Chapter 4

Drawing value from data as an infrastructure

This chapter introduces the theoretical foundation for the economic potential of data and discusses key data governance issues that need to be addressed in order to maximise data's potential and reuse across society. It begins by presenting data as an infrastructural resource and a non-rivalrous capital good. It goes on to discuss how data's value depends entirely upon context, with reuse enabling multi-sided markets in which huge returns to scale and scope can lead to positive feedback loops. The often misunderstood notion of "ownership" is discussed, and data quality is seen as multi-faceted and involving seven dimensions. The key aspects of data access, sharing, portability and interoperability are examined and presented as elements of a data governance framework that can help overcome barriers to the reuse of data.

> *I recognised that information was, in many respects, like a public good, and it was this insight that made it clear to me that it was unlikely that the private market would provide efficient resource allocations whenever information was endogenous.* (Stiglitz, 2001)

> *Through ever-expanding commerce, the nation becomes ever-wealthier, and hence trade and commerce routes must be held open to the public, even if contrary to private interest. Instead of worrying that too many people will engage in commerce, we worry that too few will undertake the effort.* (Rose, 1986)

Data have increasingly become an important source of value creation and (data-driven) innovation (DDI). More and more organisations collect, store, and process data today to expand their future production capacities (see Chapter 1 and 2 of this volume), and the productivity improvements are truly dramatic. TomTom, a leading provider of navigation hardware and software, now has more than nine trillion data points collected from its navigation devices and other sources, describing time, location, direction and speed of travel of individual anonymised users, and it now adds six billion measurement points every day.[1] The results of the data analysis are fed back to its navigation devices to inform drivers about current and predicted traffic. This can lead to significant time savings and reduce congestion. Overall, estimates suggest that the global pool of personal geo-locational data has been growing by 20% a year since 2009. By 2020, this data pool could provide USD 500 billion in value worldwide in the form of time and fuel savings, or 380 million tonnes of CO2 emissions saved (MGI, 2011).

As the use of data becomes an increasingly important economic and social phenomenon, economists and policy analysts are trying to capture the phenomenon through existing concepts and theories. Metaphors such as "data is the new currency" (Schwartz, 2000 cited in IPC, 2000; Zax, 2011; Dumbill, 2011; Deloitte, 2013) or, more recently, "data is the new oil" (Kroes, 2012; Rotella, 2012; Arthur, 2013) are often used as rhetorical means to make this emerging phenomenon better understandable to policy and decision makers. Although at first helpful to highlight the (new) economic value of data, these metaphors often fall short and are sometimes even misleading, and therefore should be used with caution (see for example Thorp, 2012; Bracy, 2013; and Glanz, 2013). For example, data are not a rivalrous good, nor are they a primary resource – such as oil, which is depleted once extracted, transformed and burned during production processes. In contrast to oil, the use of data does not exhaust the supply of data and (therefore) *in principle* its potential to meet the demands of others. All these metaphors however reflect an urgent need for a concept through which to better understand and analyse the economics of data, ideally building on familiar concepts, so as to develop better policies and strategies for data's governance.

This chapter responds to that need from a public policy perspective. It provides a framework that can guide policy makers in identifying when data warrant their attention. Not all data are of great value-added from a public policy perspective, at least at first sight: an example here would be data generated when posting on social networks such as Facebook. There are moreover controversies about the use of (e.g.) personal data. However, if the agglomeration and sharing of any data across society can respond to specific societal needs, then that data may merit policy makers' attention. The chapter begins with an analysis of the fundamental economic properties that account for data's potential as a driver of value creation and economic growth and development. These properties include: i) the (non)rivalrous nature of their consumption, ii) their (non)excludability, and iii) the economics of scale and scope in the creation and use of data. These properties lead to the conclusion that data are an infrastructural resource. Building on a rich literature base dealing with the economics of infrastructures, especially the work of Frischmann (2012), the chapter then analyses major supply- and demand-side issues that emerge from data *as* an infrastructure. Special attention is given to potential spillovers (positive externalities) that provide the major theoretical link to total factor productivity growth as highlighted by a number of scholars[2] (among them Corrado et al. [2012]) and the implications in managing data as (knowledge) commons.

4.1. Data as infrastructural resource

The economic properties of data suggest that data may be considered as an infrastructure or infrastructural resource. This may sound counterintuitive, since traditionally infrastructures typically refer to large-scale physical facilities provided for public consumption; the classic examples are transportation systems, including highway and railway systems; communication systems, including telephone and broadband networks; and basic services and facilities such as buildings and sewage and water systems (Frischmann, 2012). However, as for example recognised by the US National Research Council (NRC, 1987), the notion of infrastructure also refers to non-physical facilities, such as education systems and governance systems (including for example the court system). Frischmann (2012) highlights that "the NRC recognised three conceptual needs … first, the need to look beyond physical facilities; second, the need to evaluate infrastructure from a systems perspective; and third, the need to acknowledge and more fully consider the complex dynamics of societal demand". According to Frischmann, the broader concept of infrastructures strongly suggests that they be regarded from a functional perspective rather than from a purely physical or organisational perspective.

As defined by Merriam-Webster, infrastructures are "the basic equipment and structures … that are needed for a country, region, or organisation to function properly". According to Frischmann (2012), they provide the "underlying foundation or basic framework (as of a system or organisation)". That author goes on to state (2012) that infrastructure resources are "shared means to many ends", which satisfy the following three criteria:

1. the resource may be consumed in a non-rivalrous fashion for some appreciable range of demand (i.e. the non-rivalrous criterion)
2. social demand for the resource is driven primarily by downstream productive activities that require the resource as an input (i.e. the capital good criterion)
3. the resource may be used as an input into a wide range of goods and services, which may include private goods, public goods, and social goods (i.e. the general-purpose criterion).

As discussed in the following sections, most (though not all) data are indeed "shared means to many ends" and satisfy Frischmann's three criteria. Therefore, data can in principle be considered an infrastructural resource.

Data as a non-rivalrous good

(Non)rivalry of consumption describes the degree to which the consumption of a resource affects (or does not affect) the potential of the resource to meet the demands of others. It thus reflects the marginal cost of allowing an additional consumer of the good. A purely rivalrous good such as oil can only be consumed once. A non-rivalrous good such as data, in contrast, can be consumed in principal an unlimited number of times. But if this property is, as noted above, the source of significant spillovers that provide the major theoretical link to total factor productivity growth, it also raises questions about how best to allocate data as a resource.

While it is widely accepted that social welfare is maximised when a rivalrous good is consumed by the person who values it the most, and that the market mechanism is generally the most efficient means for rationing such goods and for allocating resources needed to produce such goods, this is not always true for non-rivalrous goods

(Frischmann, 2012). The situation is more complex, since non-rivalrous goods come with an additional degree of freedom with respect to resource management. As Frischmann (2012) highlights, social welfare is maximised not when the good is consumed solely by the person who values it the most, but when everyone who values it consumes it. Maximising access to the non-rivalrous good will in theory maximise social welfare, as every additional private benefit comes at no additional cost.

Data as a capital good

Data are often described as "the new oil". However, besides the non-rivalrous nature of data, there is another drawback with such an analogy: data are neither a consumption good such as an apple, nor an intermediate good such as oil. In most cases, data can be classified as a capital good.

Consumption goods are consumed to generate direct benefits to the consumer or firm. The United Nations System of National Accounts (SNA) defines a consumption good or service as "one that is used (without further transformation in production) by households, NPISHs [non-profit institutions serving households] or government units for the direct satisfaction of individual needs or wants or the collective needs of members of the community" (UN, 2008). In contrast, intermediate goods and capital goods are used as inputs to produce other goods. They are means rather than ends, and their demand is driven by the demand for the derived outputs. They are thus factors of production (see Saviz, 2011; Jones, 2012).

Intermediate consumption is defined by the SNA (UN, 2008) as "consist[ing] of the value of the goods and services consumed as inputs by a process of production, excluding fixed assets whose consumption is recorded as consumption of fixed capital". Capital goods, according to the OECD, are "goods, other than material inputs and fuel, used for the production of other goods and/or services".[3] Intermediate goods such as raw materials (e.g. oil) are used up, exhausted, or otherwise transformed when used as input to produce other goods; capital goods are not. Furthermore, capital goods "must have been produced as outputs from processes of production", which explains why "natural assets such as land, mineral or other deposits, coal, oil, or natural gas, or contracts, leases and licences" are not considered capital goods (UN, 2008).[4]

Data can sometimes be consumed to directly satisfy consumer demand. This is the case for example with an OECD statistic, which will inform the reader about a socio-economic fact. However, in most cases data are not a consumption good but instead are used as an input for goods or services; this is especially true of large volumes of data (i.e. "big data"), which are means rather than ends in themselves. In other words, demand for big data is not for the data itself, but for the benefits that their use promises to bring. In that sense, even pure data products such as infographics (i.e. graphic visual representations of data, information, or knowledge) are the outputs of algorithms applied to data – in the case of infographics, visualisation algorithms.

