Chapter 2

Mapping the global data ecosystem and its points of control

In exploring the rapidly evolving data ecosystem, this chapter enumerates the key actors, their main technologies and services, and their business and revenue models. It uses a layer model to identify these actors as well as strategic points of control in the system. It goes on to discuss the interaction among actors, analysing in particular the relation between competition and collaboration for DDI, and how this "co-opetition" translates in terms of horizontal and vertical dynamics. The chapter analyses the degree to which data ecosystems are open, global and interconnected. Finally, it looks at the implications of DDI for global value chains (GVCs) and trade, taxation, and competition.

The great thing about big data is that there's still plenty of room for new blood, especially for companies that want to leave infrastructure in the rearview mirror. (Harris, 2012)

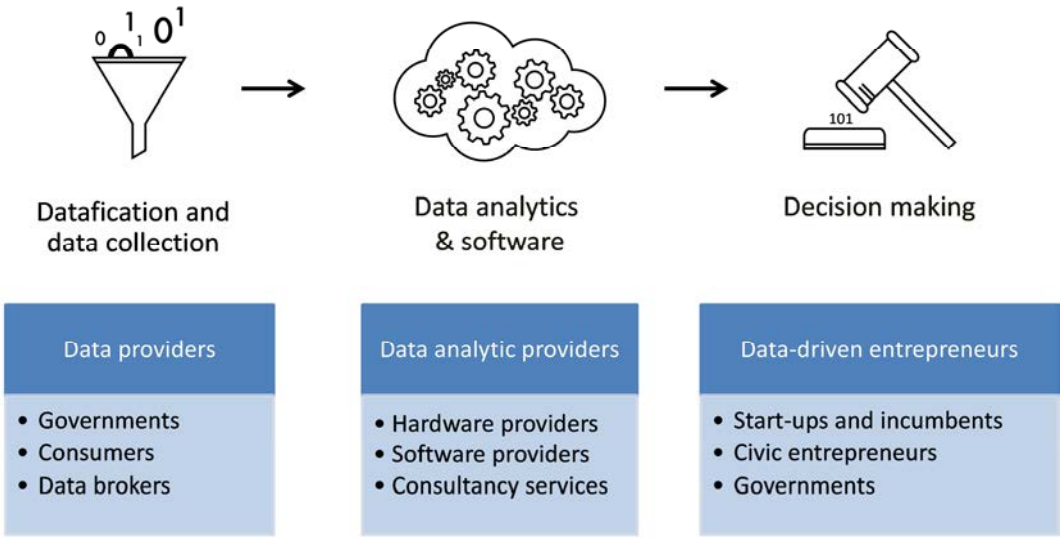
I look at this audience, and I look at VMware and the brand reputation we have in the enterprise, and I find it really hard to believe that we cannot collectively beat a company that sells books. (VMware's President and COO Carl Eschenbach, VMware Partner Exchange conference, February 2013)

Data-driven innovation – DDI, introduced in Chapter 1 of this volume – refers to the use of data and analytics to improve or foster new products, processes, organisational methods and markets. It is the concrete fulfilment of the value creation process along the data value cycle (see Figure 1.7 in Chapter 1), embarked upon in order to reach a specific goal, tackle a problem, or grasp an opportunity for which data analytics could provide (a part of) the solution. Each specific goal will require an organisation (or a consortium of organisations) to organise a value creation process along the data value cycle. It is likely that for many of the steps in this process, organisations will have to involve third parties around the world, because they lack experience, technological resources and/or talent to deal with the multidisciplinary aspects of data and analytics on their own. The resulting global value chain (GVC) is in most cases specifically tailored towards the goal that is being pursued. The combined effect is that a global data ecosystem is emerging in which, more than ever before, data and analytic services are traded and used across sectors and across national borders. For the information and communication technology (ICT) industry this represents a USD 17 billion business opportunity for 2015, with an estimated market growth of more than 40% on average every year since 2010 (see IDC, 2012; Kelly, 2013).[1]

Better analysis of both the economic and societal impacts of DDI requires a deeper understanding of the complexity and dynamics of the emerging global data ecosystem – including the interaction between the actors, their technologies and their business models, and the dynamics that structure this ecosystem. The concept of an ecological approach to describe business environments chosen for the analysis in this chapter was introduced by Moore (1993) to describe how companies should not be viewed as members of a single industry "[…] but as part of a business ecosystem that crosses a variety of industries." In these ecosystems, collaborative arrangements of firms combine their individual offerings to create coherent, customer-facing solutions (Adner, 2006). This is an appropriate perspective with which to explore the dynamics of networks of human and non-human actors, that have started to form around specific outcomes of DDI, and that may gradually link together into an all-encompassing global data ecosystem.

This chapter analyses that ecosystem, using the data value cycle introduced in Chapter 1 as a framework for identifying the different types of companies and services competing within it (Figure 2.1). The chapter also analyses other factors affecting the functioning of the data ecosystem such as key technologies, business models, and coalitions/alliances that are forming. By mapping actors and their technologies and business models using a "follow the data" approach along the data value cycle, the chapter reveals links across sectors, potential points of control in the data ecosystem, and their gatekeepers that mediate the interactions that shape the ecosystem. Newly forming GVCs will point to new actors and emerging horizontal data markets. This chapter builds on a rich mix of comprehensive expert interviews, case studies and workshops, all conducted by TNO, Netherlands Organisation for Applied Scientific Research.[2]

Figure 2.1. **Main phases of the data value cycle with their key types of actors**

