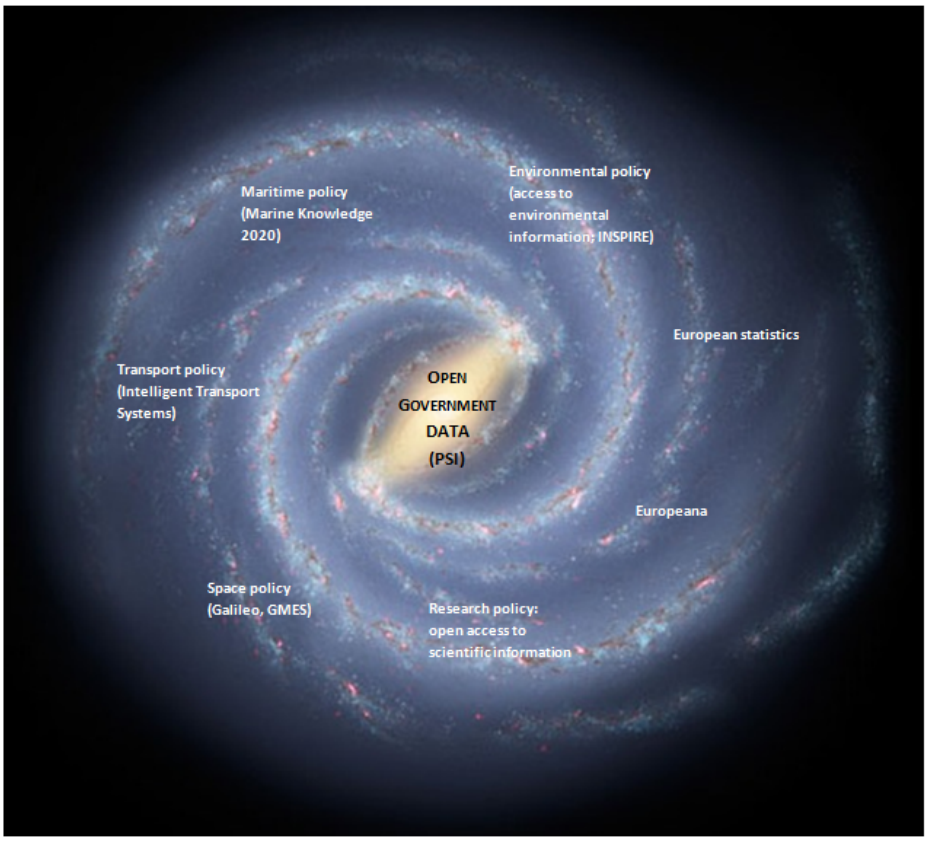
6. JOINING UP THE PIECES: REINFORCING SYNERGIES BETWEEN DATA-RELATED POLICY AND FUNDING ACTIVITIES IN THE EUROPEAN COMMISSION

In order to facilitate the achievement of the objectives of the strategy described under Points 5.1 and 5.2, as a prerequisite, synergies between data-related policy and funding activities in the European Commission need to be considerably reinforced by strengthened coordination activities.

Data is relevant for a whole range of policy areas addressed by the different services of the European Commission. One example is the range of policy areas with an open data component.

[38] http://europa.eu/rapid/press-release_SPEECH-12-282_en.htm
[39] http://ec.europa.eu/digital-agenda/en/grand-coalition-digital-jobs

Open data policies are in place or are emerging in different parts of the European Commission, targeting specific sectors or types of data, e.g. geospatial data, maritime data, environmental data (several examples are given in Annex 1).



Open Government Data Galaxy

This is in itself a positive development and underlines the key role of data in all different policy areas. However, **better co-ordination** and **mutual information** is essential in order to ensure that the policies complement and reinforce the general open data policy supported by the European Commission.

A first step towards more synergies is **mapping** and establishing an overview of the European data-related activities in DG CONNECT and in other Directorates-General.

A second step could be the creation of a – relatively light – **co-ordination mechanism** around projects and initiatives with strong data components, in order to ensure that policies complement each other, that the same activities are not funded twice and that everyone dealing with data is aware of the latest developments and actions around data.

These steps could be at the basis of a more detailed **map of the research and innovation projects** with significant data components funded by the European Commission (e.g around Smart Cities, eGovernment, eScience, eHealth, cloud computing, Internet of Things). Such a map will help to create synergies and avoid double work. It should be extended to the activities of the Member States.