Data are also not an intermediate good, as they are not exhausted when used given their non-rivalrous nature. This does not mean that data cannot be discarded after they have been used. In many cases, they are used just once. However, while the cost of storing data in the past discouraged keeping data that were no longer, or unlikely to be, needed, storage costs today have decreased to the point where data can generally be kept for long periods of time, if not indefinitely. This has increased data's capacity to be used as a capital good and production factor.

Furthermore, being a capital good does not mean that data do not depreciate like most capital goods, whose value declines "as a result of physical deterioration, normal obsolescence or normal accidental damage" (UN, 2008). In the case of data, depreciation is more complex because it is context dependent, as further described below. Data have no intrinsic value as the value depends on the context of its use. A number of factors presented in more detail in the following sections can affect that value, in particular i) the *accuracy* and ii) the *timeliness* of data. The more relevant and accurate data are for the particular context in which they are used, the more useful and thus valuable data will be (see Oppenheim, Stenson and Wilson, 2004, cited in Engelsman, 2009). This implies, however, that the value of data can perish over time depending on how they are used (see Moody and Walsh, 1999, cited in Engelsman, 2009). Data can especially depreciate in value when they begin to lose their relevance for a particular intended use. There is thus a temporal premium that is motivated by the "real-time" supply of data, for example in the financial sector.

The capital good nature of data has major implications for economic growth. As data are a non-rival capital, they can in theory be used (simultaneously) by multiple users for multiple purposes as an input to produce an unlimited number of goods and services. In practical terms this link to total factor productivity growth finds its application in data-enabled multi-sided markets, i.e. economic platforms in which distinct user groups generate benefits (externalities or spillovers) to other groups.

Data as general-purpose input

As Frischmann explains, "infrastructure resources enable many systems (markets and nonmarkets) to function and satisfy demand derived from many different types of users". They are not inputs that have been optimised for a special limited purpose, but "they provide basic, multipurpose functionality" (Frischmann, 2012). In particular, infrastructures make possible a wide range of private, public and social goods, which users are free to produce according to their capabilities.

How data are used will typically depend on the initial purpose for which they have been collected. For example, at the outset agricultural data will primarily be used for agricultural goods and services. However, in theory there are no limits with regard to the purposes for which data can be used, and many of the benefits stemming from their reuse are based on the fact that data created in one domain can provide further insights when applied in another domain. A clear illustration is provided by open public sector data, where data sets used originally for administrative purposes are reused by entrepreneurs to create services unforeseen when the data were originally created. Likewise, researchers in the areas of health care and Alzheimer's disease are considering reusing retail and social network data to study the impact of behavioural and nutritional patterns on the evolution of the disease.

The general-purpose nature of infrastructure comes with a key policy implication. The production of (ex-ante unforeseeable) public and social goods via the infrastructure could lead to the market failure of insufficient provision of the infrastructure, which would call for government intervention in some cases. As Frischmann explains, "[U]sers' willingness to pay [for the infrastructure] reflects private demand – the value that they expect to realise – and does not take into account value that others might realise as a result of their use" (social value). That "social value may be substantial but extremely difficult to measure", thus leading to a "demand-manifestation problem" which in turn may lead to an undersupply of the infrastructure and a "prioritisation of access and use of

the infrastructure for a narrower range of uses than would be socially optimal" (Frischmann, 2012). As a consequence, there can be significant (social) opportunity costs in limiting access to infrastructures. In other words: open (closed) access enables (restricts) user opportunities and degrees of freedom in the downstream production of private, public and social goods, many of which by their nature have significant spillover effects. In particular, in environments characterised by high uncertainty, complexity and dynamic changes, open access can be an optimal (private and social) strategy for maximising the benefits of an infrastructure.

This means that data markets may not be able to fully serve social demand for data where such a demand manifestation problem would occur. Although no literature is known to have discussed the data demand manifestation problem, there are plausible reasons to believe that such a problem may occur in the data ecosystem, for instance, when data is used to increase transparency in government (see Chapter 10 of this volume). In addition, the context dependency of data and information presented below and the highly uncertain, complex, and dynamic environment in which some data are used (e.g. research) make it almost impossible to fully evaluate ex ante the potential of data, and would exacerbate a demand manifestation problem.

The latter point calls for managing data based on non-discriminatory access regimes, for instance as commons or through open access regimes. Frischmann (2012) points to the following reasons; the first two are in fact closely associated with the concept of *open innovation* (see Box 4.1), as discussed in *The OECD Innovation Strategy* (2010) and the OECD (2013a) project on "Knowledge Networks and Markets":

- *Facilitating joint production or co-operation with suppliers, customers or even competitors* is not a new phenomenon. Joint research ventures or patent pools are well known examples, where firms share resources under non-discriminatory access regimes. This is "because independent research efforts are inhibited by complexity, expense, strategic concerns, transaction costs, or other impediments" (Frischmann, 2012). Sharing agreements are very often an important part of these collaboration efforts. In the particular case of data, access does not need to be open to the public, but it may be limited to the partners who share their data as commons to "overcome collective action problems, sometimes mere co-ordination problems and sometimes more difficult prisoner's dilemma problems"[5] (Frischmann, 2012).
- *Supporting and encouraging value-creating activities by users (user-/consumer-driven innovation)* can be enabled thanks to open access. Open access is an optimal strategy for organisations "when they recognise that users may be best positioned to create value" (Frischmann, 2012). In its weakest form, where users are granted access only to their own personal data, consumers are given "better visibility into their own consumption, often revealing information that can lead to changes in behavior" (MGI, 2013). In its most extreme form, where access is granted to the public, users (including consumers and citizens) are empowered to "provide input to improve the quality of goods and services" (MGI, 2013). This includes improving public services as well as the quality of data.[6]
- *Maximising the option value[7] of the organisation's infrastructural resource when there is high uncertainty regarding sources of future market value.* In contrast with the case described above, where organisations know that users are best placed to create future value, here organisations "are uncertain about the future sources of the value … what unforeseen uses may emerge, what people will want,

how much people will be willing to pay, what complementary goods and services may arise in the future, and so on" (Frischmann, 2012). They adopt open access strategies, taking "advantage of the increased value of experimentation by users, the increased range of potential value-creating services, market selection of the best services that eventually emerge, and learning over time about user preferences and possible paths for continued development". The advantage for the organisation is that it "maintains flexibility and avoids premature optimisation or lock-in to a particular development path or narrow range of paths" (Frischmann, 2012).

- *(Cross-)subsidising the production of public and social goods* requires picking winners (users or applications) by assessing (social) demand for such goods based on the (social) value they create (Frischmann, 2012). Governments can support the production of public goods i) by directly producing these goods, or ii) by supporting private firms' production of public and social goods through (e.g.) research grants, procurement programmes, contracted research and tax incentives. All these strategies raise a number of issues, including difficulties in picking winners and losers, and the fact that resources are limited. Open access regimes can be a more efficient and politically attractive "indirect intervention" to support the production of public and social goods. As Frischmann (2012) highlights, "commons management is not a direct subsidy to … users who produce public or social goods, but it effectively creates cross-subsidies and eliminates the need to rely on either the market or the government to 'pick winners' – that is, to prioritise or rank … users worthy of access and support".

Box 4.1. **Illustrations of "openness"**

Open innovation – This term refers to the "use of purposive inflows and outflows of knowledge to accelerate internal innovation, and expand the markets for external use of innovation". That includes proprietary-based business models that make active use of licensing, collaborations, joint ventures, etc. Here, "open" is understood to denote the arm's-length flow of innovation knowledge across the boundaries of individual organisations.

Open source – This term is now applied to designate innovations, often jointly developed by different contributors, available royalty-free to anyone and without significant restrictions on how they are to be used. A possible restriction is that derivative work also has to be provided on a same basis.

Open science – This term is often used to describe a movement that promotes greater transparency in the scientific methodology used and data collected; advocates the public availability and reusability of data, tools and materials; and argues for broadly communicating research (particularly when publicly funded) and its results.

Open access – This term refers to the possibility of accessing scientific literature and data "digital, online, free of charge, and free of most copyright and licensing restrictions". This term is also increasingly applied to data provided by profit-driven operators, who develop business models that enable them to obtain a source of revenue bundled alongside information provided on a free and open basis.

Open knowledge – This term coined by the Open Knowledge Foundation refers to any content, information or data that people are free to use, reuse and redistribute, without any legal, technological or social restriction.

Source: OECD (2013a).

4.2. The economics of data

Data increasing returns to scale and scope and network effects

Returns to scale are concerned with changes in the level of output as a result of changes in the amount of factor inputs used. Increasing returns to scale are realised when for example the doubling of the amount of all factors of production results in more than double the output. Returns to scope are conceptually similar to returns to scale, except that it is not the size or the scale of the factor inputs that leads to over-proportionate outputs, but the diversity of the input. In contrast, economies of scale are the cost advantages that organisations obtain thanks to the size of their outputs or the scale of their operation. As the size and scale increases, the cost per unit of output (average cost) decreases. Economies of scope are conceptually similar to economies of scale, except that – once again – it is not the size or the scale of the outputs that leads to over-proportionate reduction in the average cost (cost per unit), but the diversity of the product.