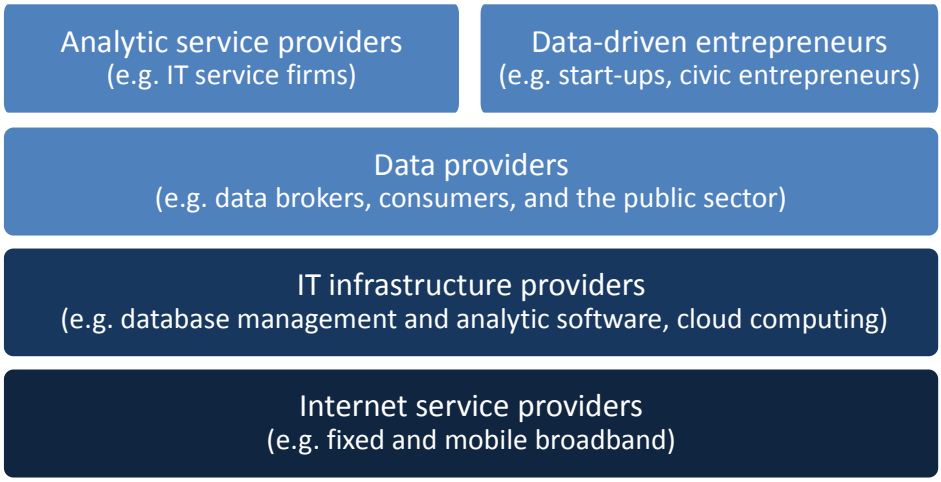


2.1. The key actors and their main technologies, services and business models

With the increase in data generated, collected and stored, and a greater variety of information that can be extracted from this data, companies in the data ecosystem are mushrooming. The number of actors, goods and services, and technologies and business models that shape the data-driven constellations creates a rapidly evolving data ecosystem. Describing this data ecosystem is problematic due to its constant evolution. It is a dynamic field as new technologies and practices are constantly being developed, driven largely by traditional information technology (IT) businesses in infrastructure and business analytics (e.g. IBM, Oracle, SAP and Microsoft) and a vast array of new start-ups. The increasingly active role of non-traditional data companies in the data ecosystem is also remarkable. One of the most telling examples is Amazon. Its role as a big data powerhouse is clearly illustrated by this chapter's opening quotation from VMware's President and COO Carl Eschenbach in response to Amazon's rise at the expense of VMware (Assay, 2013).

There are various depictions of the data ecosystem that position the different types of actors. Turck (2014), for example, illustrates the different clusters of businesses based on a detailed typology of services, products and technologies, including: i) the underlying core technologies, such as Hadoop and Cassandra; ii) the IT infrastructure (e.g. storage and computing); iii) the analytical tools (e.g. "R"); and iv) domain-specific applications. In this chapter, the data ecosystem is seen as a combination of layers of key roles of actors, where the underlying layers provide goods and services to the upper layers (Figure 2.**2**).

Figure 2.2. **The data ecosystem as layers of key roles of actors**



The following sections describe the different roles of actors, their technologies and services, and their main business models (including their revenue models) in more detail. What is provided is a generalised overview of the field as of 2014, with the understanding that as new actors enter and technologies evolve, so will the relative positions of the various players. The different roles include:

1. *Internet service providers* – Internet service providers form the backbone of the data ecosystem through which data are exchanged.
2. *IT infrastructure providers* – The second layer includes IT (hardware and software) infrastructure providers that are offering data management and analysis tools and critical computing resources, including but not limited to data storage servers, database management and analytic software, and most importantly cloud computing resources.
3. *Data (service) providers* – The third layer includes i) data brokers and data marketplaces that are selling their data across the economy, ii) the public sector with its open data initiatives (see Chapter 10 of this volume), and – last but not least – iii) consumers that are *actively* contributing their data to the data ecosystem increasingly as well, thanks to new services provided by innovative businesses but also through data portability initiatives (Chapters 4 and 5).
4. *Data analytic service providers* – The fourth layer includes businesses that provide data aggregation and analytic services, mainly to business customers. This also includes data visualisation services.
5. *Data-driven entrepreneurs*[3] – These entrepreneurs build their innovative businesses based on data and analytics available in the ecosystem. Their efforts result in DDI for science and research (see Chapter 7), health care (Chapter 8), smart cities (Chapter 9), and public service delivery (Chapter 10).

Before looking at each type of actor separately and in more detail, it is important to acknowledge five important characteristics of the data ecosystem that can only partially be reflected in an analysis based solely on the typology of the key actors.

1. Figure 2.2 implicitly suggests that firms can only be assigned to one particular role. However, a closer look at the business model of the key actors reveals that many businesses will typically play multiple roles. Internet service providers (ISPs), for example, are increasingly using data analytic services to manage their networks (OECD, 2014a),[4] but also to generate data on, for example, communication patterns that are offered to third parties. In the latter case, ISPs are acting as data service providers. For example, the French mobile ISP Orange uses its Floating Mobile Data (FMD) technology to collect mobile telephone traffic data that are anonymised and sold to third parties, including government agencies and traffic information service providers. Furthermore, IT infrastructure providers may furnish the full stack of hardware and software solutions needed for data analysis including through cloud-based services, on top of which access to third parties' data are also provided. Microsoft's cloud service Microsoft Azure, for example, is provided with the Azure Data Marketplace, where users can access data sets provided by third parties. These multiple roles of actors not only challenge measurement efforts for statistical purposes (see Box 2.1), but also point to the significant share of vertically integrated companies such as Google and Microsoft, and most importantly to the dual nature of many actors in the data ecosystem as users and producers of data and analytics.[5] This dual role suggests that the data ecosystem is a logical continuation of the Web 2.0.[6]
2. Figure 2.2 does not reflect the inherently global nature of the data ecosystem. The data ecosystem involves cross-border data flows due to the activities of key global actors and the global distribution of technologies and resources used for value creation. In particular, ICT infrastructures used to perform data analytics, including the data centres and software, will rarely be restricted to a single country, but will be distributed around the globe to take advantage of several factors; these can include local work load, the environment (e.g. temperature and sun light), and skills and labour supply (and costs). Moreover, many data-driven services developed by entrepreneurs "stand on the shoulders of giants" who have made their innovative services (including their data) available via application programming interfaces (APIs), many of which are located in foreign countries (see Chapter 3).
3. Related to the global nature of the data ecosystem is the missing representation in Figure 2.2 of the "cytoplasm" that lies between the layers of the data ecosystem and that enables the smooth *interoperability* of the different types of actors, their technologies, and services. Open Internet standards such as TCP/IP and HTTP have been and still are crucial for the global data ecosystem, which relies heavily on the open Internet for its functioning. In addition, the reuse of data and of data-driven services underlines the importance of (open) standards related to e.g. APIs and data formats (including meta-information and data), which are a requirement for the interoperability of data-driven services and the portability of data across these services.

Box 2.1. **The challenge in measuring "big data"-related industries**

In 2012, the OECD[1] undertook efforts to measure the value-added of big data-related activities identified from a National Accounts (NA) perspective. The work, based on a questionnaire submitted to 25 countries,[2] highlighted issues in attempting to derive estimates of the size of these activities.

Big data-related industries were identified as those industries collecting, processing and diffusing digital data. These industries were then retraced in the international classification of economic activities (ISIC) at the finest available level of disaggregation (i.e. at the class level, corresponding to four digits). This way of proceeding aims at producing estimates that would eventually be reliable and comparable with other NA aggregates and across countries. The result was an operational definition of big data-related industries. With reference to the latest release of ISIC (Rev.4), these activities would fall into the classes 5812: *Publishing of directories and mailing lists*; 5819: *Other publishing activities*; 6311: *Data processing, hosting and related activities*; and 6312: *Web portals*. ISIC Rev.4 offers a finer classification of information activities with respect to the previous Rev.3.1, and groups them together under Section "J". Nonetheless, the correspondence is far from perfect. Indeed, the above ISIC classes also include activities outside the data industry aggregate, such as web hosting (under class 6311), or publishing (under 5812 and 5819).

