DIRECTORATE-GENERAL FOR INTERNAL POLICIES

POLICY DEPARTMENT A
ECONOMIC AND SCIENTIFIC POLICY

Economic and Monetary Affairs

Employment and Social Affairs

Environment, Public Health and Food Safety

Industry, Research and Energy

Internal Market and Consumer Protection

Data Flows – Future Scenarios

In-Depth Analysis for the ITRE Committee

EN

2017

Data Flows – Future Scenarios

European Parliament

DIRECTORATE GENERAL FOR INTERNAL POLICIES

POLICY DEPARTMENT A: ECONOMIC AND SCIENTIFIC POLICY

Data Flows – Future Scenarios

IN-DEPTH ANALYSIS

Abstract

Prepared by Policy Department A at the request of the European Parliament's Committee on Industry, Research and Energy (ITRE), this report examines the current state of play in the open data market and the legal framework in the EU. Barriers and possible solutions are identified in the form of future scenarios to 2020-25. The key policy recommendation is to instigate a system of Open Data Licensing to drive access to open data, akin to open source software licensing.

IP/A/ITRE/IC/2017-04 November 2017

PE 607.362 EN

This document was requested by the European Parliament's Committee on Industry, Research and Energy (ITRE).

AUTHOR(S)

Colin BLACKMAN, Camford Associates Ltd; Associate Research Fellow, CEPS.
Simon FORGE, SCF Associates Ltd.

RESPONSIBLE ADMINISTRATOR

Frédéric GOUARDÈRES

EDITORIAL ASSISTANT

Laurent HAMERS

LINGUISTIC VERSIONS

Original: EN

ABOUT THE EDITOR

Policy departments provide in-house and external expertise to support EP committees and other parliamentary bodies in shaping legislation and exercising democratic scrutiny over EU internal policies.

To contact Policy Department A or to subscribe to its newsletter please write to:
Policy Department A: Economic and Scientific Policy
European Parliament
B-1047 Brussels
E-mail: Poldep-Economy-Science@ep.europa.eu

Manuscript completed in November 2017

This document is available on the Internet at:
http://www.europarl.europa.eu/supporting-analyses

DISCLAIMER

The opinions expressed in this document are the sole responsibility of the authors and do not necessarily represent the official position of the European Parliament.

CONTENTS

LIST OF ABBREVIATIONS

API	Application Programming Interface
APT	Advanced persistent threats
BSD	Berkeley Software Distribution
DLT	Distributed ledger technology
DSM	Digital Single Market
FRAND	Fair, reasonable and non-discriminatory
GDP	Gross Domestic Product
GDPR	General Data Protection Regulation
GPL	GNU Public Licence
IC	Intellectual capital
IPR	Intellectual property rights
OSS	Open source software
URL	Universal Resource Locator (global address on the World Wide Web)

LIST OF BOXES

LIST OF FIGURES

LIST OF TABLES

EXECUTIVE SUMMARY

Background

Significant progress towards the digital single market is promised with the opening of data currently held as private archives for more general use, for innovation and development, in a “data-driven” economy. This In Depth Analysis explores the issues around access to open data in the EU, particularly in the context of the European Commission’s January 2017 Communication, *Building A European Data Economy*, and the proposed Regulation on a framework for the free flow of non-personal data in the European Union.

The report examines the current state of play in the market and the legal framework concerning open data in the EU, both at European and national levels. Barriers impeding access to open data are discussed and possible solutions are identified in the form of future scenarios to 2020-25. How and to what extent the proposed measures could be implemented, at EU level, is discussed with policy recommendations. This succinct briefing is based on recently published data and current analysis.

Findings

- *On the recommended legal framework to move forward*: the optimal choice is to re-use the structures of open source software licensing for open data.
- H*ow should competition be considered in the hypothesis of total opening?*: It will be advantageous if all competitors are considered as equal beneficiaries of open data. This implies that open data is positioned in a pre-competitive category, requiring further refining through processing to move up in worth to have commercially valuable.
- *Is data that is made to optimise the manufacturing process and to achieve company specific competitiveness affected by an unlimited opening of databases?:* As emphasised by the last point, the class of content that would be put into open datasets is unlikely to affect competiveness, as it should be too uninformative, diverse and disparate of itself and so would not approach what might be termed commercially useful or confidential data. To share it widely is feasible without infringing essential IPR for the enterprise.
- *What is the value of the commercialisation of this data?*: At a simple level, if data becomes widely traded, it will be valued by the market. Currently the total EU data market is estimated at €60 billion, of which only two percent is open data. More broadly, the report examines how open data can be valued and the factors that affect this.
- *Can regulation cover all aspects of this? Should more scope be given to contractual agreements?*: The report explores four scenarios, including doing nothing, using existing contractual law, the concept of a new right in data, as well as an approach suggested by open source software licensing.
- *What are the legislative measures that could enable the guarantee of the free movement of non-personal data in the EU?*: The report proposes a novel form of contractual agreement – Open Data Licensing – with different types of licence providing permission suitable for the wishes of different donors and appropriate to the kinds of application proposed.
- *What are the critical factors in data protection and privacy and how should we simultaneously tackle data management and security?*: A critical issue is the contamination of open data by personal data. This will require vigilance that can be enforced under the current GDPR.

1. THE EU DATA ECONOMY TODAY

1.1. Data: the Fuel of the Future

The generation, collection, processing and use of data is transforming our economy and society, similar to the way in which oil reshaped the 20th century. Technology, fuelled by data, is driving innovation in the 21st century. Data-driven technologies have the potential to enhance productivity and competitiveness, and benefit citizens in areas such as education, employment and healthcare. In healthcare, for example, opportunities include machine-learning tools for early disease detection, coordinating the collection and sharing of health data for improved treatment, and applications of the Internet of Things for assisted living.

Ensuring that EU citizens realise these benefits requires that data flows as freely as possible – across borders and between sectors. Attention has focused on barriers such as data localisation, i.e. rules that dictate the location of data storage or other processing in the territory of a specific country, which harms competition and particularly hinders startups from scaling up. This issue is partially addressed in the proposed Regulation on a framework for the free flow of non-personal data in the European Union (European Commission, 2017a) (see Section 3.1.3), but building a successful European data economy will depend on much wider range of issues needs to be addressed concerning who owns or controls data, how data may be stored, who may exploit it and how, and who benefits.

This chapter explores the status of digital property and data ownership, the current state of data sharing and development of data markets, competition in open data flows, and legislation today regarding open data. Chapter 2 discusses goals for open data access and how to achieve them, while Chapter 3 examines the obstacles to realizing these goals. Chapter 4 describes four scenarios for the future, while Chapter 5 makes policy recommendations on how to best support open data flows in the EU.

1.2. The Status of Digital Property

Since data is the most valuable currency in the digital economy, it is not surprising that there is growing concern over how data is used, especially so-called “open” data which is shared. These concerns relate both to its underuse because access is restricted, and, conversely, to its overuse when it involves some form of misuse. The latter implies privacy and security hazards are likely, or alternatively, that the rights to its intellectual property are being transgressed. But just what is “open data” – the box below explains:

Box 1: Defining Open Data

The optimal, ideal open dataset has the following properties:

- Data is non-personal in content, includes public and commercial data, and is of use to manufacturing and service industries.
- Cost: open data can be accessed with insignificant charges or it may be free, often as it is already in the public domain.
- Intellectual property rights and confidentiality: there are either no limits or negligible restrictions on the access, copying, processing, addition, transfer, exploitation, alteration, reformatting and redistribution of the whole dataset. No part is reserved or precluded from processing. It may be used in any way for commercial gain by the new user.
- Accessibility: a wide range of users is permitted to access the whole open dataset, with no preferential treatment for any user and dataset parts reserved for preferred users.
- Formats and modes of access: the datasets are in a machine-readable form and so can be processed by computers, using the semantics of the data, probably via its metadata which should also be openly published, if the data is not in a well-known format.

In October 2013, the European Council recognized the importance of the digital economy, innovation and services as drivers for growth and jobs, and called for EU action to provide the right framework conditions for a single market for big data and cloud computing, to which the European Commission responded in its 2014 Communication, *Towards a Thriving Data-driven Economy* (European Commission, 2014).

Most socioeconomic activities, industrial processes and research now involve data collection and processing on a large scale. This is driving new business processes, products, services and technologies. The datasets being produced are so large and complex that processing such "big data" needs a new generation of data management tools and methods based on novel techniques, such as advanced data mining and often using cloud computing for cheaper storage.

The EU's Digital Single Market Communication in May 2015 called for "the rapid removal of key differences between the online and offline worlds to break down barriers to cross-border online activity". However, currently in the EU there is no coherent common set of regulations covering data ownership, or its use, or cross-border transfer and re-use of data. In other words, the notion of ownership is absent today for what amounts to a digital property held in data. The free exchange of such assets could form a key pillar of future EU economic development. Consequently, companies and trade organisations increasingly see data flows as essential to future trade within the digital single market (DSM). This is not just for the service sector but increasingly holds for the manufacturing sector – aerospace, advanced materials and metals, electronics and semiconductors, road vehicles, pharmaceuticals, food processing, and so on.