Networks effects, which often referred to as demand-side economies of scale, refer to the fact that the utility of a good to a user (on the demand side) depends on the use of that good by other users. An example often given is the fax machine. While a single fax machine has no utility to a single user, a fax machine starts generating benefits as more users decide to purchase a fax machine, as the technology provides a growing opportunity to communicate with an existing network of users. Many data-driven services and platforms, such as social networking sites, are characterised by large network effects where the utility of the services increases over-proportionately with the number of users. This reinforces the increasing returns to scale and scope on the supply side.

The use of data can generate large returns to scale and scope, as data are non-rivalrous capital that can be reused with positive feedback loops that reinforce each effect at the supply and demand sides. At the same time, the accumulation of data also comes with certain costs (e.g. storage) and risks (e.g. privacy violation and digital security risks). Nevertheless, the advantages are clear:

1. *Increasing returns to scale* – The accumulation of data can lead to significant improvements of data-driven services that in turn can attract more users, leading to even more data that can be collected. This "positive feedback makes the strong get stronger and the weak get weaker, leading to extreme outcomes" (Shapiro and Varian, 1999). For example, the more people use services such as Google Search, or recommendation engines such as that provided by Amazon, or navigation systems such as that provided by TomTom, the better the services, as they become more accurate in delivering requested sites and products and providing traffic information, and the more users they will attract.
2. *Increasing returns to scope* – Diversification of services leads to even better insights if data linkage is possible. This is because data linkage enables "super-additive" insights, leading to increasing returns to scope. Linking data is a means to contextualise data and is thus a source for insights and value that are greater than the sum of isolated parts (data silos). As Newman (2013) highlights in the case of Google: "It's not just that Google collects data from everyone using its search engine. It also collects data on what they're interested in writing in their Gmail accounts, what they watch on YouTube, where they are located using data from Google Maps, a whole array of other data from use of Google's Android phones, and user information supplied from Google's whole web of online services."[8] This diverse data sets enable better profiling hardly possible otherwise.

These effects are not mutually exclusive and may interact, leading to a multiplication. For instance, consumers that appreciate customised search results and ads by Google's search and webmail platform will spend more time on the platform, which allows Google to gather even more valuable data about consumer behaviour and to further improve services, for (new) consumers as well as advertisers (thus on both sides of the market). These self-reinforcing effects may increase with the number of applications provided on a platform, e.g. bundling email, messaging, video, music and telephony – as increasing returns to scope kick in and even more information becomes available thanks to data linkage. As a result, a company such as Google ends up (together with Facebook) with an almost 60% of market share in the US mobile ad market.

Data as non-rivalrous capital enabling multi-sided markets

The effects presented above need to be considered in the context of multi-sided markets that data enable. Two- or multi-sided markets are "roughly defined as markets in which one or several platforms enable interactions between end users and try to get the two or multiple sides 'on board' by appropriately charging each side" (Rochet and Tirole, 2006). These platforms enable multiple distinct groups of customers not only to interact, but also exchange possible externalities among themselves. In other words, the decisions of each group affect the outcome for the other groups. As a consequence, the prices charged to the members of each group will often reflect the effects of these externalities. If the activities of one side create a positive externality for another side (for example more clicks by users on links sponsored by advertisers), then the prices to that other side can be increased (OECD, 2014).

The reuse of data enables multi-sided markets in which huge returns to scale and scope can lead to positive feedback loops in favour of the business on one side of the market, which in turn reinforces success in the other side(s) of the multi-sided market. Established and emerging service platforms such as Google, Facebook, TomTom and John Deere have developed data- and analytics-enabled multi-sided markets, i.e. economic platforms in which distinct user groups generate benefits (externalities or spillovers) for the other side(s). In this they differ from multi-sided markets such as eBay, Amazon, Microsoft's Xbox platform, and Apple's iTunes store. eBay and Amazon, for example, provide online marketplaces for sellers and buyers, and are multi-sided by virtue of their business model (online market). This is also true of Microsoft's Xbox platform, which is positioned in between consumers and game developers, and Apple's iTunes store, which provides a platform that links consumers to application developers and musicians.

In contrast, TomTom's navigation services are provided to consumers as well as to traffic management providers. The service provided to the traffic management providers builds on the analysis of consumer data. The same applies to Google and Facebook, which provide online services to consumers while (re-)using consumer data to provide marketing services to third parties, and to John Deere, which collects agricultural data from farmers and provides them as a service to large seed companies. Data are at the core of these companies' multi-sidedness as non-rivalrous capital collected and used on one side of the market, e.g. to personalise the service, and reused on the other side(s) as input for a theoretically unlimited number of additional goods and services, such as marketing.

Context dependencies

As OECD (2012) highlighted, assessing the value of data ex ante (before use) is almost impossible, because the information derived is context dependent: data that are of good quality for certain applications can thus be of poor quality for other applications. It therefore comes as no surprise that the OECD (2011) Quality Framework and Guidelines for OECD Statistical Activities defines "data quality" as "fitness for use" in terms of user needs, underlining this context dependency (see section below on data quality and curation).

Furthermore, the information that can be extracted from data is not only a function of the data, but also a function of the (analytic) capacity to link data and to extract insights. This capacity is determined by available (meta-)data, analytic techniques and technologies; however, it is a function of pre-existing knowledge and skills. This means that there are factors beyond the data themselves that determine value:

- *Data linkage* – Information depends on how the underlying data are organised and structured. In other words, the same data sets can lead to different information depending on their structure, including their linkages with other (meta-)data.
- *Data analytic capacities* – The value of data depends on the meaning as extracted or interpreted by the receiver. The same data sets can thus lead to different information depending on the analytic capacities of the "receiver", including their skills and (prior) knowledge, available techniques, and technologies for data analysis.

4.3. Towards a data governance framework for better data access, sharing and interoperability

Given their role as the underlying framework of society, infrastructures have always been the object of public policy debates, and governments have played and continue to play a significant and widely accepted role in ensuring the provision of many infrastructures (Frischmann, 2012). The main rationale for the role of governments is justified by the significant spillovers (positive externalities) that infrastructures generate and which result in large social gains, many of which are incompletely appropriated by the suppliers of the infrastructure (Steinmueller, 1996). Spillovers of this nature provide a major theoretical link to total factor productivity growth, but they also present challenges in measuring the contribution of infrastructures or attributing economic growth to that contribution, as the OECD (2012**)** work on measuring the economic impact of the Internet has demonstrated. As Frischmann (2012) explains: "The externalities are sufficiently difficult to observe or measure quantitatively, much less capture in economic transactions, and the benefits may be diffuse and sufficiently small in magnitude to escape the attention of individual beneficiaries."

The positive externalities are also the reason why "infrastructures generally are managed in an openly accessible manner whereby all members of a community who wish to use the resources may do so on equal and non-discriminatory terms" (Frischmann, 2012). The community may, but does not necessarily include the public at large. Furthermore, this does not mean that access is free, nor that access is unregulated. The important point here is that, as Rose highlights (1986, cited in Frischmann, 2012), the positive externalities in combination with open access can lead to a "comedy of the commons", where greater social value is created with greater use of the infrastructure.

Taking commerce as an example, Rose (1986) explains that open access to roads have enabled commerce to generate not only private value that is easily observed and captured by participants in economic transactions, but also social value that is not easily observed and captured by participants (e.g. value associated with socialisation and cultural exchange). In this case, commerce is a productive downstream use of the road infrastructure that generates private as well as social surplus.

In contrast to Hardin's (1968) "tragedy of the commons", where free riding on common (natural) resources leads to the degradation and the depletion of the resources, the "comedy of the commons" is possible in the case of non-rivalrous resources such as data. It is also the strongest rationale for policy makers to promote access to data, either through "open data" in the public sector, "data commons" such as in science, or through the more restrictive concept of "data portability" to empower consumers.

The following section discusses the key challenges related to data governance. These are common challenges that individuals, businesses and policy makers face in every domain in which data are used, irrespective of the type of the data used.

Open data and data commons

A precondition for creating any economic or social value of data is access. Data are a non-rivalrous good and, as mentioned above, their use does not affect in principle their potential to meet the demands of others. As a result, data have unlimited potential to create value. On the other hand, barriers to data access can inhibit data sharing and hinder collaboration, (open) innovation, and the downstream production of data-based goods and services, many of which have significant spillover effects. As a consequence, there can be significant (social) opportunity costs due to barriers to access.

The term "open data" is increasingly used in many different contexts as a solution to promote better access to data. It may actually refer to different concepts, which share a number of commonalities. Open data for governments, for example, often refers to initiatives such as data.gov (United States), data.gov.uk (United Kingdom), or data.gov.fr (France); these enhance access to public sector information (PSI), including public sector data, as encouraged by the OECD (2008) *Council Recommendation on Enhanced Access and More Effective Use of Public Sector Information* (see Chapter 10 of this volume).