Overall, the *share* of value-added of digital data industries in the United States alone appeared to be much higher than in all the other countries considered together. This result was partly attributable to the likely higher development of these industries in the United States, but also suggested the need for a more comprehensive assessment when undertaken by NA. Additionally, data for some countries did not include those of all industries in the *digital data* aggregate. In particular, a number of respondents could not provide information on activities in the *publishing* industries classes. Regarding the relative weight of big data-related activities for other countries, one could observe that some "plausible" figures also deviated from what would be expected. Employment figures showed a pattern similar to those observed for value-added, with several data missing and wide country variations.

This first exploration provided a very preliminary perspective on the size of big data-related activities. Given the tiny size of the aggregate and the high variability across countries, these results were far from reliable, and not simply because of lack of coverage. In general, for all countries but the United States the source of data lay within the domain of Structural Business Statistics. This implied that in principle, data would include only businesses whose *main activity* fell in a given class – excluding those performing, say, data processing as a secondary activity (multi-product enterprises) or those in which data are instrumental to their main activity ("own account", as for e.g. financial firms). The coverage of publishing activities (ISIC classes 5812 and 5819) was not complete for all responding countries, while in some cases only aggregates for data processing and hosting (6311) and web-related activities (6312) were available. The above elements led to substantial underestimation of the size of big data-related activities in some countries. However, other elements of as yet unknown magnitude also influenced measurement in the opposite direction, such as the weight of activities included in the above classes that were not related to big data.

Box 2.1. **The challenge in measuring "big data"-related industries (cont.)**

At this stage, results based on NA are therefore considered an indication of the work that remains to be done, rather than a first approximation of the size of the industry. Relevant improvements could be achieved by harmonising the information produced, with specific attention paid to big data-related activities in statistical production. A more precise definition of industry boundaries, such as at the six-digit NAICS (the North American Classification), could be envisaged in the future. In principle, where estimates are to be performed by NA, this could also include secondary and own-account activities; recent evidence (including the study by Bakhshi and Mateos–Garcia, 2012) shows these to be a relevant component of the digital data aggregate. Practically speaking, NA estimates based on integration techniques are not immediately feasible. A necessary prerequisite – and at the same time a good result in itself – would be to achieve more precision in business statistics. In many cases the infrastructure is available, as published data are already based on five- or six-digit information; it could be reinforced at Kind of Activity Unit / Establishment level, be linked to product classifications, and be extended to working hours. In this respect, the exclusion of some activities such as webhosting and certain printing activities is highly recommended. Also, a thorough check (and agreement) on which activities are (to be) reported in individual classes would be beneficial, with the possible exclusion of specific activities which "by definition" are in the big data domain but seem to be substantively different. These operations, together with closer monitoring of data robustness by national statistic offices (NSOs), could help improve data quality. However, they may require significant time and efforts by NSOs.

1. The work was undertaken by the OECD Working Party on Measurement and Analysis of the Digital Economy (WPMADE, formerly the Working Party on Indicators for the Information Society, WPIIS) and presented at the December 2012 meeting of the working party.

2. Twenty-on of the responses where provided with detailed data set in ISIC Rev.3.1 and ISIC Rev.4.

4. Figure 2.2 does not reflect the fact that the data ecosystem relies on a variety of business models that may not necessarily be linked to the role of the actors within the ecosystem, but rather to the market segment targeted by these actors. To recall, a business model specifies the value proposition of a business, including its key activities and the goods and services (i.e. products) it offers. The business model also specifies the targeted market segment and most importantly the *revenue models* that describe how the business turns the value of their products into revenues. Analysis of business models of companies in the data ecosystem suggests that the selection of revenue models in the ecosystem mainly depends on whether the business model focuses on business to business (B2B) or business to consumer (B2C) offers.[7] In addition, businesses in the data ecosystem use a diversity of revenue models, some of which are often combined to maximise revenues (see Box 2.2). The resulting complexity of the mechanisms through which revenues are generated has led to a number of policy challenges, such as the challenge of value attribution (with implications for taxation) and the challenge faced by competition authorities in defining the relevant markets. Both policy issues are discussed further below.

Box 2.2. **The diversity of revenue models in the data ecosystem**

Businesses in the data ecosystem use a diversity of revenue models, some of which are often combined to maximise revenues. The most common models include the following.

Freemium – The term "freemium" is a portmanteau of the words "free" and "premium". The *freemium* revenue model, one of the most dominant in the data ecosystem, seems to be particularly attractive to start-ups: products are provided free of charge, but money is charged for additional, often proprietary features of the product (i.e. *premium*). The freemium revenue model is often combined with the advertising-based revenue model for B2C offers, where the free product is offered with advertisement while the premium offer is advertisement-free.

Advertisement – Advertisement is most frequently used for B2C offers: products are offered free of charge or with a discount to users in exchange for required viewing of paid-for advertisements (OECD, 2014d) Increasingly, advertisement is provided based on the profile and/or location of the consumers. Advertisement-based revenue models are also used in multi-sided markets together with *cross subsidies*, where a service is provided for free or at a low price on one side of the market, but subsidised with revenues from other sides of the market.

Subscription – Subscription-based revenue models are by far the models most frequently used in the data ecosystem, for B2B offers in particular (among all the B2B business models of start-ups analysed by Hartmann et al. (2014), for example, 98% were subscription based). Examples of subscription-based models include regular (daily, monthly or annual) payments for access to the Internet, as well as access to digital content including data, news, music, video streaming, etc. The category also includes regular payments for software services and maintenance, hosting and storage, and customer "help" services. Subscription-based revenue models are often combined with the *freemium revenue model*, where the premium product is provided with a subscription (see above).

Usage fees – Usage fees are the second most frequently used revenue model used by start-ups in the data ecosystem. They are also a prominent revenue model for B2B offers. Usage fees are typically charged to customers for use of a particular (online) service – including most offers that are provided "as-a-Service" (XaaS), such as cloud computing based services for example (see section on IT infrastructure providers). These services are offered through a *pay-as-you go model*, where usage fees are charged for the actual use of the service.

Selling of goods (including digital content) – Asset sale is still used in the data ecosystem, mainly by IT infrastructure providers. But it is also used by service platform providers that sell sensor-equipped smart devices (including smartphones, smart meters and smart cars) as a source for generating data and delivering value-added services. Furthermore, it includes *pay-per-download* revenue models where users pay per item of download. These could include, for instance, data sets or other digital content such as e-books, videos, apps, games and music.

Selling of services – This revenue model includes the provision of traditional B2B services such as IT consultancy services, software development and maintenance and helpdesk support. It also includes a wide range of long-term B2B services provided by Internet intermediaries such as web hosting, domain registration, and payment processing. It thus overlaps with the revenue models that are based on subscriptions and usage fees often used for IT service contracts.