The existing Database Directive, adopted in 1996, was intended to support the legal protection of databases on a sui generis basis (i.e. having unique value because of their nature and the large investment needed in assembly and organising them) and on a copyright basis. Information on the functioning of the Directive is sparse and its value is unclear. A public consultation to better understand how it is used closed in August 2017.

More broadly, it is now increasingly clear that the current legal framework around data and its use in the EU is inadequate to serve the needs of the data economy. Current EU legislation does not adequately address how to facilitate operations involving sharing and transfer of data, or provide the appropriate safeguards that are necessary. This legal uncertainty undermines the concept of the DSM and implies a significant impediment to the uptake of data sharing in the EU, which will limit both business development and innovation, a situation addressed by the European Commission in its 2017 Communication, *Building A European Data Economy* (European Commission, 2017b).

1.3. Who Owns the Data Today?

Any discussion on the use of data needs to begin with its ownership. This depends on what the data is, how it was generated, what devices were used and where it came from. Broadly speaking, four categories of data can be distinguished, which vary by owner, its use and value. These are:

- The state – traditionally the largest producer and owner of data
- The citizen who both generates and, in theory, owns their data created by their identity and sometimes by their actions and inactions, depending on Member State (e.g. personal image, in France);
- Data producing and moving companies, which includes the internet service providers, telcos, utilities, retail chains, banks, financial services and insurance data);

- Third party aggregators, including credit ratings agencies, and particularly internet companies such as Alphabet (Google), Amazon, Apple, Facebook, Microsoft, whose platforms gather enormous volumes of data.

Table 1 outlines the data generated in these four categories, its uses and places an indicative value on them.

Table 1: Data Owners, data Use and Value

Data owner	Data Generated	Data Use	Data Value
The state[a]	Public sector data: • Citizens' data • Land registry and infrastructure data • Central and local government operations for all ministries and departments • Public datasets from government agencies, e.g. broadband coverage	• Operation of the Member State • Support for citizens, e.g. health records, education • Tax harvesting • Economic planning • Social planning • Infrastructure planning, e.g. flood protection, housing	• Complete range: minimal to very high
The citizen	• Identity • Personal consumption • Internet use • Location, habitat, class • Preferences, politics • Employment and pay • Family and friends • Health data • Finance and insurance • Assets ownership, e.g. car • Physical image	• Interact with state and administrative bodies for identification (eg birth registration), taxation, voting, vehicle licensing, etc • Interactions with lifestyle entities – employer, bank, health services, utilities, consumer goods, supermarkets, retailers, etc	• Very high • High • High/medium • High • High/Medium • High • Medium • High • High • High • Medium except for ID
Commercial data producers	• Manufacturing data • Services sector data, e.g. banking data • Collected datasets by subject, e.g. shopping trends	• Marketing and sales • Production • Accounts and financials • Logistics: inbound supply chain/outbound distribution chain	• Range: low to very high • Range: low to very high • High/Medium
Third party data aggregators and resellers[b]	• Consumer profiles with: - Identity information and personal lifestyle, web surfing history, etc - Product preferences - Consumer trends - Images and videos generated by consumers • Financial market data aggregators and resellers	• Resale of data to: • Advertising industry for personalisation • Market research generally • Retailers • Product shaping and placement for goods and services providers, e.g. consumer products	• High in very large volumes • High/medium

Note: a. including public sector organizations; b. including internet service providers (ISPs), platforms, web retailers and search engines.

Among the four groups it is evident there is an imbalance – commercial data producers and aggregators and processors monetise the data, while citizens and public services give away data. Citizens provide their data (e.g. profiles from usage data) for free in exchange for a variety of services (e.g. search and social networking services). This dominance in the control over data, aided by network effects, strengthens the position of web platforms and others when dealing with their wholesale customers, i.e. advertisers, content providers, product suppliers and retailers.

1.4. The Current State of Data Sharing and Development of Data Markets

1.4.1. The EU's Data Market Today

The value of the EU market for the exchange of data-related products or services was estimated at almost €60 billion (2% of GDP) in 2016, and could grow to €106 billion by 2020 (4% of GDP) (IDC and Open Evidence, 2017).

A large and growing number of organizations serve this market, with over 250,000 data companies in the EU (i.e. organizations whose main activity is the production and delivery of data-related products or services), which could grow to 360,000 by 2020. These organizations employed 6.1 million data workers in 2016 (i.e. workers collecting, storing, managing and analysing data as their primary activity), which could reach 10.4 million by 2020 (IDC and Open Evidence, 2017).

1.4.2. Most Well-developed Data Markets are Commercial

The value of information of a non-personal nature over the centuries has been exploited for gain, in politics and war as well as in business. But today's trade in wholesale databases has drastically increased that trend, for instance in the share and bond market movements for all listed securities in all global markets. This has resulted in data markets for an assorted range of subject data, which is paid for and whose collection may require payment, e.g. a feed from the share trading floor of a market, or from a trading platform, if permitted.

A number of data marketplaces operate currently, either focused on the trading of predominantly personal data for marketing purposes or specialised in trading specific data types generated within one sector, or a single market or environment. There are also commercial market intelligence data resellers on everything from processed food to sales of robots, but few of these offer full access to the raw, unprocessed data for re-use. Few examples have been found of independent two-sided marketplaces with platforms offering supplier and user companies full access to the market and trading data on negotiations and pricing across the different sectors (European Commission, 2017b).

Naturally the data market is complemented by a wide range of software publishers offering data mining analytics (e.g. Microsoft, Pyramid, IBM, Oracle, SAS, SAP, etc), aimed at the business intelligence market, while the large internet platforms (e.g. Amazon, Google, and Facebook) and the financial industry leaders may prefer to invest in their own bespoke tools development.

1.4.3. Examples of Open Sharing of Data are Rare in Today's Market

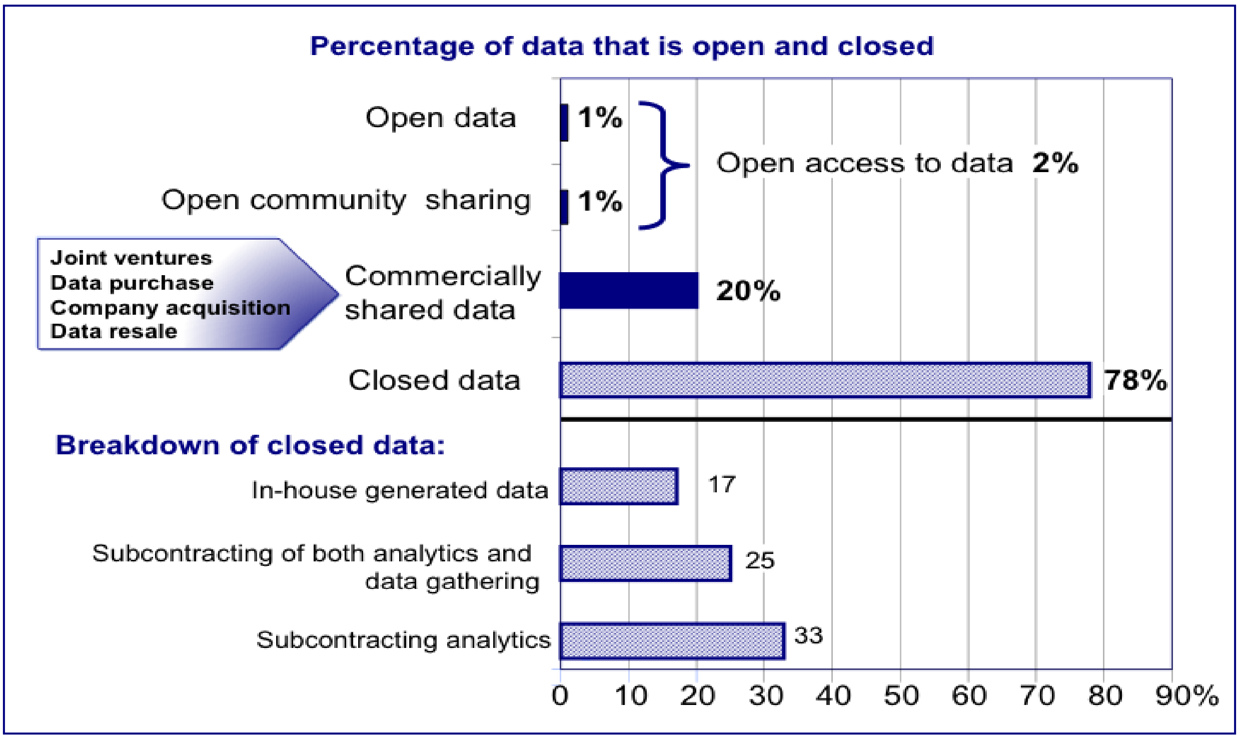
Data is also available as so-called open data, that is data being shared with no or few restrictions, either for public benefit (e.g. humanitarian intervention and disaster management) or corporate social responsibility, or with the prospect of cooperative benefits in a commercial arrangement.

A typical public sector open data source may be published (e.g. electoral rolls with names and addresses). Categorising or refining of such data may add value, providing a resale opportunity. The Directive on the re-use of public sector information (2003/98/EC), defined the rules around unfettered database re-use. Also, note that such data may be in the public domain but not come from the public sector, for instance open street maps in some Member States, which rely on citizens sharing information.