The term "open data" in the scientific community refers to open access to scientific data, as promoted for example by the OECD (2004) *Declaration on Access to Research Data from Public Funding* and the OECD (2006b) *Council Recommendation concerning Access to Research Data from Public Funding*.[9] All these OECD instruments highlight openness as the first key principle (see Chapter 7 and 10). Last but not least, "open data" is also often associated with movements such as the open source movement, which became particularly popular in the context of open source software (OSS) such as Linux. According to Wikipedia, "Open source as a development model, promotes a) universal access via free license to a product's design or blueprint, and b) universal redistribution of that design or blueprint, including subsequent improvements to it by anyone."

It is important to note that the concept of open data is not limited to the public sector. UN Global Pulse (2012), for example, introduced the concept of "data philanthropy", whereby the private sector shares data to support more timely and targeted policy action, and to highlight the public interest in shared data. In this context two ideas are debated: i) the "data commons", where some data are shared publicly after adequate anonymisation and aggregation; and ii) the "digital smoke signals", where sensitive data

are analysed by companies but the results are shared with governments. The Open Data Institute (ODI), a not-for-profit organisation based in the United Kingdom, is also promoting the release of open data in the private sectors, including but not limited to finance and health care.

Most definitions for open data point to a number of criteria or "principles". According to the OECD (2005) *Recommendation on Principles and Guidelines for Access to Research Data from Public Funding*, for example, openness means i) access that should be granted on equal or non-discriminatory terms, and ii) access costs that should not exceed the marginal cost of dissemination. As another example, at a meeting of open data advocates in 2007,[10] participants agreed on "8 Principles of Open Government":

- *Complete* – All public data are made available. Public data are data that are not subject to valid privacy, security or privilege limitations.
- *Primary* – Data are as collected at the source, with the highest possible level of granularity, not in aggregate or modified forms.
- *Timely* – Data are made available as quickly as necessary to preserve their value.
- *Accessible* – Data are available to the widest range of users for the widest range of purposes.
- *Machine processable* – Data are reasonably structured to allow automated processing.
- *Non-discriminatory* – Data are available to anyone, with no registration requirement.
- *Non-proprietary* – Data are available in a format over which no entity has exclusive control.
- *Licence-free* – Data are not subject to any copyright, patent, trademark or trade secret regulation. Reasonable privacy, security and privilege restrictions may be allowed.

Other definitions that followed focused on a smaller set of criteria. The Open Data White Paper of the United Kingdom Cabinet Office (2012), for example, highlights three of the principles listed above as criteria for open data: i) "accessible (ideally via the Internet) at no more than the cost of reproduction, without limitations based on user identity or intent", ii) "in a digital, machine readable formation for interoperation with other data", and iii) "free of restriction on use or redistribution in its licensing". A recent report by MGI (2013), which defines open data as "the release of information by government and private institutions and the sharing of private data to enable insights across industries", also based its definition on these three criteria, highlighting however access costs as a fourth criterion. A comprehensive discussion of the principles governing open data can be found in Ubaldi (2013).

Among the criteria listed in the above definitions, non-discriminatory access (or "access on equal terms", as stated in the OECD [2005] Recommendation on Principles and Guidelines for Access to Research Data from Public Funding) is central to open data. Non-discriminatory access is about "terms that do not depend on the users' identity or intended use" (Frischmann, 2012; see also United Kingdom Cabinet Office, 2012). As highlighted above, access independent of identity and intent can be crucial for

maximising the value of data across society, as it keeps the range of opportunities as wide as possible.

All other criteria listed above are factors affecting the level of non-discriminatory access, and thus the degree of openness. Three criteria deserve to be highlighted, as they significantly affect the degree of openness (ordered by their increasing magnitude of influence):

- *Technological design* is a broad concept that includes all technical aspects affecting the (re-)use and distribution of data. These factors were presented in Berners-Lee's (2006b) proposed "5 Star Deployment Scheme for Open Data": 1) "make your stuff available on the Web (whatever format) [under an open licence]"; 2) make it available as structured data (e.g. Excel instead of an image scan of a table); 3) "use non-proprietary formats (e.g., CSV [comma-separated values] instead of Excel)"; 4) "use URIs [uniform resource identifiers] to identify things, so that people can point at your stuff"; 5) "link your data to other data to provide context". In essence, the scheme points to the following key technological factors affecting the degree of data openness: i) data availability (ideally online), ii) machine readability (of structured data), and iii) data linkability. It should be noted that factor (i) is required for factor (ii), which in turn is a requirement for factor (iii).
- *Intellectual property rights (IPRs)* – Data can be subject to legal regimes, copyright as well as other IPRs applicable to databases (Box 4.2.) and trade secrets, which need to be respected as highlighted in the OECD (2008) *Council Recommendation on Enhanced Access and More Effective Use of Public Sector Information*. These rights can in some cases limit or prevent the (re-)use and distribution of open data. Some open data initiatives therefore explicitly state that open data should be free of any IPRs (see the 8 Principles of Open Government above). In other cases, innovative IP regimes are used and even promoted through open data regimes, as long as they do not restrict the rights of users to reuse and sometimes redistribute the data. In 2010, for example, the United Kingdom created the *Open Government Licence*[11] to release public sector information (including data) for free without restricting (re-)use or distribution, with the only requirement being attribution. This new licence scheme was based on the *Creative Commons* (CC) licences, another licence scheme widely used for open data.[12] Another example of open licence schemes used for data is the Open Data Commons Open Database License (ODbL), which is for example used for OpenStreetMap data.[13] (For further discussion on IPRs see OECD [2015], Inquiries into Intellectual Property's Economic Impact, OECD, forthcoming).
- *Pricing* – Although pricing will have less of an impact on the degree of openness than technological design and IPRs, it can nevertheless be one of the most challenging factors, because optimal pricing can be hard to determine. Many governments, for example, wish to engage in cost recovery, partly for budgetary reasons and partly based on the principle that those who benefit should pay. But the calculation of benefits can be problematic due to significant spillover effects through the creation of public and social goods based on open data. Furthermore, as Stiglitz et al. (2000) have argued, if government provision of a data-related service is a valid role, generating revenue from that service is not. Many open data initiatives therefore encourage the provision of data "at the lowest possible cost, preferably at no more than the marginal cost" as stated in the OECD (2005)

Recommendation on Principles and Guidelines for Access to Research Data from Public Funding. The OECD (2008) *Council Recommendation on Enhanced Access and More Effective Use of Public Sector Information* further specifies that "where possible, costs charged to any user should not exceed marginal costs of maintenance and distribution, and in special cases extra costs for example of digitisation". While marginal cost pricing is often considered the best option for the public sector, that option is seen as unattractive for the private sector, for which at least cost recovery is a necessity. This can lead to average cost pricing as an alternative pricing model, or can even require complex revenue models including subscription fees, freemium[14] and voluntary donations, in combination with cross-subsidies.

Box 4.2. **Database protection**

Databases are protected by copyright under certain circumstances, but in some countries – namely in the European Union, Japan and South Korea – they are also protected by a so-called sui generis database right (SGDR) aimed at protecting the investment.

The Berne Convention does not mention databases, but provides protection for collections of literary or artistic works such as encyclopaedias and anthologies that, by reason of the selection and arrangement of their contents, constitute intellectual creations.[1] The plain meaning of that provision seems to exclude from protection collections that do not consist of works, which is to say that collections of data (databases) are not covered by Art. 2(5). It has been argued that collections of data are in fact covered by the general provision of Art. 2(1) as "literary and artistic works".

In any event, currently the protection afforded to databases (as collections of data or other elements) is established – or confirmed – by both Art. 10(2) of the Agreement on Trade-Related Aspects of Intellectual Property Rights (TRIPS) and the almost identical Art. 5 of the WIPO Copyright Treaty: "Compilations of data or other material, whether in machine readable or other form, which by reason of the selection or arrangement of their contents constitute intellectual creation shall be protected as such..."[2]

An additional layer of protection is found in some countries and is afforded to databases regardless of the intellectual creation (i.e. "selection or arrangement") that may or may not be present. What is protected here is the investment in generating the database, i.e. in the obtaining, verification or presentation of the data. This type of right, also known as the sui generis database right mentioned above, is found in the EU Database Directive and the laws of a number of other countries, and will be dealt with below. It should be borne in mind that while the protection afforded to original databases focuses on the arrangement or selection without extending to the content of the database, the SGDR offers protection against the copying of substantial parts of the database – that is to say it extends, at least to some extent, to the data themselves.

1. See Art. 2(5) of the Berne Convention available at www.wipo.int/treaties/en/text.jsp?file_id=283698.

2. See Art. 10(2) of the TRIPS Agreements at www.wipo.int/wipolex/en/other_treaties/text.jsp?file_id=305907.

Source: OECD (2015), *Inquiries into Intellectual Property's Economic Impact?* (forthcoming), Chapter 7, "Legal Aspects of Open Access to Publicly Funded Research".