Licensing – This revenue model is often used to generate revenues from intangible assets that are protected through intellectual property rights (IPRs), such as patents and copyrights. Licensing may thus be used to monetise software and software components including algorithms, libraries and APIs. It may also be used for databases. However, evidence suggests that licensing may not be an essential revenue model for start-ups, although it may be an important for well-established IT providers including in particular software companies (among the 100 start-ups analysed by Hartmann et al. (2014), none has indicated licensing as a source of revenue).

Box 2.2. **The diversity of revenue models in the data ecosystem (cont.)**

Commission fees – This is mainly used in B2C markets by intermediaries that use data analytics to better match supply and demand. Payment often will be calculated on the basis of a percentage of the price of products supplied, and it will only be obtained when successfully matching supply and demand – that is, when successfully providing businesses with customers.

5. Finally, although most illustrations of the data ecosystem such as Figure 2.2 provide an extensive and useful overview of the most relevant roles of actors in the ecosystem, they tend to be strongly ICT sector biased. They describe data-related technologies, the various types of data-related products and services, and the companies that provide them. However, the analyses conducted by TNO (2013) and in other chapters of this volume strongly suggest that DDI is not just a technological (ICT supply-side) challenge. DDI also presents serious demand-side challenges: working processes, attitudes, changes in management and human resource (HR) policy. However, the services or products that support these organisational challenges are rarely represented, and often also too complex to be fully represented, in a simplified model of the data ecosystem such as Figure 2.2. Legal consultation for example is very important, especially for organisations that deal with personal data, and this kind of service is often provided by external legal advisors.

It should therefore be acknowledged that in focusing solely on technological aspects, the analysis of the data ecosystem presented in this chapter only accounts for a relatively small share of all the interactions and relationships within the data ecosystem. Any assessment of the ecosystem – in particular, quantitative assessment of its total market size – that does not consider this limitation risks underestimating its full size and impact.

Internet service providers

In general, Internet service providers (ISPs) build and operate networks, typically at the regional level. They grant subscribers (businesses and consumers) access to the Internet through physical transport infrastructure as they have the equipment and telecommunication network required for a point-of-presence on the Internet. This is necessary to allow users to access content and services on the Internet and content providers to publish or distribute data and information online (OECD, 2011a). ISPs thus help build the foundation of the data ecosystem as they provide local, regional and/or national (fixed and mobile) broadband coverage, or deliver backbone services for other ISPs.

Some ISPs are extending their product offer with for example web hosting, web-page design and consulting services related to networking software and hardware (OECD, 2011a). In this case, well-established ISPs can benefit from their established reputation to place themselves in new markets such as the IT service market (including e.g. cloud computing) in which consumers' trust plays an important role (Koehler, Anandasivam and Dan, 2010). Since 2010, Telefonica, Orange and Deutsche Telekom have launched cloud computing services targeting in particular small and medium-sized enterprises (SMEs) (Arthur D. Little, 2013). Some ISPs are going further up the value chain by providing data and analytic services. For example, the French mobile ISP Orange is acting as a data service provider by using its *Floating Mobile Data* (FMD) technology to collect mobile telephone traffic data; these determine speeds and traffic density at a given point in the road network, and deduce travel time or the formation of traffic jams. The

anonymised mobile telephone traffic data are sold to third parties, including government agencies, to identify "hot spots" for public interventions, but also to private companies such as Mediamobile, a leading provider of traffic information services in Europe.[8]

Another example is Telefónica, which in 2012 launched its new "big data business unit", Telefónica Dynamic Insights. This business unit, based in the United Kingdom, operates as an analytic service provider with the goal of providing companies and governments around the world with analytical insights based on mobile network and machine-to-machine (M2M) data. Its first product, *Smart Steps*, uses "anonymised and aggregated mobile network data to enable companies and public sector organisations to measure, compare, and understand what factors influence the number of people visiting a location at any time" (Telefónica, 2012).

The subscription model is the prevalent revenue model in the majority of OECD countries in which ISPs act as traditional Internet service providers. ISPs mostly charge a periodic – daily, monthly or annual – fee to subscribe to an unlimited service (OECD, 2011a). Other revenue models include those prepaid-based, or a combination of both subscription and prepaid. Prepaid models are commonly used by ISPs that meter their services, for example when mobile Internet access is offered. The price paid by the consumer is based on actual usage rates or a monthly subscription fee, with an additional amount charged for a data package (OECD, 2011a).

However, ISPs are currently debating whether the flat rate model will still be applicable in the future. Some ISPs have proposed differentiating among classes of Internet traffic (e.g. gold, silver bronze) or dedicating specific broadband capacity to certain applications. These plans are motivated by the rise of Internet traffic volume in particular due to the increased usage of video. Helping drive that increase are online streaming, such as Netflix offers in the United States and other countries, and online television, such as the BBC iPlayer in the United Kingdom and the Swedish company Magine TV that offers its service in Sweden, Germany and Spain (van der Berg, 2014). It has been argued that, if investments in networks continue to be made, the growth in traffic will not overwhelm networks since the growth rate of data traffic is strong but decreasing in relative terms (OECD, 2014a). In addition, ISPs appear to have been mostly unsuccessful in promoting a discriminatory pricing scheme. One reason put forward by content providers for not purchasing these services is, that their impact is mostly unknown as the ISPs control only part of the network. Furthermore, in a competitive market content providers may judge that ISPs will upgrade their networks when quality degrades to remain competitive with other ISPs (van der Berg, 2014).

IT infrastructure providers

The market for IT infrastructure comprises providers of both hardware and software. But most important for DDI are providers of databases and related technologies and services (management, security, transport, storage). These include in particular providers of platforms for distributed parallel data processing – such as Hadoop, which has almost become the standard technology to deal with more complex, unstructured large-volume data sets (Box 2.3). The importance of databases and related technologies and services is also reflected in estimates by IDC (2012), which suggest that "big data technology and services" will grow from USD 3 billion in 2010 to USD 17 billion in 2015. This represents a compound annual growth rate (CAGR) of almost 40%. Data storage technologies and services are estimated to be the fastest growing segment, followed by networking, and IT services, which explains the increasing role of IT equipment firms in this relatively new market (see section below on mergers and acquisitions, M&A).