Much rarer, but possible, is open sharing by commercial data owners, sometimes referred to as data philanthropy. They may see value in opening their databases to further processing. This may be the case in banking, mobile telecommunications and even information services (European Commission, 2017c). These include a number of companies who offer open access via an Application Programming Interface (API) for simplified database access by third party applications usually for non-commercial purposes, e.g. Elsevier.[1]

How far this will develop is as yet unclear. As yet there are only a few examples of what are effectively commercial enterprises sharing private data within the definition of open data. In the utilities sector, for example, such data sharing is based on the recognition that there is a case for wider access to such data to help suppliers, other utilities and industry bodies. Some data portals are also being developed to encourage the discovery of data published by non-public actors and to allow collaboration of such data. Evidence from a survey of business models (European Commission, 2017c) shows that business models based on data are still emerging and constantly evolving. This makes it difficult to present a stable picture of the state of data sharing in Europe across all sectors. However, research being conducted on behalf of the European Commission has produced a first classification of the distribution of data-sharing models. Figure 1 shows that only two percent of shared data today is open.

Figure 1: How Much Data is Open Today?



Source: Authors, based on European Commission (2017c).

Some sectors may have a strategic interest in having their data embedded in as many combined datasets applications as possible, as a form of advertising. For instance, in the transport and logistics sector, public publishing of regular passenger schedules or shipments can be combined with other data to act as a form of marketing. Citizens may also participate with time given to collect and store data. Community-driven initiatives like OpenStreetMap build on citizens volunteering to record data and share it to construct joint mapping displays. In the health sector, collective use of statistical patient data on illness conditions, symptoms

[1] https://www.elsevier.com/about/our-business/policies/text-and-data-mining.

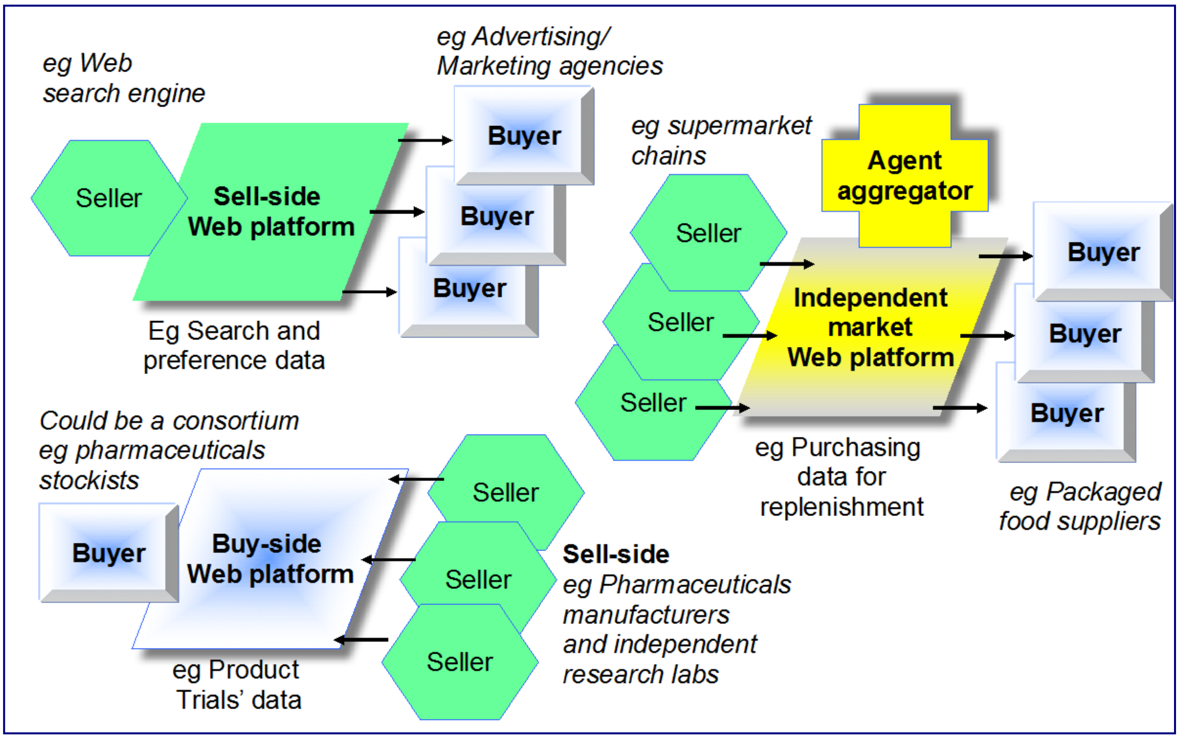
and the degree of success of treatments/procedures/trials is essential to modern public health systems. Figure 1 suggests that in the vast majority of cases (78% of the companies surveyed) data is generated and analysed in-house by the company or by a subcontractor. Vertical integration remains the principal strategy in the sectors surveyed. Thus generally, data tends to stay within an organisation and is not traded with third parties.

1.5. Competition in Open Data Flows

The big players in data do not generally participate in open data flows, the most significant being the web platforms in search services, consumer goods retailing and social networking, whose business model is based on offering services based on harvested consumer data to product and marketing firms. For example, Google's business model is based on applications such as AdWords, its major revenue generator, which enable profiling for targetted advertising through analysis of the words used in search patterns and in email content on its various platforms (Schneier, 2015, Crawford, 2013). However, Google does make some data available through Google Trends and also releases a small amount of their research data[2] but most of its data is used to generate its revenue based on targeting advertising.

Today's data market is defined by data owner, data user and business model. The market mainly comprises consumer data, most usually for personal data aimed at marketing to consumers and for credit scoring them (see Figure 2). Such data is often passed to credit checking agencies by financial services. There is some business-to-business data that may be non-personal in niche areas, for instance airlines share data on their jet engine operations and performance to their engine manufacturers in enormous volumes, as part of their maintenance agreements. All of this implies that open data should generally be considered as in a *pre-competitive* category, requiring further refining through processing to potentially become commercially valuable (see Section 3.1.1).

Figure 2: The Consumer Data Market Today



Source: Authors, derived from European Commission, 2017c.

[2] See https://trends.google.com/trends/, and https://github.com/google-research-datasets.

1.6. Current Legislation for Open Data Flows

The current state of European legislation on open data today is summarised below (Van Asbroeck, Debussche and César, 2017a):

- There is no EU legislation that specifically regulates the question of ownership of data. However, there is significant legislation that has an impact on data or that may confer some kind of protection to certain types of data or on datasets (i.e. copyright, database rights and trade secrets).
- Case law at EU level does not recognise explicitly an ownership right in data. However, according to some legal experts, the 2012 *UsedSoft* decision by the European Court of Justice (CJEU) implies that there is a specific ownership attributed to intangible goods like software downloaded via the internet. Despite this ruling and the possible interpretation deriving from it, legal uncertainty remains.
- Meanwhile, numerous Member States have their own legislation that impacts data ownership and data transfer and may confer some kind of protection to certain types of data and datasets, specifically for copyright, database rights and trade secrets (Osborn Clark, 2016).
- Intellectual property rights, in their various forms of copyrights and trade secrets, are the limits of most ownership-like rights currently available in the EU. However, none of these is likely to provide suitable rights for ownership of data.
- Copyright is also inappropriate for datasets as it requires originality of creation, has territorial limits, often time limits and is exclusive (the work’s authors). Nevertheless, copyright offers some features beneficial to protection of data, for example, disclosure of data is possible, as is sharing by permission, and broad exclusivity rights. However, these features are unlikely to be sufficient regarding open data.
- Trade secrets’ protection, in general was not created to give the comprehensive protection of entire datasets. Moreover, it mandates that the data remain a secret (and cannot be widely distributed without let or hindrance).
- There is debate on whether personal data is owned by the individual, so that an "ownership" right to data by appointed data controllers or processors cannot be excluded. But such ownership could be subject to the individual's control over their personal data. This varies greatly by Member State; in some it is clearer, principally France and Germany.
- While protection of individuals' privacy has advanced in definition with the adoption of the General Data Protection Regulation, the issues related to privacy and data protection have not been considered in depth by the Commission in relation to open data. Although the Commission’s 2017 Communication offered an approach to ownership of data, be it non-personal or personal data, this is absent in the proposal for a Regulation of September 2017 (European Commission, 2017b).

2. A FRAMEWORK FOR OPEN DATA ACCESS

Given the current lack of clarity, what should be done regarding the status of digital property and data ownership? Building on and extending the Commission's 2017 Communication, *Building European Data Economy*, this chapter considers the objectives of a framework for open data access, and how they might be achieved.