The three factors presented above (technological design, IPRs and pricing) determine the degree of openness, which can range from *closed* (access only by the data controller) to *open to the public* at its two extremes. In between, access may be restricted to i) individual stakeholders who can affect or are affected by the use of the data, with

access typically being granted on discriminatory bases, and to ii) specific communities (see the OECD 2005 Recommendation on Principles and Guidelines for Access to Research Data from Public Funding), with access being restricted to the "international research community". This leads to a three-level definition of open access, as illustrated in Figure 4.1.

Figure 4.1. **The data common continuum**

Overall, open data can be an optimal (private and social) strategy for maximising the benefits of data, in particular in environments characterised by high uncertainty, complexity and dynamic evolution such as climate change, urban development and health care research. These complex systems are often characterised by complementary effects; non-discriminatory access can be a means of internalising them by encouraging "experimentation and innovation among complementary applications" (Frischmann, 2012).

There are a number of other factors affecting the degree of openness: confidentiality and privacy considerations may be justifications for limiting data access in some cases as well. Furthermore, access problems and issues at the international level can emerge due to differences in culture and legislations. OECD (2013d) discusses the following factors in the particular context of science, but they are valid for other domains as well:

- *Legal and cultural barriers* – Depending upon the perceived sensitivity of the data and/or the legal framework governing data-sharing arrangements, some departmental "gatekeepers" can regulate access conditions tightly.
- *Public concerns* – To date there has been relatively little public engagement to explain the potential of data linkage, or the methods that are used to protect individual confidentiality when such linkages are made.
- *Technical barriers* – While various models for secure data access exist in some countries, the expertise, hardware and software to implement secure access is unevenly distributed among countries.

Finally, the provision of high-quality data can require significant time and up-front investments before the data can be shared. These include the costs related to i) datafication, ii) data collection, iii) data cleaning and iv) data curation. Effective knowledge sharing is, however, not limited to sharing data. In many cases a number of complementary resources may be required, ranging from additional (meta-)data to data models and algorithms for data storage and processing, and even secured IT infrastructures for (shared) data storage, processing, and access. For example, data from the distributed array telescope may create large data sets, which however require additional data on the direction of the telescopes to be interpreted correctly.

Given these significant costs, creators and controllers of data do not necessarily have the incentives to share their data. One reason is that the costs of data sharing are perceived as higher than the expected private benefits of sharing. Also, since data are in principle non-exclusive goods for which the costs of exclusion can be high, there is the possibility that some may "free ride" on others' investments. The argument that follows is that if data are shared, free-riding users can "consume the resources without paying an adequate contribution to investors, who in turn are unable to recoup their investments" (Frischmann, 2012). In science and research the situation poses even more incentive problems, as scientists and researchers traditionally compete to be first to publish scientific results, and may (a third disincentive) not enjoy or even perceive the benefits of disclosing the data they could further use for as yet uncompleted research projects (see Chapter 7 of this volume).

The root of these incentive problems can be summarised as a positive externality issue: data sharing may benefit others more than it benefits the data creator and controller, who cannot privatise these benefits and as a result may not sufficiently invest in data sharing or may even refrain completely. However, the idea that positive externalities and free riding always diminish incentives to invest has been challenged by some:

> *There is a mistaken tendency to believe that any gain or loss in profits corresponds to an equal or proportional gain or loss in investment incentives, but this belief greatly oversimplifies the decision-making process and underlying economics and ignores the relevance of alternative opportunities for investment. The conversion of surplus realised by a free rider into producer surplus may be a wealth transfer with no meaningful impact on producers' investment incentives or it may be otherwise, but there is no theoretical or empirical basis for assuming that such producer gains are systematically incentive-relevant.* (Frischmann, 2012)

Such an assumption therefore cannot be generalised, and needs careful case-by-case scrutiny. Indeed, free riding is sometimes the economic and social rationale for providing access to data. Open data, for example, is motivated by the recognition that users will free ride on the data provided, and in so doing will be able to create a wide range new goods and service that were not anticipated and otherwise would not be produced. In that sense, according to Frischmann, "free riding is pervasive in society and a feature, rather than a bug" (2012).

Data portability and interoperability

Data are rarely harmonised across sectors or organisations, as individual units collect and/or produce their own set of data using different metadata, formats and standards. Even if access to data is provided, the data may not be able to be reused in a different context for new applications. Reusability will typically be limited if data are not machine readable and cannot be reused across IT systems (interoperability). Some data formats that are considered machine readable are therefore based on open standards, such as RDF (Resource Description Framework), XML (eXtensible Markup Language), and more recently JSON (JavaScript Object Notation). Other standards include file formats such as CSV (comma-separated values) and proprietary file formats such as Microsoft Excel. Unresolved interoperability issues are, for example, still high on the e-government agendas of many OECD countries (see Chapter 7 of this volume). For instance, interoperability of data catalogues, or the creation of a pan-European data catalogue, is an important challenge currently faced by EU policy makers.

An important development in the context of data portability and interoperability is the increasing role of consumers in the data-sharing ecosystems. In enabling their personal data to flow across organisations, consumers are playing an important role that derives from their access to their own data under the Individual Participation Principle of the OECD (2013b) *Recommendation of the Council concerning Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data* (OECD Privacy Guidelines). Furthermore, the Individual Participation Principle grants individuals the right "to challenge data relating to [them] and, if the challenge is successful to have the data erased, rectified, completed or amended" (subject to regulatory obligations, e.g. to keep billing information, etc.). This is a right they could exert when porting their data from one controller to another.

Government initiatives are promoting data portability and thus contributing to the free flow of data as well. In 2011, a government-backed initiative called midata was launched in the United Kingdom to help individuals access their transaction and consumption data in the energy, finance, telecommunications and retail sectors. Under the programme, businesses are encouraged to provide their customers with their consumption and transaction data in a portable, preferably machine readable format. A similar initiative has been launched in France by Fing (Fondation Internet Nouvelle Génération), which provides a web-based platform, MesInfos,[15] for consumers to access their financial, communication, health, insurance and energy data that are being held by businesses. Both the UK and French platforms are outgrowths of ProjectVRM,[16] a US initiative launched in 2006 that provides a model for Vendor Relationship Management by individual consumers. Last but not least, the right to data portability proposed by the European Commission in the current proposal for reform of its data protection legislation aims at stimulating innovation through more efficient and diversified use of personal data, by allowing users "to give their data to third parties offering different value-added services" (EDPS, 2014).

The initiatives discussed above show promise in terms of helping individuals make informed decisions and increasing trust in the data-intensive services that organisations seek to deliver. But such programmes may also bring significant costs with regard to both developing and maintaining the mechanisms for enhanced data access and complying with relevant regulations (Field Fisher Waterhouse, 2012). The question arises: who should bear these costs?

Data linkage and integration

The value of data is, as stated above, highly context dependent – it increases when the data can be linked with and integrated into other data sets. As data are placed in a larger context, they can reveal additional insights that otherwise were not possible to gain. This is for instance true with linked micro data sets, as the example of the Micro-Data Lab of the OECD Directorate for Science, Technology and Innovation (DSTI) demonstrated, where data on firms' innovation performance (e.g. patent applications) are linked with data on their economic performance (e.g. financial statements). Linked data thus create super-additive value, which is greater than the sum of its parts (i.e. of data silos).

There are various reasons why linking data across different silos may be challenging. Some are obviously related to the legal, cultural and technical barriers to data access and sharing, as highlighted above. Others may be related to skills barriers. As OECD (2013d) highlights: "even though techniques for record linkage are now well developed, and are used by numerous organisations regularly, the capacity with which to carry out successful

linkages may be in short supply". Also, some of the barriers to data linkage are legitimate, since linkage can undermine privacy protective measures such as anonymisation and pseudonymisation, as highlighted in Chapter 5 of this volume.

Data quality and curation

The information that can be extracted from data depends on the quality of the data, and data quality in turn depends on the intended use. "If data [are] accurate, they cannot be said to be of good quality if they are produced too late to be useful, or cannot be easily accessed, or appear to conflict with other data" (OECD, 2011). Thus, data quality needs to be viewed as a multi-faceted concept. The OECD (2011) defines the following seven dimensions:

1. Relevance – "is characterised by the degree to which the data [serve] to address the purposes for which they are sought by users. It depends upon both the coverage of the required topics and the use of appropriate concepts".
2. *Accuracy* – is "the degree to which the data correctly estimate or describe the quantities or characteristics they are designed to measure".
3. *Credibility* – "the credibility of data products refers to the confidence that users place in those products based simply on their image of the data producer, i.e. the brand image. Confidence by users is built over time. One important aspect is trust in the objectivity of the data".
4. *Timeliness* – "reflects the length of time between their availability and the event or phenomenon they describe, but considered in the context of the time period that permits the information to be of value and still acted upon. … Real-time data [are] data with a minimal timeliness".
5. *Accessibility* – "reflects how readily the data can be located and accessed", as discussed in the previous section on data access and sharing.
6. *Interpretability* – "reflects the ease with which the user may understand and properly use and analyse the data". The availability of metadata plays an important role here, as they provide for example "the definitions of concepts, target populations, variables and terminology, underlying the data, and information describing the limitations of the data, if any".
7. *Coherence* – "reflects the degree to which they are logically connected and mutually consistent. Coherence implies that the same term should not be used without explanation for different concepts or data items; that different terms should not be used without explanation for the same concept or data item; and that variations in methodology that might affect data values should not be made without explanation. Coherence in its loosest sense implies the data are 'at least reconcilable'".