Box 2.3. **Internet spillovers enabling data-driven innovation across the economy: The case of Hadoop**

Internet firms, in particular providers of web search engines, have been at the forefront in the development and use of techniques and technologies for processing and analysing large volumes of data. Google, in particular, inspired the development of a series of technologies after it presented *MapReduce*, a programming framework for processing large data sets in a distributed fashion, and *BigTable*, a distributed storage system for structured data, in a paper by Dean and Ghemawat (2004) and Chang et al. (2006) respectively. In 2006, the open source implementation of *MapReduce,* called *Hadoop,* emerged. Initially funded by Yahoo, *Hadoop* is now provided as an open source solution (under the Apache License) and has become the engine behind many of today's big data processing platforms. Beside Yahoo, Hadoop is ushering in many data-driven goods and services offered by Internet firms such as Amazon, eBay, Facebook, and LinkedIn. As mentioned above, even traditional providers of databases and enterprise servers such as IBM,[1] Oracle,[2] Microsoft[3] and SAP[4] have started integrating Hadoop and other related open source tools into their product lines, making them available to a wider number of enterprises including Walmart (retail), Chrevon (energy), and Morgan Stanley (financial services).

The key innovation of MapReduce is its ability "to take a query over a data set, divide it, and run it in parallel over many nodes" (Dumbill, 2010), often using (low-cost) commodity servers that can be distributed across different locations. This distribution solves the issue of data being too large to fit onto and to be processed by a single server. The data used for MapReduce also do not need to be relational or even to fit a schema, as is the case with the conventional (relational) SQL databases. Instead, unstructured data can be stored and processed. The standard storage mechanism used by Hadoop is therefore a distributed file system, called HDFS (Hadoop Distributed File System). On top of being distributed, HDFS is a fault tolerant file system that can scale up to dozens of petabytes (millions of gigabytes) of storage and can run with high data throughput on all major operating systems (Dumbill, 2010). However, other file systems are also supported by Hadoop, such as the Amazon S3 file system (used on Amazon's cloud storage service).

To simplify the use of Hadoop (and HDFS), additional open source applications have been developed or existing ones have been extended, some through the initiative of top Internet firms. HBase, for example, is an open source, non-relational (i.e. NoSQL) distributed database, also under the Apache Licence. HBase was modelled after Google's BigTable, and can run on top of HDFS or Hadoop. HBase is now, for example, currently used by Facebook for its Messaging Platform, which in 2010 had to support 15 billion person-to-person messages and 120 billion chat messages per month (Muthukkaruppan, 2010). Another example is Hive, an open source data warehouse infrastructure running on top of Hadoop, which was initially developed by Facebook to simplify management of structured data using a SQL-based language (HiveQL) for queries. Finally, analytical tools such as R, an open-source environment for statistical analysis, are increasingly being used in connection with Hive or Hadoop to perform big data analytics. The evidence suggests that R is becoming a more preeminent tool for data analytics (Muenchen, 2014).

The resulting ecosystem of big data processing tools can be described as a stylised stack of storage, MapReduce, query, and analytics application layers. Increasingly, the whole stack is provided as a cloud-based solution by providers such as Amazon (2009) and Microsoft (2011). One could argue along with Dumbill (2010) that this evolving stack has enabled and democratised big data analytics in the same way "the commodity LAMP stack of Linux, Apache, MySQL and PHP changed the landscape of web applications [and] was a critical enabler for Web 2.0" (Dumbill, 2010).

1. IBM is offering its Hadoop solution through InfoSphere BigInsights. BigInsights augments Hadoop with a variety of features, including textual analysis tools that help identify entities such as people, addresses and telephone numbers (Dumbill, 2012b).

Box 2.3. **Internet spillovers enabling data-driven innovation across the economy: The case of Hadoop (cont.)**

2. Oracle provides its Big Data Appliance as a combination of open source and proprietary solutions for enterprises' big data requirements. The appliance includes, among others, the Oracle Big Data Connectors to allow customers to use Oracle's data warehouse and analytics technologies together with Hadoop, the Oracle R Connector to allow the use of Hadoop with R, an open-source environment for statistical analysis, and the Oracle NoSQL Database, which is based on Oracle Berkeley DB, a high-performance embedded database.

3. In 2011, Microsoft began integrating Hadoop with Windows Azure, Microsoft's cloud computing platform, and one year later with Microsoft Server. It is providing Hadoop Connectors to integrate Hadoop with Microsoft's SQL Server and Parallel Data Warehouse (Microsoft, 2011).

4. In 2012, SAP announced its roadmap to integrate Hadoop with its real-time data platform SAP HANA and SAP Sybase IQ.

Until a few years ago, the nascent Hadoop space was dominated by a few products and their providers, such as the open-source Apache Hadoop distribution, the independent Hadoop distribution provider Cloudera and Amazon's Elastic Map Reduce (Harris, 2011a). But this space has rapidly become densely populated. According to Harris (2011b), the infrastructure market is already near its point of saturation:

> *The market for horizontally focused products is filling up fast with both start-ups and large vendors […] Yes, there's still room for start-ups to get in here, but the door looks to be closing fast. It's not just Hadoop, either; other techniques, from traditional data warehouses to, arguably, predictive analytics, all are nearing the saturation point in terms of vendors selling the core technologies (Hariss, 2011b).*

On the one hand, new independent Hadoop distribution actors emerged, such as Hadapt, HortonWorks (a Yahoo spin-off) and MapR. On the other hand, traditional infrastructure vendors that offer servers, storage and database technologies, moved into this space as well. IBM, EMC, Cisco, Oracle, HP and VMware have all adopted Hadoop in order to provide big data solutions to their customers – sometimes in partnership with the independent Hadoop distribution providers. They align their Hadoop products with the rest of their database and analytical offerings for business intelligence (Dumbill, 2012a).

Although Hadoop has proved to be very popular – especially for big, unstructured data challenges – classical (relational) database technologies are still important, as are next-generation massive parallel processing database technologies and their related analytical tools. Companies that provide these analytical platforms that combine databases and analytical tools are Vertica (owned by HP), Asterdata (owned by Teradata), SAP (with Hana), ParAccel, Attivo and Datastax, to name but a few. However, these products are often used in combination with Hadoop.[9]

In addition, data analytic solutions that help extract insights from data are also provided by IT providers in particular specialised analytic software companies such as SAS, The MathWorks, and RapidMiner. As highlighted in Chapter 3 of this volume, open source software (OSS) based on free software licences – such as the MIT License,[10] the BSD License,[11] the Apache License[12] and the GNU general public license (GPL v2 or v3)[13] are attracting an increasing number of business and consumer users. A well-known example is R mentioned in Box 2.3. A GPL licenced open-source environment for statistical analysis, R is increasingly used (sometimes together with Hadoop) as an

alternative to commercial packages such as SPSS (IBM) and SAS (Muenchen, 2014). The high popularity of R has even pushed traditional providers of commercial databases and enterprises servers (and competitors) such as IBM, Oracle, Microsoft, and SAP to integrate R (together with Hadoop) into their product lines, and to compete with the specialised analytic software companies.