According to the 2017 Communication (European Commission, 2017b) a future EU framework for open data access would contain the following objectives:

- **Improve access to anonymous machine-generated data**: Through sharing, reuse and aggregation, machine-generated data becomes a source of value-creation, innovation and diversity of business models.
- **Facilitate and incentivise the sharing of such data**: Any future solution should encourage effective access to data, taking into account, for example, possible differences in bargaining power between market players.
- **Protect investments and assets**: Any future solution should also take into account the legitimate interests of market players that invest in product development, ensure a fair return on their investments and thereby contribute to innovation. At the same time, any future solution should ensure a fair sharing of benefits between data holders, processors and application providers within value chains.
- **Avoid disclosure of confidential data**: Any future solution should mitigate the risks of disclosing confidential data, especially to current or potential competitors. Thus, any solution should enable proper data classification to be performed first, prior to the assessment of whether or not specific data can be shared.
- **Minimise lock-in effects**: The unequal bargaining power of companies and private individuals should be taken into account. Lock-in situations, especially for SMEs and start-ups and private individuals, should be avoided.

The Commission also believes that safeguarding procedures are needed for open access to machine-generated data, with varied levels of intervention:

- **Data producer's right**: A right to use and authorise the use of non-personal data could be granted to the "data producer", i.e. the owner, or long-term user (and lessee) of the data is the Commission's view. The aim would be to clarify the legal situation while giving more choice to the data producer. It should open up access for other users to exploit machine-generated data. Any exceptions in the dataset would need to be clearly specified, in particular the grant of non-exclusive access rights to the data by a range of possible users, on a non-discriminatory basis**.**

- **Personal data:** would need to be rendered anonymous, so the individual is unidentifiable before use by any other party. We note that the General Data Protection Regulation (GDPR) continues to apply to any personal data (whether machine generated or otherwise) until it has been anonymised. However, significant problems with anonymisation are becoming apparent (see Box 2) as more advanced techniques of data mining on large datasets are now available (Hern, 2017). Just as encryption is being broken by faster and more powerful parallel machines, so is anonymisation. There is also the stipulation that if personal data is involved with the person's explicit permission, then the individual would retain their right to withdraw consent at any time after authorising its use.
- **Incentivising businesses to share data**: the Commission may issue guidelines on how non-personal data control rights should be addressed in contracts. The aim would be to mitigate the effects of divergent national regulations and provide increased legal certainty for companies. Guidance would be based on existing legislation, in particular

the transparency and fairness requirements laid down by EU marketing and consumer law, the Trade Secrets Directive and copyright legislation, notably the Database Directive.

- **Encourage development of technical solutions for reliable identification and exchange of data**: Traceability and clear identification of data sources are a precondition for real control of data in the market. The definition of standardised protocols for reliable, persistent identification of data sources should create trust in the system. APIs to access the data in standard ways could also help firms and public authorities to identify and profit from different types of re-uses of the data they hold. APIs could foster creation of an ecosystem of application and algorithm developers interested in the data held by companies. Broader use of open, standardised and well-documented APIs should be pursued, with technical guidance and publication of best practice. Guidelines on making data accessible should include machine-readable formats and definition of the associated metadata.
- **Default contract rules**: For contracts relating to data, default rules could be applied. These would define a balanced set of usage conditions, and include the Fitness Check which is part of the Unfair Contract Terms Directive (93/13/EEC). These rules should also be linked to an unfairness control in business-to-business contractual relationships. That control would invalidate contract clauses that deviate excessively from the default rules. This should also be complemented by a set of recommended standard contract terms, designed by stakeholders (data users and data suppliers). This approach should be aimed at lowering the legal barriers for small businesses to use open data and reduce the imbalance in bargaining positions (while still allowing a significant degree of contractual freedom).
- **Access for public interest and scientific purposes**: Exceptions for public services could be granted, to give access to data where this would be in the "general interest" and would considerably improve the functioning of the public sector, for example, access for national statistics offices to business data, or the optimisation of traffic management systems on the basis of real-time data from private vehicles. Another possibility is in public health systems where statistics on effects and side effects of drugs, or surgical procedures could form a guidance database for medical practitioners. Access to datasets and then the powers to combine the data from different sources is critical for scientific research in fields such as medical, social and environmental sciences.
- **Access to data for a remuneration**: fair, reasonable and non-discriminatory (FRAND) terms could form a legal framework developed for data holders, such as manufacturers, service providers or other parties for open access to their data. That would provide certain key principles, such as, to provide access to the data they hold against remuneration after anonymisation. Relevant legitimate interests, as well as the need to protect trade secrets, would need to be taken into account. The consideration of different access regimes for different sectors and/or business models could also be envisaged in order to take into account the specificities of each industry.

These overall objectives are summarised in Table 2 with potential ways forward to achieve them and the types of measures to be taken, i.e. legislative or non-legislative measures.

Table 2: Objectives for the Future and the Types of Measures Required

Objectives		Possible Ways Forward
Improve access to anonymous machine-generated data	Non-legislative measures	Guidance on incentivising businesses to share data
Facilitate and incentivise the sharing of such data		Fostering the development of technical solutions for reliable identification and exchange of data
Protect investments and assets		Model contract terms
Avoid disclosure of confidential data	Legislative measures	Default contract rules
Minimise lock-in effects		Access for public interest and scientific purposes
Clarify who owns what, and who is a data producer, e.g. a person's image is owned by the person; consumer owns their identity and consumption data		Well-defined data producer's rights
Enforcing right to remuneration of consumers for use of their data		Access must be set against remuneration
Enforce disclosure of a) holding personal data; b) use of personal data		Make part of data producer's rights and data user's obligations

Source: Authors; European Commission, 2017; Van Asbroeck, Debussche, and César, 2017b.

3. OVERCOMING BARRIERS TO OPEN DATA

There are many different barriers to building a European data economy based on open data, which may be technical, socioeconomic or legal in nature (EuDeCo, 2015). This chapter discusses the most significant issues and ways of overcoming them.

3.1. The Value of Data

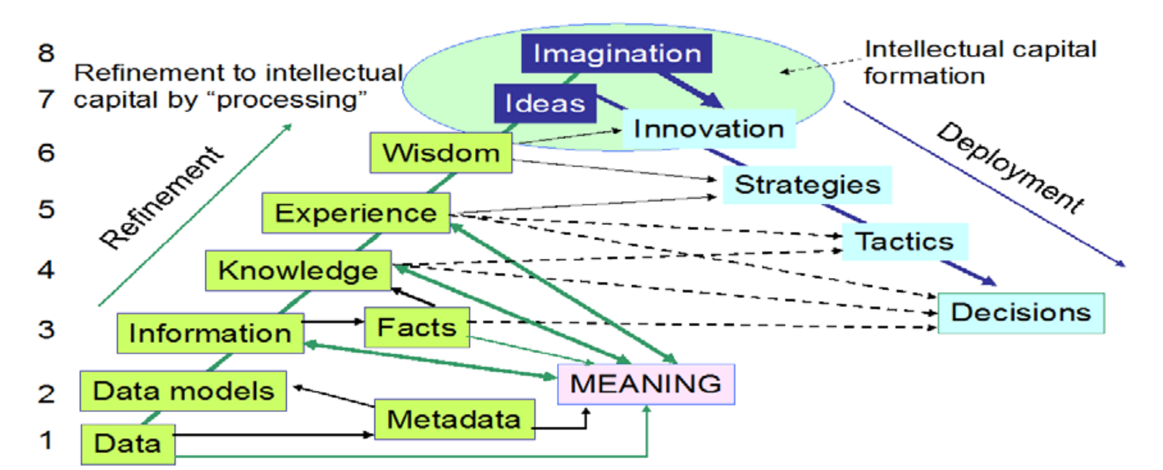
Although information has become a strategic asset for many organizations, all too often it is considered as intangible and not valued, for example, it does not appear in company balance sheets. It is clear that widespread sharing or trading of open data will be held back unless a way is found to value datasets and information-based assets. But how can this be achieved?

3.1.1. Can Commercially Useful Data be Valued?

At a simplistic level, one might say that if there is a market for data, then data will find its value in the marketplace. For instance, according to the *Financial Times*,[3] "general information about a person, such as their age, gender and location is worth a mere $0.0005 per person, or $0.50 per 1,000 people. A person who is shopping for a car, a financial product or a vacation is slightly more valuable to companies eager to pitch those goods." The latter illustrates two fundamental points: the value of data depends on how rich it is and also on what you want to do with it.

One way of attributing value is by applying the concept of intellectual capital to encompass all forms and levels of raw intelligence with material value. This is a much wider concept than intellectual property rights (IPR) and patents and includes data, information, knowledge and wisdom. Applying effort in accumulating and organising lower level material increases its value, so it moves up the value scale. In this way, a hierarchy of value can be identified in the sense that refining lower levels of intellectual capital give richer results, which have more value (see Figure 3).

Figure 3: The Value of Data in the Hierarchy of Intellectual Capital Assets



8. Imagination – creative thinking using new ideas and generating new ideas and innovations
7. Ideas – new thinking for defining concepts which (re)set the company's directions
6. Wisdom – long term accumulation of knowledge with deductions (on operations and behaviour)
5. Experience – accumulation of knowledge into useful guidelines (eg for running the business)
4. Knowledge – processed information which enables decisions; may be encapsulated in IPR
3. Information – the meaning of data, which forms facts
2. Data models – data structure with meaning of elements such as the record fields defined by metadata
1. Data – raw unstructured, or structured, perhaps as records with fields, without cognisance of their meaning

Source: Forge, 2004a.