At the same time, efforts need to be made to **capitalise on the knowledge and know-how of the groups** that European Commission brings together in specific areas, e.g. transport data, health data, and Smart Cities. They could, for example, be used to validate the results of research projects.

In addition to fostering internal communication on data between the Commission DGs and Services, special attention is to be paid to external communication on data. In order to reach the highest possible number of current and potential stakeholders in the context of the data-related activities of the European Commission, it is essential to find the most adequate communication tools and channels that reach the right audience.

Reinforced synergies between the different domains at the level of the European Commission will result in a clear map of competences, responsibilities, communication channels (input/output flows) and working frameworks around data. This will definitely have a further reaching impact not only on the various Directorates-General of the European Commission and the different policies but on all the economic sectors concerned by data. Through facilitated synergies and collaboration between the data-intensive sectors in the EU, Europe can be positioned at the forefront of data-intensive industries globally speaking.

Proposed action:

- Run a European Commission-wide analysis of the various policy fields from the perspective of data, starting in 2013, and then, as a second step, prepare an extensive map of research and innovation projects with significant data components[40] from 2014 onwards.

[40] Almost all research and innovation projects have data components in one way or the other. The mapping will focus on activities where a better use of substantial amounts of data is recognised as an explicit goal. These are the projects that can make a real difference.

ANNEX 1

EUROPEAN COMMISSION'S OPEN DATA POLICY AND THE DIFFERENT POLICY DOMAINS

Within the European Commission, there is already a range of policy initiatives in the various policy domains that build on and complement on the Open Data policy. The following list contains the policy initiatives where the links with the Open Data policy are solid[41]:

- the Access to Environmental Information and INSPIRE Directives[42], aimed at the widest possible dissemination of environmental information and the harmonisation of key datasets;

- the European Commission Communication on Marine Knowledge 2020[43], aiming amongst other things to make marine data easier and less costly to use;

- the DG MOVE initiatives following the 2008 Action Plan[44] for the Deployment of Intelligent Transport Systems (ITS), looking at, amongst other things, access for private service providers to travel and real-time traffic information;

- the European Commission's policy on open access to scientific information[45], which includes the principle of open access to publications resulting from projects funded by the European Union and a pan-European, participatory e-Infrastructure of Open Access Repositories; The JRC publications repository is also relevant in this context;

- the policies for the digitisation of cultural heritage and the development of Europeana, Europe's digital library, archive and museum, aiming to ensure the widest possible use of digitised cultural material and the related metadata;

In addition to the above policy initiatives, the gradual deployment of the European Commission's Open Data policy is expected to have an impact on a number of policy domains which do not yet have concrete relating policy initiatives but which will undoubtedly profit from the benefits of opening up a wide range of public and business data linked to their specific domain (education policy, tourism policy, consumer protection policy, public health policy, etc.)

[41] COM (2011) 882 final
[42] Directives 2003/4/EC, OJ L 41/26, 14.2.03 and 2007/2/EC, OJ L108/1, 14.3.2007
[43] COM(2010) 461 final
[44] COM(2008) 886 final/2
[45] C(2012) 4890 final;

ANNEX 2

DATA INTER-OPERABILITY, FORMATS AND STANDARDS

Formats and standards can be seen as coordination devices, in particular for data dependent activities. When two data producers/users exchange data according to a commonly agreed format or standard, the value of process in which they jointly engage increases because all costs of data interpretation, validation and conversion are simply eliminated.

In addition, when the format they use is also a standard, this benefits the economy as a whole because this reduces lock-in to proprietary solutions and thus encourages the arrival of new entrants into the market, to the benefit of competition and thus of quality of service and reduced prices.

Removing friction in the data exchange practices among European industrial actors will be a source of strategic advantage as it will support the deepening and diversification of the supply chain for data goods and services.

Examples of friction in data exchanges

- In your database you mention "IBM", in my database I mention "International Business Machines": when we merge the two databases somebody will need to spend some effort to realise that we are talking about the same entity and to choose a standard name for it;
- A French furniture manufacturer keeps a catalogue of 'chaises' while its English corporate owner wants to know how many 'chairs' have been sold last year. Somebody will need to figure this out manually.
- When I purchased my plan with cloud provider A, I started logging all the movements of my car fleet using the format of the geographical information system (GIS) which was available there. Now I want to move all my data to a different cloud provider B, which makes available a different GIS. Somebody will have to work on the conversion between the two data formats.

We can help EU actors in the data value chain remove this friction with medium term and long term plans, leading into one another.