Two trends in the business models of IT infrastructure providers can be observed. First, the business model of IT infrastructure providers is increasingly characterised by the *freemium* revenue model described in Box 2.2, where products are provided free of charge, but money is charged for additional, often proprietary, features (*premium*). This model commonly used in the open source software industry, play a major role in the data ecosystem – most likely because of the prominence of open source software solutions such as Hadoop and R, as described above. IT infrastructure vendors such as Hortonworks, Cloudera and MapR, for example, are providing at least one basic version their products for free. Revenues are then generated either based on premium versions of the product or based on value added complementary services. Hortonworks, for example, provides just one version of its Hadoop solution, called "Hortonworks Data Platform", at no cost to download. Around two-thirds of its revenues are generated based on annual subscription services contracts, which are the equivalent of maintenance and supports contracts customarily provided by virtually all businesses in the software industry (Kelly, 2013). The remaining third of the revenues is generated based on professional training services. Cloudera, as another example, provides its Hadoop solution "Cloudera's Distribution Including Apache Hadoop" (CDH) with a proprietary software component for free. The full version of the package "Cloudera Enterprise" is available for an annual for-pay subscription, however.

The second key trend that has substantially changed the business models of IT infrastructure providers is *cloud computing* (see Chapter 3 of this volume). Cloud computing has been described as "a service model for computing services based on a set of computing resources that can be accessed in a flexible, elastic, on-demand way with low management effort" (OECD, 2014c). Cloud computing can be classified into three different service models according to the resources it provides: infrastructure as a service (IaaS), platform as a service (PaaS) and software as a service (SaaS) (see Chapter 3 for further information).[14]

As cloud-based services increasingly become viable alternatives to (parts of the) infrastructure, actors such as Amazon, Google and Microsoft, and other new entrants specialised in cloud computing, are continuously challenging the predominant business models of IT infrastructure providers. Cloud computing providers are also extending their services in the market; some offer not only data storage and management solutions as IaaS, but also data analytic solutions as either PaaS or SaaS. The key business model innovation of cloud computing providers is the fact that their services are offered through a pay-as-you go model (a usage fee-based revenue model), which enables cloud users to act more responsively to their needs and their customers' demand without much initial investment in IT infrastructure. That innovation lowers the entry barriers for start-ups and small and medium-sized enterprises (SMEs), but also for governments that cannot or do not want to make heavy upfront investments in ICTs; it consequently makes the markets more competitive and more innovative (see Chapter 3).

Data (service) providers

The stacks of technologies, analytics platforms, and applications are all tailored to process data, transforming them into valuable information and insights or otherwise actionable output. But at the heart of the current data ecosystem lies data, which some have characterised as the "life blood" or the "oil" of the ecosystem (see Chapter 4). The sections that follow discuss various groups of data providers. Their business model could be described in analogy to the cloud computing value proposition as Data-as-a-Service (DaaS; see Chen et al., 2012). Their revenue models however can vary significantly.

Data brokers

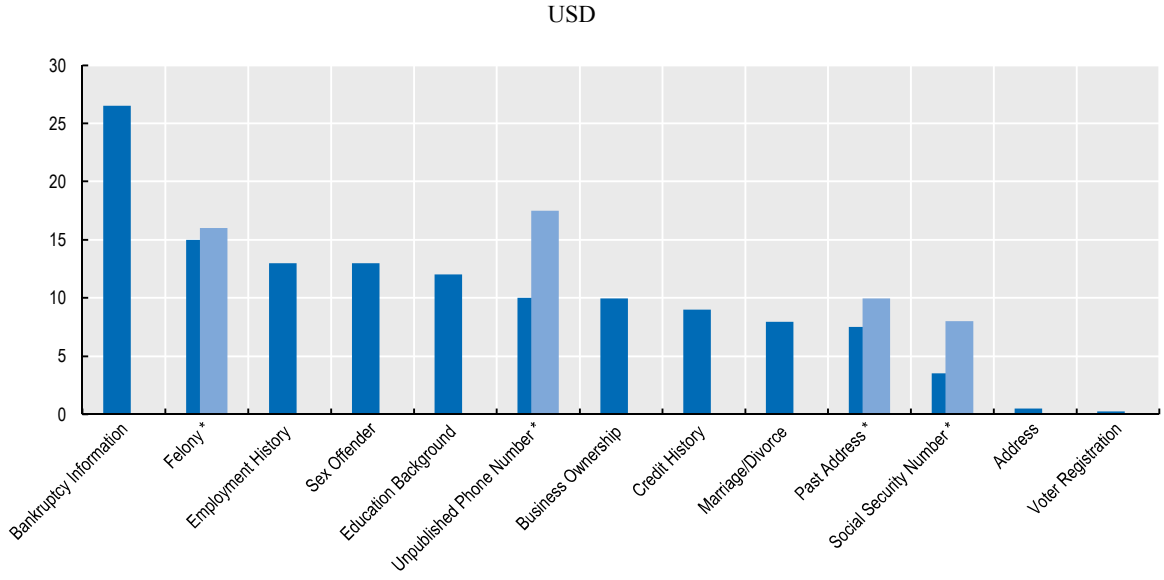
The core business objective of data brokers is to collect and aggregate data, including personal data (FTC, 2014). Data brokers such as Bloomberg, Nielsen, STATS (sports data) and World Weather Online tap into a variety of data sources that are used for data-related services. These include, for example, data that are disclosed or provided by individual firms and citizens; data from firms that install sensors; data crawled from the Internet; and data from non-profit and public sector agencies (e.g. earth observation data and demographic, health and other statistics). Some data brokers also analyse their data sets to provide information and intelligence services to their clients in wide range of domains for a variety of purposes, including verifying an individual's identity, product marketing, and fraud detection. This is where the boundary between data brokers and data-driven entrepreneurs may be blurring (see the section below entitled "Combining internal and external data sets").

Most data brokers focus on the B2B market segment. Businesses can for instance purchase the email addresses of potential customers from data brokers for marketing purposes. In addition, analytical products sold by data brokers can, for example, provide insights on the media channel to be used for advertisement (e.g., online or newspapers) and/or the geographic region to be targeted (FTC, 2014). Other data brokers provide "people search" websites through which users can search for publicly available information about potential consumers extracted from social media or other online content; this allows them to find old friends or obtain court records or other information about consumers or job applicants (FTC, 2014). But the activities of data brokers are not limited to the commercialisation of personal data. Data brokers such as Bloomberg offer, for instance, professional data services based on financial data to businesses. World Weather Online as another example sells global weather forecast and weather content to websites, businesses and the travel industry.