The OECD Privacy Guidelines also provides a number of criteria for data quality in the context of privacy protection. The Recommendation states that "personal data should be relevant to the purposes for which they are to be used, and, to the extent necessary for those purposes, should be accurate, complete and kept up-to-date". So the data quality dimensions would have to include completeness as an eighth dimension according to the OECD Privacy Guidelines. Furthermore, the cost efficiency with which data are collected could also be considered as a measure for data quality. "Whilst the OECD does not regard

cost-efficiency as a dimension of quality, it is a factor that must be taken into account in any analysis of quality as it can affect quality in all dimensions" (OECD, 2011).

Data curation embodies those data management activities needed to assure long-term data quality across the data life cycle. Data curation thus includes activities affecting the eight dimensions of data quality presented above. As OECD (2013c) highlights, however, "these particular activities [...] are often beyond the scope and timeframe of original [...] projects" for which the data were initially collected and used. This can lead to disincentives for data curation and put at risk long-term access and reuse of data. In science and research, where the long-term quality of data is essential, data curation is seen as a key part of the provision of research infrastructure (OECD, 2013c).

Data ownership and control

Data ownership is a concept that is often misunderstood and/or misused. With businesses, for example, data ownership is often used to assign responsibility and accountability for specific databases (the "data owners"). In this context, ownership is perceived as a means of assuring data quality and curation, as well as data protection and security along the complete data life cycle. However, ownership is assigned without IPRs being granted to the "data owner" (Scofield, 1998; Chisholm, 2011). Scofield (1998) therefore suggests replacing the term "ownership" with "stewardship", as this better captures the responsibility that organisations are actually looking to promote with the ownership concept.

Granting private property rights is often suggested as a solution to the incentive problems related to free riding. The concept of ownership typically means "to have legal title and full property rights to something" (Chisholm, 2011). Data are an intangible asset; like other information-related goods, they can be reproduced and transferred at almost zero marginal costs. So in contrast to the concept of ownership of physical goods, where the owner typically has exclusive rights and control over the good – including for instance the freedom to destroy the good – this is not the case for intangibles such as data. For these types of goods, IPRs are typically suggested as the legal means to establish clear ownership. In the case of data in particular, legal regimes such as copyright as well as other IPRs applicable to databases and trade secrets can be used (see Box 4.2). Furthermore, technologies such as cryptography have dramatically reduced the costs of exclusion, and thus are often used as a means to protect data (see Chapter 5 of this volume).

However, in contrast to other intangibles, data typically involve complex assignments of different rights across different data stakeholders, requiring of some stakeholders "the ability to access, create, modify, package, derive benefit from, sell or remove data, but also the right to assign these access privileges to others" (Loshin, 2002). So in many cases, no single data stakeholder will have exclusive rights. Different stakeholders will typically have different powers depending on their role. As Trotter (2012) highlights in the case of health patient data, all stakeholders (including patient, doctor and programmer) "have a unique set of privileges that do not line up exactly with any traditional notion of 'ownership'. Ironically, it is neither the patient nor the [doctor] who is closest to 'owning' the data. The programmer has the most complete access and the only role with the ability to avoid rules that are enforced automatically by electronic health record (EHR) software". Loshin (2002) identifies the following data stakeholders that could claim data ownership:

- *creator* – the party that creates or generates data
- *consumer* – the party that uses the data
- *compiler* – the party that selects and compiles information from different information sources
- *enterprise* – all data that entering the enterprise or created within the enterprise is completely owned by the enterprise
- *funder* – the user that commissions the data creation and therefore claims ownership
- *decoder* – in environments where information is 'locked' inside particular encoded formats, the party that can unlock the information becomes an owner of that information"
- *packager* – the party that collects information for a particular use and adds value through formatting the information for a particular market or set of consumers
- *reader as owner* – the value of any data that can be read is subsumed by the reader and, therefore, the reader gains value through adding that information to an information repository
- *subject as owner* – the subject of the data claims ownership of that data, mostly in reaction to another party claiming ownership of the same data
- *purchaser/licenser as owner* – the individual or organisation that buys or licenses data may stake a claim to ownership.

In cases where the data are considered "personal data" the situation is even more complex, since certain rights of the data subject cannot be waived. For example, the Individual Participation Principle of the OECD (2013c) *Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data* recommends that individuals have "the right: a) to obtain from a data controller, or otherwise, confirmation of whether or not the data controller has data relating to him; b) to have communicated to him, data relating to him within a reasonable time; […] c) to be given reasons if a request made under sub-paragraphs (a) and (b) is denied, and to be able to challenge such denial; and d) to challenge data relating to him …". The rights of the data subject limit any possibility for exclusive right on the storage and use of the data.

There are also economic reasons why granting private property rights may not be the optimal solution in the case of data. As highlighted above, social welfare is maximised when a rivalrous good is consumed by the person who values it the most, while social welfare through the consumption of non-rivalrous goods is maximised when the good is consumed by everyone who values it. This additional degree of freedom suggests that other institutions such as commons and "data citations" (see Chapter 7 of this volume) may be more effective in maximising welfare while still providing sufficient incentive for the production and release of data. Furthermore, the free riding story can be "translated in game-theoretic terms into a prisoners' dilemma, another good story, although one that does not necessarily point to private property as a solution to the cooperation dilemma" (Frischmann, 2012).

Overall, "the concept [of ownership] doesn't map well to the people and organisations that have relationships with that data" (Trotter, 2012). Data ownership can be a poor starting point for data governance, and can even be misleading. As Croll (2011) points out: "The important question isn't who owns the data. Ultimately, we all do. A better

question is who owns the means of analysis? Because that's how […] you get the right information in the right place. The digital divide isn't about who owns data – it's about who can put that data to work".

Data value and pricing

The discussion has underlined that data have no intrinsic value; their value depends on the context of their use. In fact, information – more than any other good – is an experience good, i.e. a good that consumers must experience in order to value. "Virtually any new product is an experience good"; however, "information is an experience good every time it's consumed" (Shapiro and Varian, 1999). Data pricing schemes can thus be complex. In particular, the context dependency of data challenges the applicability of market-based pricing: that pricing assumes that markets can converge towards a price at which demand and offer meet, and such is not always the case.

As the OECD (2012) study "Exploring the Economics of Personal Data: A Survey of Methodologies for Measuring Monetary Value" showed, the monetary valuation of the same data set can diverge significantly among market participants. For example, while economic experiments and surveys in the United States indicate that individuals are willing to reveal their social security numbers for USD 240 on average, the same data sets can be obtained for less than USD 10 from US data brokers such as Pallorium and LexisNexis. Data pricing schemes based on cost structure seem to be a more common approach. As noted above, the OECD (2005) *Recommendation on Principles and Guidelines for Access to Research Data from Public Funding* – and the OECD (2008) *Council Recommendation on Enhanced Access and More Effective Use of Public Sector Information* – both encourage the provision of data "at the lowest possible cost, preferably at no more than the marginal cost" which can include the cost for "maintenance and distribution, and in special cases extra costs for example of digitisation".

4.4. Key findings and policy conclusions

Data are an infrastructural resource – a capital good that cannot be depleted and that can be used for a theoretically unlimited range of purposes. In particular, data enable multi-sided markets – which, combined with increasing returns to scale and scope – provide businesses with significant growth opportunities (see Chapter 4 of this volume). There are, however, data demand manifestation problems, which may lead to under-provision of data or the prioritisation of access and use for a narrower range of uses than would be socially optimal.

This calls for managing data based on non-discriminatory access regimes, including commons or open access regimes, because:

1. these regimes facilitate joint production or co-operation with suppliers, customers or even competitors
2. they support and encourage value-creating activities by users
3. they maximise the option value of data and data-related products when there is high uncertainty regarding sources of future market value
4. they are (cross-)subsidising the production of public and social goods, which otherwise would require governments or businesses to pick winners (users or applications) by assessing the right (social) demand for such goods based on the (social) value they create.

The provision of high-quality data can require significant up-front investments. These costs can sometimes exceed the private benefits expected from data sharing, and thus present a barrier to data sharing. The possibility of "free riding" on others' investments is sometimes seen as a source of additional incentive problems, although there are many cases where free riding had no significant disincentive effects on producing or sharing data (e.g. open data).

"Ownership" is a questionable appellation when it comes to data. In contrast to other intangibles, data typically involve complex assignments of different rights across different data stakeholders. Those different stakeholders will typically have different power over the data, depending on their role. In cases where the data are considered "personal data", the concept of data ownership by the party that collects personal data is even less practical since privacy regimes grant certain explicit control rights to the data subject, as for example specified by the Individual Participation Principle of the OECD (2013c) *Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data*.