[3] http://ig.ft.com/how-much-is-your-personal-data-worth/#axzz2z2agBB6R.

3.1.2. Factors Affecting the Value of Data

When considering the nature of data, analysts commonly refer to five attributes – the five "Vs": Volume, Variety, Velocity, Veracity, and Value. In essence, we can think of the first four as factors that affect the value of data.

- Volume: High-power computer processing combined with cost-effective data storage means that broader and deeper analysis of data across different data dimensions and multiple years of historical context is possible. Data no longer needs to be sampled as entire datasets can be analysed. In other words, the more data, the better in terms of accuracy of results and insights.
- Variety: As it is now possible to analyse data combined from diverse sources, in the digital economy, acquisition of varied data becomes more valuable.
- Velocity: Having the most up-to-date information has always been important in providing competitive edge, but is becoming even more so. Data that can be analysed in real time is particularly valuable.
- Veracity: The more trustworthy the data, the more valuable it will be. However, this aspect requires greater judgement than other factors because it requires understanding of biases, noise and abnormalities in the data. Assuring veracity also requires measures to protect open data from malicious manipulation by *addition* of false data (e.g. for introduction of new biases) and *falsification* of existing data, so that it becomes faked in some way. Checks for integrity are traditionally used but more advanced error correction techniques may not only detect missing and changed binary coded data, but in some cases, can repair it.

3.2. Data Localisation

Cross-border data transfer is needed for trade in services as well as in manufactured goods. Jurisdiction over data, i.e. restricting where data may be stored or processed, is a significant impediment to the free flow of open data. Data localisation rules are on the rise around the world, fragmenting data flows, which increases costs for businesses, especially for startups seeking to scale, and consumers. A recent study identified 22 data localisation measures where European Union Member States impose restrictions on the transfer of data to another Member State (Bauer, Ferracane, Lee-Makiyama and van der Marel, 2016), with restrictions most often concerning company records, accounting data, banking, telecommunications, gambling and government data. The study also identified more than 35 restrictions on data usage that could indirectly localise data within a certain Member State.

The European Commission considers that removing these restrictions is the most important factor for growth in the data economy, and a public consultation found broad support for its approach and removing restrictions via a legislative instrument.[4] In consequence, in September 2017, the Commission proposed a new set of rules to govern the free flow of non-personal data in the EU by, among other things, limiting the scope of data localization requirements currently imposed by Member States. In summary, the draft proposal puts forward measures which:

- Reduce the range of restrictions for data localization;
- Enhance legal certainty;
- Facilitate the availability of data on a cross-border level;
- Improve the conditions to switch data storage for users or port data back to IT systems for service providers; and

[4] Synopsis report of the public consultation on building a European data economy, https://ec.europa.eu/digital-single-market/en/news/synopsis-report-public-consultation-building-european-data-economy

- Reinforce the trust and security of cross-border data storage and processing.

Under the proposal, Member States would be required to notify the Commission about any draft legislation or measure introducing new data localization requirements or any plans to repeal existing national data localization requirements within a year of the application date of the draft proposal. The proposal also encourages service providers and professional users to develop codes of conduct detailing the information on data porting conditions.

Regarding future cooperation, Member States will have to designate single points of contact to cooperate with the Commission on the application of the Regulation, and a "Free Flow of Data Committee" would also be established.

It is too early, of course, to assess how the proposal will be received, but we would make the following observations:

- Surprisingly, the proposed Regulation does not define the meaning of "open data" (see our definition in Box 1).
- The proposal is concerned with non-personal data, but seems to assume that open data will be free from personal data. However, as we have explained, the distinction between personal and non-personal data is often blurred and it may be necessary to include a mechanism to check and guarantee that personal data is excluded.
- It seems to apply largely to government data already in the public domain and not necessarily for privately held data that might be shared under specific conditions, which is really what would unlock the data economy.
- The Commission will have a Free Flow of Data Committee to advise it but its aims and terms of reference are not specified.

3.3. Competition Policy and Open Data Flows

3.3.1. Does Control over Data Breed Data Monopolies?

There is an ongoing debate among academics and policy analysts as to whether control over data confers unfair competitive advantage in the digital economy.[5] Some argue that the exploitation of the large amounts of data that web platforms are able to gather amounts to an abuse of a monopolistic position in the market, and that those who have control over large amounts of data should be required to share it (Prüfer and Schottmuller, 2017). According to the *Financial Times*, Japan's Fair Trade Commission is considering regulation to prevent "digital monopolies" abusing their market position by amassing data (Harding, 2017). Measures being contemplated include blocking mergers that will increase data monopoly, prosecuting companies that keep data from competitors, and banning social, shopping and search platforms from collecting extraneous additional consumer data beyond consumers' direct requirements.

Others contend that, by itself, big data is unlikely to be valuable, as there are many alternative sources of data available because users leave multiple digital footprints on the internet (Lambrecht and Tucker, 2015). The brief history of the digital economy offers many examples, such as Airbnb, Uber and Tinder, where insight into customer needs enabled entry into markets where incumbents already had access to large amounts of data. Rather than simply amassing data, building sustainable competitive advantage in the data economy depends on managerial competence and the technical ability to analyse and interpret data

[5] See, for instance, CEPS Digital Forum workshop on "Competition policy in the digital economy: towards a new theory of harm", 1 June 2016, https://www.ceps.eu/events/competition-policy-digital-economy-towards-new-theory-harm.

and deliver an attractive customer proposition. In other words, it’s not about the data but what you do with it!

The way in which digital technologies are evolving are undoubtedly challenging some fundamental tenets of competition law and policy, and it is becoming increasingly difficult to distinguish anti-competitive behaviour from normal business strategies. It would be wise to be cautious as the debate on competition and harm in the digital economy is still unresolved.

3.3.2. Making Open Data Accessible Needs Re-usability Mechanisms

A competitive open data economy requires standards: re-usability depends on readability. The actual degree of readability of public data is fairly easily measured using the government dataset sources within the EU such as the European Data Portal (2017). The two key requirements for open data re-use are machine-readability and the suitable formats or distributions used to present the content of the data. The latter should be in one of the more common distribution formats.[6] If the data is unreadable, except by some non-standard conversion to a standard format that is then readable by common applications, it is effectively encrypted, and so should not be considered as open data.

3.4. Open Data Flows Pose Some Threats to European Society

Open data flows do pose some threats today, the most significant being:

- The non-personal nature of open datasets is difficult to guarantee
- Open data sets could be a highly attractive target for economic sabotage

3.4.1. Mixing Personal and Non-personal data

If open data flows are pursued, there are clear dangers over the possible lax use of personal data, which is often consumer related. Internet companies, such as Amazon, Google, Facebook, Netflix, Microsoft, etc, claim that with a blending of data points without personal attributes the data that was generated by an individual person becomes one sample in enormous fields of zetabytes and cannot be attributed. This is somewhat disingenuous, as German researchers revealed at the August 2017 DEF CON conference (Hern, 2017).

In consequence, there are many ways in which open data could be “contaminated” by personal data, either in error or by further processing and combining with other datasets to *deanonymise* it. Therefore, it will be necessary to assure citizens about how their data that may be used. This is particularly important in those sectors where electronic personal record-keeping is common, which includes health services, for medical records, and financial services. These are however two quite different cases.

The GDPR would protect stored personal details, and transactions for financial services, which should prevent any sharing of such data, while anti-fraud legislation would also apply, e.g. for identity theft. Medical records would also fall under this category with privacy and security conditions, with two possible exceptions. First, sharing medical records between medical institutions and practitioners, for care of the patient by several institutions, implies that *closed* medical data systems are required with highly secure interfaces. Second, for the statistical analysis of bulk medical records, e.g to analyse side effects of a prescribed treatment, open data categories might be allowed if it were restricted to bona fide medical research users for which it was intended with decoupling of personal details to provide some protection for anonymous trials, with data deletion afterwards.

[6] Searching data on the European Data Portal, shows over 49 different file formats are used (European Data Portal, 2017).

Box 2: "Anonymous" Browsing Data Can Be Analysed to Identify an Individual

German researchers presented findings at the DEF CON hacking conference in Las Vegas in August 2017, demonstrating that "anonymous" browsing data can be easily analysed for user identification. Using what was described as "an anonymous data set" of three million German citizens' surfing history over 30 days, it was easy to de-anonymise many users. For instance, just 10 URLs can be enough to uniquely identify someone with modern data mining tools. The researchers described various methods to find an individual in the noise, just using a long list of URLs and timestamps. Some sites even guide the search directly to the individual by storing identifiers with the website URL. A more probabilistic approach can de-anonymise most of the other users.

It was also noted that this is common in the industry. In one example given from 2008, de-anonymisation of a set of ratings from Netflix was used to help its researchers improve the recommendation algorithm then in use. It compared "anonymous" ratings of films with public profiles on IMDB. Researchers were able to identify Netflix users – including one woman, who sued Netflix for the privacy violation. Another route discovered for data collection was Google Translate, which stores the text of every query put through it in the URL. From this, the researchers were able to uncover operational details about a German cybercrime investigation, since the detective involved was translating requests for assistance to foreign police forces. The researchers stated that in general they considered personal data anonymisation was nearly impossible. Much of the data used in the research came free, from browser plug-in suppliers, such as the "safe surfing" tool *Web of Trust*, which sells the data it collects on its users, but has not been telling them.