Analysis of the data available on B2B data brokers suggests that they primarily use the following key revenue models: (i) pay-per-use, (ii) licensing, and (iii) freemium (with a subscription-based premium) model. Data brokers that sell personal data for example for advertisement purposes mostly use the *pay-per-data-set* model. Figure 2.3 summarises some estimates that are derived from various online data brokers.[15] These estimates provide some insight into the relative market values of different pieces of personal data (OECD, 2013a). Another example of the *pay-per data set* model is World Weather Online, which provides weather data via its application programming interface (API). The total amount paid by customers is based mainly on their requests per day. Other data brokers such as Nielsen are selling their data sets (and specific analysis results) via their online stores. The pricing scheme is not completely transparent but it seems that the more specific the insights provided are, the higher the price.

For online advertising, revenues are also generated based on auctions. Data exchanges, for example, are marketplaces where advertisers bid for access to personal data about customers. Tracking the Internet activity of consumers is essential for this business model. Within seconds of visiting a website affiliated with a tracking company, detailed data on a web surfer's activity may be auctioned on a data exchange, such as that run by BlueKai (Angwin, 2010; cited in OECD, 2013a). According to their website, the BlueKai Exchange is the world's largest data marketplace, with data on more than 300 million users offering more than 30 000 data attributes; it processes more than 750 million data events and transacts over 75 million auctions for personal information a day (OECD, 2013a). The freemium revenue model is commonly used by data brokers in combinations with other revenue models. The sports data provider STATS, for example, offers a premium service based on monthly subscription to different applications in various price categories. World Weather Online is another example where a freemium model is used in addition to pay-per-data-set model.

Figure 2.3. **Market prices per record for personal data by type, 2011**



* Two different prices provided by different providers.

Sources: Locate Plus (address, unpublished phone number, felony); Pallorium (address, past address, unpublished phone number, social security number); KnowX via Swipe Toolkit (past address, marriage/divorce, bankruptcy information, business ownership); LexisNexis via Swipe Toolkit (education background, employment history, social security number, felony, sex offender); Experian (credit history); and Voters online.com (voter registration).

But data brokers are not limited to collecting and processing data for the B2B market. Some data brokers are also focusing on the business to consumer (B2C) segment, providing consumers with insights on consumer goods and services (Hartmann, et al., 2014). Examples of consumer-oriented data brokers include AVUXI and CO Everywhere; these provide data on local businesses to consumers, including data on restaurants and bars, based on a variety of data collected from the web. The revenue models of these consumer-oriented businesses are primarily advertisement and/or commission fees, which are obtained when successfully providing businesses with customers. Given the high importance of value-added services provided by many B2C data brokers, making the distinction between B2C data brokers and data-driven entrepreneurs – in particular data explorers – is often difficult.

Businesses and consumers can benefit from the services of data brokers, but at the same time are exposed to many risk factors due to the often sensitive nature of the data collected, analysed and provided. While the service of data brokers may help to prevent fraud, improve product offerings and deliver tailored advertisements to consumers, there can be significant negative side effects arising from (e.g.) misguidance of consumers, discrimination, and violation of consumer privacy (see Chapter 5 of this volume). For example, the scoring processes used in some marketing products are not transparent to consumers, rendering them incapable of preventing possible negative effects of data analytics (see FTC, 2014). There may also be a lack of transparency in the revenue schemes. Different clients may have to pay different prices, depending on the type of client (e.g. researcher, firm or government), the size of the client, the markets in which the client is active, and the purpose for which the data are expected to be used. It is equally important to acknowledge that data brokers also provide data to other data brokers, which develop new combinations of data products.

Public sector

Governments are important actors in the data ecosystem; their multiple roles can include data provision and use, investment, and provision of legal frameworks and regulation. As Chapter 10 of this volume highlights, the public sector is one of the economy's most data-intensive sectors. In the United States, for example, public sector agencies stored on average 1.3 petabytes (millions of gigabytes) of data in 2011,[16] making them the country's fifth most data-intensive sector (OECD, 2013b). The public sector is not only a key user of data and data analytics, but also a major *source* of data. However, the circumstances under which the public sector should provide value-added products from its data assets continue to be debated (see Chapter 10).

The public sector has nevertheless led the way in opening up its data to the wider economy through various "open data" initiatives (Ubaldi, 2013) and thanks to government initiatives promoting improved access to and reuse of public sector data (PSI).[17] The OECD (2008) *Council Recommendation for Enhanced Access and More Effective Use of Public Sector Information* (PSI), which is currently under review, describes a set of principles and guidelines for access to and use of PSI including public sector data (see Annex of Chapter 10).

As highlighted in the OECD (2008) PSI Recommendation, public sector data should be provided free of charge. When data are not provided free of charge, pricing should be transparent and consistent across different organisations and not exceed marginal costs of maintenance and distribution to ensure reuse and competition. Evidence shows that reduced pricing (e.g. allowing non-commercial reuse at zero cost and reducing the charges for commercial use) significantly increases the use of open government data (Capgemini Consulting, 2013), and cross-country research underlines the firm-level benefits from free or marginal cost pricing (Koski, 2011). Potential revenue models for the public sector are therefore variants of the freemium model where higher value-added data product or services are sometimes provided in addition to, and for cross subsidising, the free basic data product, if at all. EC (2013) also discusses alternative sources of revenue, including public funding, usage fees and advertisement. Chapter 10 presents an in-depth analysis of open government and PSI initiatives.

Individuals (consumers)

Even individual end users and consumers can become (active) data providers. A number of start-ups, such as Personal, are currently offering so called "data lockers" where people can gather, store and manage their personal data. These services allow people to take control of their personal data and "re-use it to their own benefit" (The Economist, 2012). Another possible development that could prove interesting is the rise of personal data marketplaces. Start-ups like Handshake and Enliken offer platforms where users can sell their personal data to interested parties (Lomas, 2013), and there are many alternative and more open initiatives of consumer participation through crowdsourcing as well. The social traffic app Waze, acquired by Google in June 2013, collects and aggregates data generated by its users to create real-time traffic information. In the Netherlands over 10 000 iPhone owners joined the collaborative research project iSPEX to measure aerosols via their mobile phones.

When analysing business models related to consumers it can be noted that consumers for the most part do not profit in direct monetary terms when providing their data to companies. As discussed before, they may instead profit from "free services" in exchange for their personal data, and they may also be confronted with advertisement. Nevertheless, there are a number of interesting developments with companies such as Handshake that are promoting a marketplace for personal data in which consumers are immediately rewarded financially when providing their personal data. The increasing demand of consumers to gain more control over their own data is also reflected by initiatives such as Diaspora* (OECD, 2012a). Initiated by four students in the United States, Diaspora* aims to highlight the significant discrepancy between the relatively low value that social networking sites provide and the privacy that users are required to give up in return (Dwyer, 2010; Suster, 2010). The project provides a decentralised platform that allows users to save their personal data on their own servers (at home or at the web-hosting provider), and thus makes it possible for the users to own and control their information.