Lack of data portability and interoperability are among the most challenging barriers to data reuse. This is particularly the case where data are not provided in a machine readable format and thus cannot be reused across IT systems. Individuals (consumers) play an important role in promoting the free flow of their personal data across organisations. Government and private sector initiatives such as midata (United Kingdom), MesInfos (France), and the proposed reform of EU data protection legislation are promoting data portability – and thus promoting the free flow of data across organisations – as a means of empowering individuals and consumers and strengthening their participation in DDI processes.

Even within organisations, especially large ones, data silos are perceived as a barrier to intra-organisational data sharing. According to a survey by the Economist Intelligence Unit (2012a), almost 60% of companies stated that "organisational silos" are the biggest impediment to using "big data" for effective decision making. Executives in large firms (with annual revenues exceeding USD 10 billion) are more likely to cite data silos as a problem (72%) than those in smaller firms (with revenues less than USD 500 million, 43%).

Better data governance regimes are needed to overcome barriers to data access, sharing and interoperability (subject to legitimate restrictions, such as privacy). These barriers are often faced by individuals, businesses and policy makers alike across sectors. Data governance regimes can have an impact on the incentives to share and the possibility of data to be used in interoperable ways. The elements to be considered for an effective data governance regime include:

- data access and reuse
- data portability and interoperability
- data linkage and integration
- data quality and curation
- data "ownership" and control
- data value and pricing.

Coherent guidelines are needed to promote better data governance across the economy. Many of the barriers to data access and reuse, for example, are common across

domains, including science and research (Chapter 7 of this volume), health care (Chapter 8) and smart cities (Chapter 9), and the public sector (see Chapter 10). Existing frameworks that promote better access to data, some of which are sector specific, may need to be reviewed and eventually consolidated to foster coherence among public policies related to data access, linkage and reuse. This would also include the OECD Council Recommendations promoting better access to data, including in particular the OECD (2008) *Recommendation of the Council for Enhanced Access and More Effective Use of Public Sector Information* of 30 April 2008, and the OECD (2006b) *Recommendation of the Council concerning Access to Research Data from Public Funding* of 14 December 2006, both of which are currently under review.

Notes

1 See www.tomtom.com/en_gb/licensing/products/traffic/historical-traffic/custom-travel-times.

2 See Klenow and Rodriguez-Clare (2005) for an excellent review of the most relevant theoretical models of technological spillovers and economic growth.

3 See OECD, *Main Economic Indicators*, "Sources and Definitions", http://stats.oecd.org/mei/default.asp?lang=e&subject=1, accessed 9 April 2015.

4 The System of National Accounts uses the term "fixed capital" (in contrast to circulating capital, such as raw materials) to refer to capital goods.

5 The prisoner's dilemma is a central part of game theory. It describes a game with two players ("prisoners") that have the opportunity to collaborate to achieve a high payout, or to betray each other for a lower payout. Both players make their choice wihout knowing the choice of the other player, and in case of no collaboration, it is the player who betrays that profits strongly. The prisoner's dilemma is therefore used to illustrate why "rational" individuals might not cooperate, even if collaboration is in their best interests.

6 Altogether, over 50% of the total potential value of open data (more than USD 3 trillion annually) is estimated to be generated from consumer and customer surplus (MGI, 2013). The total value of open data must exceed by far the benefits highlighted in MGI (2013), which attributes the largest share of the total benefits of open data to better benchmarking, "an exercise that exposes variability and also promotes transparency within organizations" (MGI, 2013). Better benchmarking would enable "fostering competitiveness by making more information available and creating opportunities to better match supply and demand" as well as "enhancing the accountability of institutions such as governments and businesses [to] raise the quality of decision [making] by giving citizens and consumers more tools to scrutinize business and government" (MGI, 2013).

7 "Costs and benefits are rarely known with certainty, but uncertainty can be reduced by gathering information. Any decision made now and which commits resources or generates costs that cannot subsequently be recovered or reversed, is an irreversible decision. In this context of uncertainty and irreversibility it may pay to delay making a decision to commit resources. The value of the information gained from that delay is the option value or quasi-option value." (OECD, 2006a)

8 The "super-additive" nature of linked data is of course not without its challenges as well. In particular, linked data sets can undermine confidentiality and privacy protection measures such as anonymisation and pseudonymisation.

9 See also OECD (2005) *Principles and Guidelines for Access to Research Data from Public Funding*, www.oecd.org/sti/sci-tech/38500813.pdf, accessed 12 June 2014.

10 The meeting was organised by Tim O'Reilly of O'Reilly Media and Carl Malamud of Public. Resource.Org. See https://public.resource.org/8_principles.html, accessed 7 November 2013.

11 See www.nationalarchives.gov.uk/doc/open-government-licence/version/2/.

12 See data.australia.gov.au, data.gv.at, and *Google Ngram Viewer,* http://storage.googleapis.com/books/ngrams/books/datasetsv2.html.

13 See www.openstreetmap.org/copyright.

14 This business model offers free service to customers, and a premium level of the service is available for a fee (see for example Dropbox).

15 See: http://fing.org/?-MesInfos-les-donnees-personnelles-&lang=fr.

16 See: http://cyber.law.harvard.edu/projectvrm/Main_Page.

References

Arthur, C. (2013), "'Data is the new oil': Tech giants may be huge, but nothing matches big data", *Raw Story*, 24 August, www.rawstory.com/rs/2013/08/24/data-is-the-new-oil-tech-giants-may-be-huge-but-nothing-matches-big-data/, accessed 6 April 2015.

Berners-Lee, T. (2006a), "Isn't it semantic?", www.bcs.org/content/conWebDoc/3337, accessed 4 May 2015.

Berners-Lee, T. (2006b), "Linked data", www.w3.org/DesignIssues/LinkedData, accessed 6 April 2015.

Bracy, J. (2013), "Changing the conversation: Why thinking 'Data Is the New Oil' may not be such a good thing", *Privacy Perspectives*, International Association of Privacy Professionals (IAPP), 19 July, https://privacyassociation.org/news/a/changing-the-conversation-why-thinking-data-is-the-new-oil-may-not-be-such/, accessed 6 April 2015..

Chisholm, M. (2011), "What is data ownership?", www.b-eye-network.com/view/15697, accessed 5 November 2014.

Corrado, C., C. Hulten and D. Sichel (2009), "Intangible Capital and U.S. Economic Growth", *Review of Income and Wealth*, Series 55, No. 3, September, www.conference-board.org/pdf_free/IntangibleCapital_USEconomy.pdf.

Croll, A. (2011), "Who owns your data", 11 January, http://news.yahoo.com/owns-data-20110112-030058-029.html, accessed 5 November 2014.

Deloitte (2013), "Data as the new currency: Government's role in facilitating the exchange", *Deloitte Review*, Issue 13, 24 July, http://cdn.dupress.com/wp-content/uploads/2013/07/DR13_data_as_the_new_currency2.pdf.

Dumbill, E. (2011), "Data is a currency: The trade in data is only in its infancy", *O'Reilly Radar*, http://radar.oreilly.com/2011/02/data-is-a-currency.html, accessed 6 April 2015.

EDPS (European Data Protection Supervisor) (2014), "Privacy and competitiveness in the age of big data: The interplay between data protection, competition law and consumer protection in the Digital Economy", March, https://secure.edps.europa.eu/EDPSWEB/webdav/site/mySite/shared/Documents/Consultation/Opinions/2014/14-03-26_competitition_law_big_data_EN.pdf, accessed 6 April 2015.

Engelsman, W. (2009), "Information assets and their value", University of Twente, http://referaat.cs.utwente.nl/conference/6/paper/6807/information-assets-and-their-value.pdf, accessed 6 April 2015.

Field Fisher Waterhouse (2012), "Will access to midata work?", http://privacylawblog.ffw.com/2012/will-access-to-midata-work, accessed 6 April 2015.

Frischmann, B.M. (2012), *Infrastructure: The Social Value of Shared Resources*, Oxford University Press.

Glanz, J. (2013), "Is big data an economic big dud?", *New York Times*, 17 August, www.nytimes.com/2013/08/18/sunday-review/is-big-data-an-economic-big-dud.html, accessed 6 April 2015.

Hardin, G. (1968), "The tragedy of the commons", *Science* (American Association for the Advancement of Science, AAAS) Vol. 162, No. 3859, pp. 1243-48, www.sciencemag.org/content/162/3859/1243.full.pdf.

IPC (Information and Privacy Commissioner of Ontario) (2000), "Should the OECD Guidelines apply to personal data online?", Report to the 22nd International Conference of Data Protection Commissioners, Venice, Italy, September, www.ipc.on.ca/images/resources/up-oecd.pdf, accessed 6 April 2015.