3.4.2. Open Data Flows May Act as Cyber-Attack Vectors

At a macro-economic level, cyber-attacks are now being targetted on economic disruption by nation states posing advanced persistent threats (APTs). The most advanced today use common business intermediaries such as accountancy software, law firms and other business services. Such targeting could make open data flows into a subtle vector for ransomware and intentional database destruction, for disruption of business activities inside the EU, an economic threat that some nation states seem willing to pursue.

For example, the Petya cyberattack in the Ukraine in June 2017 quickly spread via an update to a popular accounting package from Ukrainian business software vendor MeDoc. The APT was targetted at large closed business networks – not necessarily with internet access. It was designed to destroy databases, not just collect ransom, and so sabotage businesses and the economy, specifically that of the Ukraine. However, there was much collateral damage – among the many multinational victims of Petya were the makers of Cadbury chocolate and Oreo biscuits with lost invoicing and shipping, WPP, the world's largest advertising group, and Maersk, the world's largest shipping operator. Effective protection for open data from malefactors embedding malware of various forms in open data is beyond the scope of this paper but will need to be carefully considered.

3.5. What Data Protection and Privacy is Needed?

The key questions are what conditions need to be met, and what security is necessary, in view of the intention to have open data? A fundamental dimension is whether data is public or private, i.e. non-attributable "averaged" data versus personal profiles.

Baes on our analysis, for personal data to enter the open data arena in any form, however "anonymised", the only way is to manage it for the benefit of its original owners. As has often been pointed out, that implies that citizens as consumers should own the rights to their own data (Robinson, 2017). They would have to explicitly give permission for its use and receive payment in return. Also, there would have to be minimal safeguards:

- First, in the anonymisation procedures, i.e. that reach standard levels of guaranteed anonymity and that standard would have to be of a level that complies with Article 17 of the GDPR on the right to be forgotten.
- Second, regarding the use to which the data is put. For example, citizens might agree to share their “anonymous” health data for health research or public purposes, but not for insurance purposes or for other purely commercial uses. That means that the unrestricted use and openness of a dataset would be severely compromised, but that would be the penalty for attempting to use personal profiling data. Judgements on the use of open data that could include personal data, however remotely, is needed.

4. FUTURE SCENARIOS

4.1. Defining How Data May Be Used

Five key questions need to be considered for an ideal solution:

1. What rights should the various actors have?
2. What differences in rights are sensible in terms of the powers these rights bestow?
3. What is ownership and what is access?
4. How should personal and non-personal data be treated, especially if they are mixed through de-anonymisation?
5. Should there be freedom to use, on some form of payment either monetary or in kind (i.e. data access in return or access to the processed results) to those who offer their data? Should payment be commensurate with market pricing, i.e. on a FRAND basis.

Several different strategies have been suggested to address these questions. Below, we summarise three current views, plus we add a further potential solution drawn from existing practices in the IT market.

4.1.1. Do Nothing: the Legal-libertarian View

A study by the European Commission's Joint Research Centre (Duch-Brown, Martens, and Mueller-Langer, 2017) concluded that there are, at present, no compelling economic arguments for regulatory intervention. Doing nothing is also supported by Drexl (2016), who argues that the existing toolkit of trade secrets' protection, contract and technological protection measures offers data producers ample means of securing de jure or de facto exclusivity in "their" data, if they so desire.

Note also that the European Court of Justice has clarified that copyright cannot apply to databases per se but only to the data. The general rule is that copyright requires acts of human authorship (Hugenholtz, 2017). The Court held that investing significant amounts of skill and labour to collect the data does not justify a finding of originality. This originality criterion removes any copyright protection for a database founded only on effort and investment to arrange the data. The collection activity and database structure is not copyrightable. This decision ends prior contrary rulings in the Member States, notably the Netherlands and the UK.

4.1.2. Use Current Contract Law

An alternative would be to rely on contracts as the legal framework, the approach proposed by the Commission in its 2017 Communication:

> Therefore, comprehensive policy frameworks do not currently exist at national or Union level in relation to raw machine-generated data which does not qualify as personal data, or to the conditions of their economic exploitation and tradability. *The issue is largely left to contractual solutions. The use of existing general contract law and competition law instruments available in the Union might be a sufficient response.*

However, the Communication follows on to note:

> In addition, voluntary or umbrella agreements covering certain sectors might be envisaged.
>
> Nevertheless, where the negotiation power of the different market participants is unequal, market-based solutions alone might not be sufficient to ensure fair and innovation-friendly results, facilitate easy access for new market entrants and avoid lock-in situations.

This implies that contracts can be a solution and sectoral level agreements are a way forward. However, while this might seem to provide the necessary simplicity and flexibility, this path has significant obstacles, including:

- The lack of harmonisation of contract law across the EU,
- Issues of validity of data-related agreements in the various Member States
- Limits of contract validity in arrangements with respect to third parties.

Moreover, it is difficult to see how sectoral agreements would not arrive at a patchwork of data flows, with imbalances in the conditions of use, openness, rights to the derived data and protection from contamination by personal data – implying intervention under the GDPR.

Relying on contract law is clearly far from perfect. Ideally an open data framework would apply across the EU, and not just operate by sector with specific conditions. It must protect the citizen and be efficient, with use of a provable provenance to distinguish personal data from open, exchangeable non-personal data. Thus, a solution is needed which is legally secure, yet efficient and universal – a new framework is needed.

4.1.3. A New “Right in Data”

The rise in the data economy has been accompanied by calls from some industrial quarters, academics and lawyers to introduce a novel property right in data, a concept also tentatively floated in the 2017 Communication. Such a new right would have the following features (Van Asbroeck, Debussche, and César, 2017b):

- it would be a right in individual pieces of data that will naturally extend to the entire datasets which those individual pieces of data are part of;
- the right would be non-exclusive;
- the right would be paired with an obligation of compulsory transfer of the data, subject to certain conditions; and
- the right and transfer obligation would be supplemented and enabled by a traceability obligation.

Whereas database right protects data on the double condition that the data are structured in a “database” and the database is the result of “substantial investment”, any new right would directly protect machine-generated data without any material prerequisite. A “data producer’s right” would also go beyond any protection currently offered by EU copyright law.

While such an approach may have some attractions, it may not be practical as there are some critical uncertainties left to be defined:

- The limits and rights of the proposed “non-exclusive ownership” of datasets are not yet clearly defined. They seem to imply use of a dataset by a closed user group, on a specific agreement between the data originator and those who wish to use the dataset. Perhaps, with time, that could be developed into clear divisions of what is permitted and what is excluded in the concept, but the proposed legal framework is still to be defined. It seems to propose access rights rather than an ownership and associated access rights model.
- It is recommended that its legal basis depend on soft law (codes of conducts, guidelines that the European Court of Justice would follow and extend). In the cut and thrust of commercial business that is unlikely to be enough to be respected, as it depends on a court’s future ruling and decision processes may be protracted.

More fundamentally, however, introducing a “data producer’s right” to protect the data in industrial databases from unauthorised access to its data would seem to directly contradict one of the Commission’s goals for the DSM. Moreover, Hugenholtz (2017) argues that introducing such an all-encompassing property right in data would also undermine the current European system of intellectual property law, contravene fundamental freedoms enshrined in the European Convention on Human Rights and the EU Charter, distort freedom of competition and freedom of services in the EU, restrict scientific freedoms and generally undercut the promise of big data for European economy and society. Having said that, two useful concepts might be borrowed:

- Access to one dataset element or a subset implies access to the whole dataset – but this does require definition of what is excluded from the dataset.
- Enforcement of the transfer of data as an obligation (to whom needs to be defined).

4.1.4. Could Practice in the Software Industry Offer a Way Forward?

Over the past fifty years of software development, one model of re-use or sharing of software programmes and data has emerged – open source software (OSS). In contrast to proprietary software, OSS may be defined as software whose source code is published and made available to the public, enabling anyone to copy and use it and, depending on the software licence, to modify and redistribute the source code without paying royalties or other fees (Forge, 2004b). It is free in the sense of “free speech”, rather than “free-of-charge”, as it can be charged for in some models, even if it is a nominal charge. It can also be used to apply licences to content, which may be used for managing content distribution, e.g. Open Content License, OpenOffice.org Public Documentation License, Artistic License, etc.

OSS is considered by many to be a stable and sustainable model of software development and distribution, making the trend towards its use almost inevitable. It has been recognised officially and developed in formalised groups since 1984, but its development goes back to the development in the 1960s of ARPANET (the precursor of today’s Internet). Open source software is already important to the EU economy, since most software anywhere uses either programmes or protocols, including datasets, made available under an OSS licence. For instance, the two key operating systems used today in smartphones, Apple’s iOS and Google’s Android, are both derivatives of open source software. Note that the EU already has its own software licence, for distributing the software from its EU funded research projects, the European Union Public Licence (EUGPL), which is similar to the GNU Public Licence.