In the domain of research data, the European Commission in collaboration with NSF (US) and ANDS (Australia) is supporting the set-up of the Research Data Alliance[46] through the iCORDI project[47] aiming at driving interoperability between emerging global data infrastructures for research.

[46] http://rd-alliance.org/

[47] http://www.icordi.eu/Pages/Home.aspx

There are at least three distinct aspects along which the development of European data standards ought to be fostered:

1. Standardised entity identifiers
2. Standardised, compositional concept systems (thesauri, taxonomies, ontologies)
3. Standardised formats

In the long term, standardised identifiers would allow entities of interests (from corporations to people to the tiniest identifiable component of complex artefacts, to individual fruits picked from an orchard) to be reliably traced across independently established processes, managed by millions of organisations, and across very complicated supply or logistics chains. To exemplify, standardise identifiers would allow a farmer to record when and where an individual pear was picked, a transportation company when and where the pear was delivered and a supermarket to record when and where it was sold and guarantee that all three actors are actually always referring to the same individual pear. Most barcoding and entity identifier systems today work at best within the domain of a single enterprise. Breaking this bottleneck with the use of universally shareable identifiers would allow for unprecedented levels of process integration. In the short term we could support the development of mapping software application that allow interested actors to establish that two or more distinct identifiers really denote the same entity (these are, incidentally, the same kind of application that in a different context undermine personal privacy via de-anonymisation). Also, it is likely that certain kinds of identifiers (e.g. identifiers for legal entities such as corporations) will prove to be more central to the EU data value chain than other types (e.g. identifiers for archaeological artefacts). It is in fact quite possible that a few identifier types will by themselves be responsible for the majority of cases of economic significance in the data value chain. If consultations with the stakeholders support this conjecture, our goal should be to foster the development and adoption of such identifier systems in their priority order.

Standardised systems of concepts would allow for semantic integration: being able to express unambiguously and across languages the fact that a given object (for which one may have a standard identifier) is a pear as opposed to an apple or a hammer would allow for anyone who needs to handle it to do so appropriately (keep it refrigerated, sell it by a certain date, handle it gently, etc…) An example of these systems of concepts is the Common Procurement Vocabulary http://simap.europa.eu/codes-and-nomenclatures/codes-cpv/codes-cpv_en.htm which allows publishers of calls for tender to specify the exact nature of the good or service they wish to procure. Similar standards, voluntarily negotiated across industries or supply chains would allow for data collected within a given process to be analysed for the benefit of a larger number of other processes, within or across industries. Equally important, the existence of such standards would allow customers of a given company to recover their data from said company and transfer them to a competing company for better service. This would increase competition and benefit European consumers. In the short term we could support the development of mapping software application that allow interested actors to establish that two or more distinct labels or expressions really denote the same set of entities (machine

translation will have a major role to play when the core issue is linguistic diversity). As in the previous case, it is likely that certain kinds of concept systems (e.g. products and services as used in e-commerce and procurement) will prove to be more central to the EU data value chain than other types (e.g. classifications of literary genres). As before, our goal should be to foster the development and adoption of such concept systems in their priority order.

Finally, even assuming the existence of standardised identifiers and systems of concepts, standards must be established for the actual formats in which data will be recorded on a physical substrate or communicated across a computer network. In the absence of such standards it will be impossible to have vendor-neutral machine readable data. Software engineering abounds with such formats (each optimised for various operating conditions): it will be important for Europeans involved in the data value chain to agree on and systematically use formats that are optimised for the integration of all actors on that chain as this will remove friction (various processing costs) along the chain. In the short term this can be addressed by supporting high performance format conversion software libraries. The general usefulness of such libraries makes them similar to public goods, which might best developed as open source components with very free licences as far as commercial reuse is concerned.

Just as it is important for data representation formats to be standardised, it will be important for data processes to be standardised so as to ensure the desired guarantees of data accuracy, freshness, coverage, etc.. This is analogous to what happens in mechanical engineering, where the materials properties of components of complex artefacts are certified by the adherence to the process used to manufacture them.

In which specific domains can we expect for the need of data standards to be most pressing? Reasoning from first principles, there are three important categories:

1. Domains in which the number of data producers that must coordinate their output is growing very quickly. This includes networked sensors of the type discussed in the context of the Internet of Things, drones collecting large amounts of 3D or photographic information, public transportation systems, cars fitted with advanced sensors, Smart Cities, RFID endowed objects, etc…

2. Domains in which data integration across processes carries large liability risks (e.g. e-health or finance)

3. Domains that directly support political objectives of the European Union such as cross-border e-commerce/procurement or safety of the food chain.