Jones, S. (2012), "Why 'Big Data' is the fourth factor of production", *Financial Times*, 27 December, www.ft.com/intl/cms/s/0/5086d700-504a-11e2-9b66-00144feab49a.html, accessed 26 January 2013.

Klenow, P. and A. Rodriguez-Clare (2005), "Externalities and Growth", in A. Philippe and S. Durlauf ed, *The Handbook of Economic Growth*, Elsevier, Amsterdam.

Kroes, N. (2012), "Digital agenda and open data: From crisis of trust to open governing", European Commission, SPEECH/12/149, 5 March, Bratislava, http://europa.eu/rapid/press-release_SPEECH-12-149_en.htm , accessed 6 April 2015.

Lazer, D. et al. (2014), "The parable of Google flu: Traps in big data analysis", *Science*, Vol. 343, 14 March, http://scholar.harvard.edu/files/gking/files/0314policyforumff.pdf.

Loshin, D. (2002), "Knowledge integrity: Data ownership", 8 June, www.datawarehouse.com/article/?articleid=3052, accessed 6 April 2015.

Loukides, M. (2014), "The backlash against big data, continued", *O'Reilly Radar*, 11 April, http://radar.oreilly.com/2014/04/the-backlash-against-big-data-continued-2.html, accessed 6 April 2015.

McNamee, R. (2009), "Obama needs to think bigger about infrastructure", Huffington Post, The Blog, 2 July, www.huffingtonpost.com/roger-mcnamee/obama-needs-to-think-bigg_b_156126.html, accessed 6 May 2015.

Merriam-Webster (2014), "Infrastructure", *Merriam-Webster.com,* 24 October, www.merriam-webster.com/dictionary/infrastructure, accessed 6 April 2015.

MGI (McKinsey Global Institute) (2013), "Open data: Unlocking innovation and performance with liquid information", McKinsey & Company, October, www.mckinsey.com/insights/business_technology/open_data_unlocking_innovation_and_performance_with_liquid_information, accessed 3 March 2014.

Moody, D. and P. Walsh (1999), "Measuring the value of information: An asset valuation approach", Seventh European Conference on Information Systems (ECIS'99), Copenhagen Business School, http://si.deis.unical.it/zumpano/2004-2005/PSI/lezione2/ValueOfInformation.pdf, accessed 6 April 2015.

Newman, N. (2013), "Taking on Google's monopoly means regulating its control of user data", *Huffington Post*, The Blog, 24 September, www.huffingtonpost.com/nathan-newman/taking-on-googles-monopol_b_3980799.html, accessed 6 April 2015.

NRC (1987), *Infrastructure for the 21st Century: Framework for a Research Agenda*, Committee on Infrastructure Innovation, National Research Council, National Academy Press, Washington, DC..

O'Neil, C. (2013a), "K-nearest neighbors: Dangerously simple", 4 April, *Mathbabe*, http://mathbabe.org/2013/04/04/k-nearest-neighbors-dangerously-simple/, accessed 6 April 2015.

OECD (2014), *Addressing the Tax Challenges of the Digital Economy*, OECD/G20 Base Erosion and Profit Shifting Project, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264218789-en.

OECD (2013a), "Knowledge networks and markets", OECD Science, Technology and Industry Policy Papers, No. 7, OECD Publishing, Paris, http://dx.doi.org/10.1787/5k44wzw9q5zv-en.

OECD (2013b), Recommendation of the Council concerning Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data, 11 July, C(2013)79, www.oecd.org/sti/ieconomy/2013-oecd-privacy-guidelines.pdf.

OECD (2013c), "New data for understanding the human condition: International perspectives", OECD Global Science Forum Report on Data and Research Infrastructure for the Social Sciences, February, OECD Publishing, Paris, www.oecd.org/sti/sci-tech/new-data-for-understanding-the-human-condition.pdf.

OECD (2012), "Exploring the economics of personal data: A survey of methodologies for measuring monetary value", *OECD Digital Economy Papers*, No. 220, OECD Publishing, Paris, http://dx.doi.org/10.1787/5k486qtxldmq-en.

OECD (2011), Quality Framework and Guidelines for OECD Statistical Activities, OECD Publishing, Paris, 17 January, http://search.oecd.org/officialdocuments/displaydocumentpdf/?cote=std/qfs%282011%291, accessed 6 April 2015.

OECD (2010), *The OECD Innovation Strategy: Getting a Head Start on Tomorrow*, OECD Publishing, Paris.

OECD (2008), Recommendation of the Council for Enhanced Access and More Effective Use of Public Sector Information, 30 April 2008, [C(2008)36], OECD Publishing, Paris, www.oecd.org/internet/ieconomy/40826024.pdf.

OECD (2006a), "Quasi Option Value", in OECD, *Cost-Benefit Analysis and the Environment: Recent Developments*, OECD Publishing, Paris. DOI: http://dx.doi.org/10.1787/9789264010055-11-en.

OECD (2006b), OECD Recommendation of the Council concerning Access to Research Data from Public Funding, 14 December 2006, [C(2006)184], OECD Publishing, Paris.

OECD (2005), *Principles and Guidelines for Access to Research Data from Public Funding*, OECD Publishing, Paris, www.oecd.org/sti/sci-tech/38500813.pdf.

OECD (2004), Declaration on Access to Research Data from Public Funding, OECD Publishing, Paris, www.oecd.org/science/sci-tech/sciencetechnologyandinnovationforthe21stcenturymeetingoftheoecdcommitteeforscientificandtechnologicalpolicyatministeriallevel29-30january2004-finalcommunique.htm, accessed 6 April 2015.

Oppenheim, C., J. Stenson and R. Wilson (2004), "Studies on information as an asset III: Views of information professionals", *Journal of Information Science*, Vol. 30, No. 2, pp. 181-90.

Rochet, J.-C. and J. Tirole (2006), "Two-sided markets: A progress report", RAND Journal of Economics, RAND Corporation, Vol. 37, No. 3, pp. 645-67, http://ideas.repec.org/a/bla/randje/v37y2006i3p645-667.html.

Rose, C. (1986), "The comedy of the commons: Custom, commerce, and inherently public property", Faculty Scholarship Series, Paper 1828, http://digitalcommons.law.yale.edu/fss_papers/1828, accessed 6 April 2015.

Rotella, P. (2012), "Is data the new oil?", *Forbes*, 2 April, www.forbes.com/sites/perryrotella/2012/04/02/is-data-the-new-oil/, accessed 6 April 2015.

Savitz, E. (2011), "The new factors of production and the rise of data-driven applications", *Forbes*, www.forbes.com/sites/ciocentral/2011/10/31/the-new-factors-of-production-and-the-rise-of-data-driven-applications/, accessed 13 December 2013.

Schwartz, J. (2000), "Intel exec calls for e-commerce tax", *The Washington Post*, 6 June.

Scofield, M. (1998), "Issues of data ownership", www.information-management.com/issues/19981101/296-1.html, accessed 6 April 2015.

Shapiro, C. and H.R. Varian (1999), *Information Rules: A Strategic Guide to the Network Economy*, Harvard Business Press, Boston, MA.

Steinmueller, W.E. (1996), "The US software industry: An analysis and interpretative history", in David C. Mowery (ed.), *The International Computer Software Industry*, Oxford University Press.

Stiglitz, J., P. Orszag and J. Orszag (2000), "Role of government in a digital age", Computer and Communications Industry Association, October, www.ccianet.org/CCIA/files/ccLibraryFiles/Filename/000000000086/govtcomp_report.pdf, accessed 10 October 2013.

Thorp, J. (2012), "Big data is not the new oil", *HBR Blog Network*, 30 November, http://blogs.hbr.org/cs/2012/11/data_humans_and_the_new_oil.html, accessed 6 April 2015.

Trotter, F. (2012), "Who owns patient data? Look inside health data access and you'll see why 'ownership' is inadequate for patient information", O'Reilly Strata, 6 June, http://strata.oreilly.com/2012/06/patient-data-ownership-access.html, accessed 6 April 2015.

Ubaldi, B. (2013), "Open government data: Towards empirical analysis of open government data initiatives", *OECD Working Papers on Public Governance*, No. 22, OECD Publishing, Paris, http://dx.doi.org/10.1787/5k46bj4f03s7-en.

UN (2008), *System of Nation Accounts 2008*, United Nations, http://unstats.un.org/unsd/nationalaccount/docs/SNA2008.pdf, accessed 6 April 2015.

UN Global Pulse (2012), "Big data for development: Opportunities & challenges", United Nations Global Pulse, May, www.unglobalpulse.org/sites/default/files/BigDataforDevelopment-UNGlobalPulseJune2012.pdf, accessed 6 April 2015.

United Kingdom Cabinet Office (2012), "Open data white paper: Unleashing the potential", June, http://data.gov.uk/sites/default/files/Open_data_White_Paper.pdf, accessed 6 April 2015.

Zax, D. (2011), "Is personal data the new currency?", *MIT Technology Review*, 30 November, www.technologyreview.com/view/426235/is-personal-data-the-new-currency/, accessed 6 April 2015.