4.2. OSS Licences Bring Principles Applicable to Open Data

OSS licences vary greatly in their usage conditions, for instance on what type of software OSS can be linked to or combined with, and the control over derived works. There may be rules on returning OSS code modifications to the software “commons” – the development community – or there may be no rules at all on this. There is wide range of different *types* of conditions (Raymond, 1997).

A form of OSS licence for open data makes sense for anyone contributing to collaborative projects, especially those creating technology platforms requiring large amounts of input data. Companies, institutions and individuals could contribute datasets freely and as peers, secure in the knowledge that they may give to a larger dataset work without it being hijacked by one community member for commercial ends, except under agreed conditions. The forms of open data rights that could be essential are:

- The right to see and access the original dataset
- The right to make copies of the original and store them locally
- Rights to make derivatives, modifications and improvements by analytic processing
- The right to use the datasets under agreed conditions for commercial or public ends
- The rights to distribute derivatives following processing and modifications
- Optionally, obligation to return processed data to the user community of that dataset.

The key principle that the OSS software world turns on is a range of licences to use the original source code. Various kinds of OSS licence are already well accepted in law. Their different types are well suited to managing data flows. What we propose here is an extension of the OSS model for software programming to large datasets. Related objects already exist in the OSS world for certain "information sets", specifically documents and document formats (such as open document format, ODF) and also for formats of data for programmed use where relevant.

4.2.1. The Importance of Backwards Compatibility Over Time

A particularly important OSS area is electronic document formats, specifically for code. OSS offers stable, efficient formats for a period of time chosen by the user, which may be over 50 years for public sector and some business agreements. Most commercial electronic document formats over ten years old cannot be read by today's commercial applications, except occasionally in a limited form, but certainly not across all the different operating system environments. This concept of standard stable formats can be incorporated in open datasets.

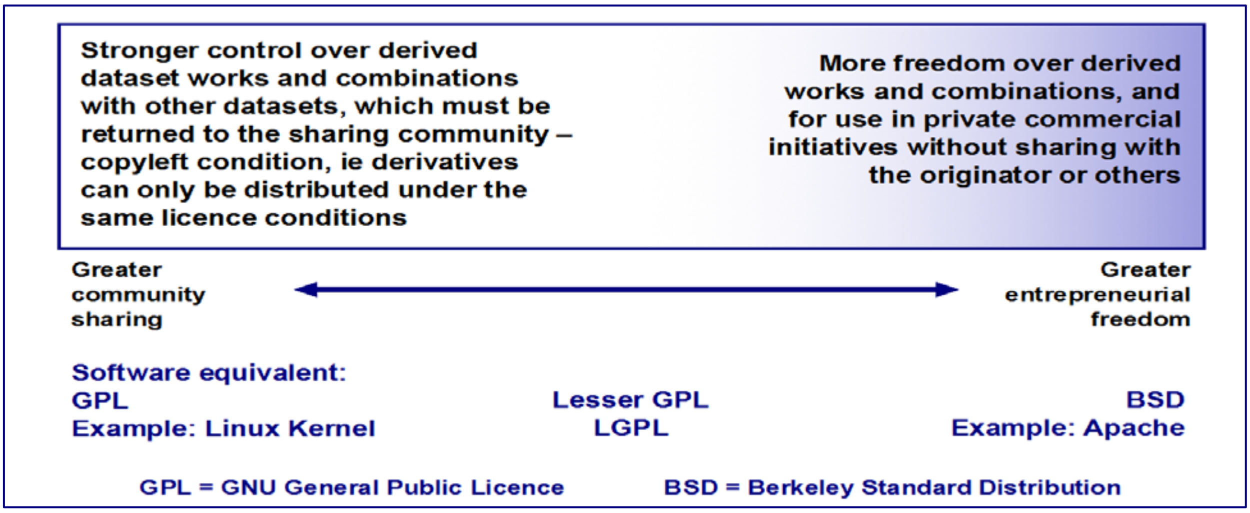
4.2.2. Adding Traceability with Distributed Ledger Technology

However, data is not software. Software is less likely to require a strong effort to record its genealogy and evolution, except in cases of the surreptitious addition of malware, which has yet to be addressed by logging its current actual contents and its evolution path (but could be in the future, to help counter cyberattacks, especially for software updates and email attachments). The problems incipient in data sharing, of understanding its origins and its processing history are necessary to understand whether it is complete, how it has been processed, not just the veracity of its origins. The use of distributed ledger technology (DLT) is an evident possibility.

4.2.3. A Suitable Licence is Available for Every Application

The actual licences used in OSS are highly variable in the degree of freedom to re-use the intellectual capital involved in the source code. Two general types of requirement are included – those that mandate return of the extension or derivatives to the whole user community and those that permit completely free use of the derived work, including the issue of licences as in a commercial product, but still based on the open source software origins in concepts and code. The range for open datasets in general and their usage rights is illustrated below:

Figure 4: The Range of Licences for Open Datasets and Their Usage Rights



Stronger control over derived dataset works and combinations with other datasets, which must be returned to the sharing community – copyleft condition, ie derivatives can only be distributed under the same licence conditions

More freedom over derived works and combinations, and for use in private commercial initiatives without sharing with the originator or others

Software equivalent:

GPL
Example: Linux Kernel

Lesser GPL
LGPL

BSD
Example: Apache

GPL = GNU General Public Licence

BSD = Berkeley Standard Distribution

Source: Forge, 2004b.

In general, when applying OSS concepts to an open dataset licence, there could be of at least five general forms of licence, which would be chosen as required by the originators and further users:

Box 3: Five Types of Open Data Licence

- **Type 1: Access and copy only:** The dataset can only be used as it is presented literally, to a particular new user, without permitting processing or creating derivatives for whoever obtains a licence to access the dataset. The licence may be charged for, or be free, depending on the originator's wishes.
- **Type 2: Private processing permitted – Research licence:** A dataset can be used as the source of a derivative dataset by some form of processing, which can be used by the processing entity who has paid for access with a licence and the derived dataset(s) will be kept for the processing entity's private internal use only and cannot be shared. The licence permits access, copying, local storage and post processing with local storage of the derived dataset within the processing entity. The derived dataset may be mixed with other datasets.
- **Type 3: Shared derivative:** As for Type 2 but the derived set must be shared with the original source entity of the primary dataset. Any party holding a licence, as well as the originator, can use the derivative dataset for their private use only.
- **Type 4: Community licence:** A dataset can be used as the source of a derivative dataset, which can be used by the processing entity and the original entity and also may be passed to any new users who subsequently obtain a licence. New users who then process the dataset or the derived dataset must return their new derived version to the whole user community.
- **Type 5: Commercial exploitation licence:** A dataset can be used as the source for a derived dataset, which can then be stored, accessed and used by the processing entity for its own purposes, without sharing its derived content with the originator or any other entity and it may exploit the derived version commercially for profit.

In the OSS world, licences are examined and approved by ISO, giving them international status through the approvals process by its OSI agency (the Open Standards Institute). If distributing software under an OSI approved license, the OSS source entity is permitted to state that its software is "OSI Certified Open Source Software." The same procedure and sign of approval might be applicable for open datasets.

In addition, it would be possible to further adapt or vary these licences, for example by adding conditions to prevent anonymised open data being combined with other data sets to enable deanonymisation of personal data.

In summary, Table 4 shows how the open data licence approach responds to the various contractual issues.

Table 3: The Open Data Licence Approach Solves Contractual Issues

Contractual Issue	Proposed Solution
The rights of the various actors – data donors and subsequent users	The licences (Types 1 to 5) define the differences in rights that are reasonable and sensible in terms of the powers these rights bestow on dataset users and donors.
What is ownership and what is access to the dataset?	This is the key content of the licence.
How can personal and non-personal data be treated, especially if they are mixed through de-anonymisation?	GDPR and also Member State regulation should enter here. Tests for contamination by personal details in the data will be necessary with confirmation.
Should there be freedom to use, on some form of payment either monetary or in kind (i.e. data access in return or access to the processed results) to those who offer their data? And should the payment be commensurate with market pricing i.e. on a FRAND basis of fair, reasonable and non-discriminatory access and payment levels?	Each of the forms of licence outlined above gives options on payments and can define the rights over re-use of data and return of processed data to the user community, or its retention for private commercial use.

5. RECOMMENDATIONS

5.1. Recommendations for Opening Data Flows to Industry Across the EU

5.1.1. The Aim of the Recommendations

With any set of recommendations, the overall goals and intention should be clear. Here the aim is to support openness and free flow of datasets with no restrictions on access when under specific commercial conditions, to all those parties who should be given a right of use to the dataset, and to process it for their own ends. The recommendations should cover various issues, in particular, how to implement an environment for open data sets to be simply and freely exchanged, which requires:

1. The contractual forms needed between open data sharing parties
2. The data protection and privacy safeguards necessary
3. The security needed, in view of the intention to have open data, and that it can be stolen or have malware embedded in it, or be corrupted
4. The data management necessary to protect the dataset's security and privacy, through a governance and operational framework. That framework calls for various preparatory actions:
   - Creation of the regulation necessary for enforcement of rules of use and storage as a protective legal environment for access and exchange – it should cover the security and privacy of datasets as well as the contractual licences between donors and users.
   - Measures to ensure exclusion of personal data, following the GDPR on privacy of personal data and the right to be forgotten, and licence conditions to prevent contamination of open data with personal data.
   - Data traceability obligations, as a safeguard (which can include the use of various forms of distributed ledger technology (DLT), where practical).
   - Creation of functional guidelines for physical infrastructures for the exchange of open datasets (i.e. guidelines on the forms of database and networked, especially for cloud-based datasets, to ensure protection).

5.2. Recommendations to Provide Answers with Practical Measures

First, a framework for governance of open datasets that use open data licences may be necessary with the components being:

- *A body to hold open dataset IPR in a commons at a European level*, so that any subsequent recourse is to that body, not the data donor. The body might be financed by the vertical sectors involved and its associated users.
- *A reference base and possibly a documentation repository* for European enterprises, with a set of templates for datasets that includes the standard data formats and a suite of standard licences acceptable in courts across the EU. It might also hold the metadata for individual datasets (but not the dataset itself).
- *A reference model for the form of DLT that is to be* used – that is optional as other designs may be used.

Second, support for a range of funded programmes, or simple policy support for industry funding in the following areas:

- *Shared vertical sector business platforms* – to encourage sector and cross-sector use of open data for shared business activities such as private design and innovation, and open collaborative innovation (i.e. pre-commercialisation). The policy should ensure that the commons model really does deliver what the commercial private/secret data model can never provide – additional wealth and employment across many sectors from a common shared open database platform without commercial restrictions.

- *Industrial research* – Create an industrial research programme year for ad hoc development communities, funded by the European Commission, for instance in:
  - Engineering and manufacturing
  - eHealth
  - Process industries (chemicals, food, biosciences, pharmaceuticals, etc)
  - Education and training - encourage education and vocational training in using open data, at all levels, to form a new generation of students well versed in its use and to harness their creative force and ideas for the EU community.

The open dataset licence should cover most aspects of the commercialisation of data, defining what data can be used freely and with what protections of its digital rights to ensure that non-eligible data (e.g. personal private data) cannot be used.

5.3. Other Safeguards Requiring Regulatory Controls May Be Desirable

These include:

- Creation of checks for the presence of personal data and proof of its eradication
- Regulation on compliance of datasets, primarily on observing the terms of the open data licence, including updates to the DLT records, if used, that act as the reference log for the history of use and the permissions allowed in the licence.

Table 4 ranks the recommendations proposed above according to their likely impact.

Table 4: Recommendations Ranked According to Their Likely Impact

Recommendation	Ranking
Promote use of open data licences to build trust	1
Support testing for contamination of open data by personal data	2
Promote sharing of private data within vertical sectors and across sectors to increase the volume of open data	3
Ensure abuse of SMP does not compromise the open data initiative	4

REFERENCES

- Bauer, Matthias, Ferracane, Martina F., Lee-Makiyama, Hosuk, van der Marel, and Erik (2016), *Unleashing Internal Data Flows in the EU: An Economic Assessment of Data Localisation Measures in the EU Member States*, ECIPE Policy Brief, No. 03/2016, http://ecipe.org/app/uploads/2016/12/Unleashing-Internal-Data-Flows-in-the-EU.pdf.
- Crawford, Alejandro, and Chau, Lisa (2013), "Why Google's business model works", *US News and World Report*, https://www.usnews.com/opinion/blogs/economic-intelligence/2013/06/25/why-googles-business-model-works.
- Drexl, Josef (2016), *Designing Competitive Markets for Industrial Data: Between Propertisation and Access*, Max Planck Institute for Innovation & Competition, Research Paper, No. 16-13, http://www.ip.mpg.de/en/publications/details/designing-competitive-markets-for-industrial-data-between-propertisation-and-access.html.
- Duch-Brown, Nestor, Martens, Bertin, and Mueller-Langer, Frank (2017), *The Economics of Ownership, Access and Trade in Digital Data*, Digital Economy Working Paper 2017-01, JRC Technical Reports, https://ec.europa.eu/jrc/sites/jrcsh/files/jrc104756.pdf.
- EuDeCo (2015), Initial Heuristic Model of EuDeCo of the European data Economy, http://data-reuse.eu/wp-content/uploads/2016/09/D2.1_InitialHeuristicModel-v1_2015-11-06.pdf.
- European Commission (2014), "Towards a thriving data-driven economy", Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and The Committee of the Regions, Brussels, COM(2014) 442 final, 7 July, http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52014DC0442&from=EN.
- European Commission (2017a), "Framework for the free flow of non-personal data in the European Union", Proposal for a Regulation from the European Parliament and of the Council, Brussels, COM(2017) 495 final, 13 September, http://ec.europa.eu/transparency/regdoc/rep/1/2017/EN/COM-2017-495-F1-EN-MAIN-PART-1.PDF.
- European Commission (2017b), "Building A European Data Economy", Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and The Committee of the Regions, Brussels, COM(2017) 9 final, 10 January, http://ec.europa.eu/newsroom/dae/document.cfm?doc_id=41205.
- European Commission (2017c), "Commission Staff Working Document on the free flow of data and emerging issues of the European data economy", SWD(2017) 2 final, 10 January, http://ec.europa.eu/newsroom/dae/document.cfm?doc_id=41247.
- European Data Portal (2017), *Open Data Maturity 2016: Insights into the European State of Play*, Report for the European Commission, https://www.europeandataportal.eu/sites/default/files/edp_landscaping_insight_report_n2_2016.pdf.
- Forge, S. (2004a), "So where is the value? The role of intellectual capital in business", *Information Economics Journal*, March.
- Forge, S. (2004b), *Open Source Software: Importance for Europe*, Institute for Prospective Technological Studies, JRC, European Commission, Seville, March, http://cordis.europa.eu/pub/ist/docs/opensourcesoftware-report.pdf.

- Harding, Robin (2017), “Japan eyes tough anti-monopoly rules on data”, *Financial Times*, 17 July.
- Hern, Alex (2017), “'Anonymous' browsing data can be easily exposed, researchers reveal”, *The Guardian*, 1 August, https://www.theguardian.com/technology/2017/aug/01/data-browsing-habits-brokers.
- Hugenholtz, P.B. (2017), “Data property in the system of intellectual property law: welcome guest, or misfit, paper presented at conference”, paper presented at Trading Data in the Digital Economy: Legal Concepts and Tools, Muenster Colloquium on EU Law and the Digital Economy, University of Muenster,4 May.
- IDC and Open Evidence (2017), *European Data Market*, SMART 2013/0063, http://www.datalandscape.eu/study-reports.
- Lambrecht, Anja and Tucker, Catherine E. (2015), “Can Big Data Protect a Firm from Competition?”, https://ssrn.com/abstract=2705530.
- Osborn Clarke (2016), *Legal study on Ownership and Access to Data*, A study prepared for the European Commission, https://publications.europa.eu/en/publication-detail/-/publication/d0bec895-b603-11e6-9e3c-01aa75ed71a1.
- Prϋfer, J., and Schottmuller, C. (2017), Competing with Big Data, CentER Discussion Paper, Vol 2017-007, Center for Economic Research, Tilburg, https://pure.uvt.nl/ws/files/15514029/2017_007.pdf
- Raymond, Eric S. (1997), *The Cathedral and the Bazaar*, Linux Kongress, Wurzburg.
- Robinson, Stephen Cory (2017), “What’s your anonymity worth? Establishing a marketplace for the valuation and control of individuals’ anonymity and personal data”, *Digital Policy, Regulation and Governance*, Vol 19, No 5.
- Schneier, Bruce (2015), *Data and Goliath: The Hidden Battles to Collect Your Data and Control Your World*, WW Norton, New York.
- Van Asbroeck, Benoit, Debussche, Julien, and César, Jasmien (2017a), *Building the European Data Economy: Data Ownership,* White Paper, Bird & Bird, 1 January, https://www.twobirds.com/en/news/articles/2017/global/data-ownership-in-the-context-of-the-european-data-economy.
- Van Asbroeck, Benoit, Debussche, Julien, and César, Jasmien (2017b), *Data Ownership: A new EU right in data*, Supplementary Paper, Bird & Bird, 31 March, https://sites-twobirds.vuture.net/1/773/uploads/white-paper---data-ownership---a-new-eu-right-in-data.pdf.

CAT: QA-06-17-181-EN-C (paper)
CAT: QA-06-17-181-EN-N (pdf)

DIRECTORATE-GENERAL FOR INTERNAL POLICIES

POLICY DEPARTMENT ECONOMIC AND SCIENTIFIC POLICY A

Role

Policy departments are research units that provide specialised advice to committees, inter-parliamentary delegations and other parliamentary bodies.

Policy Areas

- Economic and Monetary Affairs
- Employment and Social Affairs
- Environment, Public Health and Food Safety
- Industry, Research and Energy
- Internal Market and Consumer Protection

Documents

Visit the European Parliament website:
http://www.europarl.europa.eu/supporting-analyses

PHOTO CREDIT:
iStockphoto.com; Shutterstock/beboy

ISBN 978-92-846-2170-5 (paper)
ISBN 978-92-846-2169-9 (pdf)

doi: 10.2861/014109 (paper)
doi: 10.2861/91583 (pdf)

Publications Office

Data Flows – Future Scenarios