However, anecdotal evidence suggests that the share of businesses aiming to empower individuals to play a more active role in the use of their personal data is still very low compared to the share of businesses aiming at exploiting personal data. The main reason for this can be assumed to be the relatively poor potential for profit in the case of privacy-enhancing services. This is illustrated by the case of Buyosphere, a start-up based in Canada. When the company began operations in 2010, its aim was to help individuals take control of their shopping history, while giving them the possibility of organising it, sharing it and tracking how they influence others. Buyosphere's initial business model was based on a consumer-to-business (C2B) communication flow: rather than businesses gathering personal data on users' behaviour and pushing advertising at them, Buyosphere would give consumers the power to share their preferences with the companies they choose directly. Furthermore, Buyosphere would let consumers port their own data so they could use the data for their own purposes. Tara Hunt, the CEO and co-founder, did admit however that running a C2B retail company had a significant downside by saying, "Well, we can't promote what would make us the most money" (O'Dell, 2011). And so in the course of its first year, 2010, the company redefined its business model, and now provides online product search through a combination of social search and intelligent use of data. Only once embarked on this pivotal transformation was the company able to raise additional USD 325 000 in venture capital (VC).

Analytic service providers

The field of data and IT infrastructure would seem to leave limited room for new entrants, as it is dominated by traditional vendors and a few independent Hadoop distribution providers. Yet there is indeed room for an explosion of start-ups that focus on data analytic services (including the development of software applications and visualisation tools based on data analytics).

The services of these start-ups and SMEs sit on top of the foundation layer of IT infrastructures such as database, Hadoop and analytic software solutions. Cloudera's CEO Mike Olson (Harris, 2011a), discussing the future of Cloudera and its products, noted that he sees great potential for specialised companies in this layered construction. These new companies focus on specific analytical or visualisation solutions, targeting specific industries or even specialised tasks within an industry.

There are pragmatic reasons for this, which are inherent in start-ups. Because of their focused approach, these smaller companies can offer value and ease of use that generic tools lack. As shown by Criscuolo, Nicolaou and Salter (2012), new technologies and innovations are often first commercialised through start-up companies as they are not as captured by the *innovator's dilemma* as incumbents (Christensen, 1997). They can instead leverage the advantage of starting without the legacy of an existing business and customer base to experiment and create a rich variety of presumably new business models (Hartmann et al., 2014). Expert interviews by TNO (2013) emphasise the limitations of generic off-the-shelf tools provided by incumbent IT suppliers. According to Clive Longbottom, founder of the analyst house Quocirca, many IT suppliers have a tendency to sell one-size-fits all offerings, whereas these new start-ups try to cater to very specific data needs (Heath, 2012).

As the need for business intelligence becomes more focused on real-time insights rather than historical and periodical information, the demands from the users of data analytics have changed; there is now higher demand for advanced specialised data analytic services (see Chapter 3). In addition, it is becoming increasingly important not only to generate the best actionable output, but also to present it in such a way that it is aligned with the business process that it strives to support in order to establish competitive differentiation (Dumbill, 2011). As discussed in Chapter 3, it is expected that for the next couple of years most of the value of data will be added by advanced analytical techniques, in particular predictive analytics, simulations, scenario development, and advanced data visualisations (Russom, 2011). These are the most important growth areas for the near future that data analytic service companies are now targeting. The generic analytical tools that are often provided by many IT suppliers can be important building blocks, but as the threshold for competitive, differentiating data analytics increases, data analytic applications need to be optimised for the context in which they will be used, and analytic service providers are often positioning themselves as specialised service providers to do that job.

Analysis of existing data-driven business models by Hartmann et al. (2014) suggests that two types of business models characterise analytic service providers, which they refer to as "analytics-as-a-service" and "aggregation-as-a-service". Data analytic service companies provide advanced data aggregation and/or analysis services to their customers, which are primarily businesses. But the main characteristic of data analytic service providers that distinguishes them from (e.g.) data brokers is that their activities are primarily based on data provided by their customers rather than obtained from crawling the web or collected from third parties including other data brokers. Where external data sources are collected and integrated by data analytic service companies, this is done mainly to enhance the results

of prior analysis of their customer's data. Furthermore, data analytic service providers will typically act as subcontractors to the data controller (i.e. data processor), while data brokers typically act as independent data controllers (in the B2B and B2C market).

The revenue model of data analytic service providers is therefore often based on service contracts. But increasingly, Internet start-ups are providing their services – including the analytic results – via APIs and visualisation platforms (Hartmann et al., 2014). These start-up companies can (and therefore do) use alternative revenue models, including in particular subscription and usage based revenue models. For example, the start-up company Welovroi, based in Madrid, Spain, is a marketing company that provides monitoring and analysis tools for data provided by customers via the Internet. Welovroi offers its services on a monthly subscription basis, the amount of which depends on the number of employees of its customers, and the number of web services that the customers use.

Another interesting development in data analytic services is the crowdsourcing of data analysts. These services enable organisations that include businesses and governments, as well as individuals all over the world to post their data and let others compete to produce the best analytic results. Crowdsourcing of data analytic activities can lead to faster results, on unprecedented scales, and with better quality control than any individual or small research group can attain. Given its open, informal structure, crowdsourcing is cross-disciplinary by design. In some cases, even gifted amateurs and people without direct experience with the problem provide valuable insights and solutions.

InnoCentive, one of the first companies to crowd-source in the chemical and biological sciences, today has more than 300 000 registered "solvers"', who stand to gain rewards of between USD 5 000 and USD 1 million if their solution works. Key to the success of InnoCentive's crowd-sourcing has been: i) a carefully defined governance structure designed to protect intellectual property from both the seeker and the solver; ii) reduced barriers to participation, so that the challenge scales quickly; and iii) global reach, increasing the likelihood of solutions coming from very unexpected directions. Another popular example of a start-up providing crowdsourced analytic services is Kaggle, which in November 2011 raised USD 11 million from a number of investors (Rao, 2011).[18] Hal Varian, Google's Chief Economist, described Kaggle as "a way to organise the brainpower of the world's most talented data scientists and make it accessible to organisations of every size" (Rao, 2011). According to Kaggle, more than 200 000 data scientists have registered worldwide, from fields such as computer science, statistics, economics and mathematics. These data scientists are competing for prizes as high as the USD 3 million *Heritage Health Prize* for the most accurate prediction of the patients who are most likely to require a hospital visit within the next year (The Economist, 2011).

While firms such as InnoCentive and Kaggle aim at data analysts that have advanced skills in data analytics, other crowdsourcing platforms are designed in such a way that the data analytic problem is masked and presented to Internet users in a very simplified way, often it takes the form of a game, which when won leads to the solution of the original data analytic problem. Foldit, for example, is a popular online citizen-science initiative, in which individuals are scored on the structure of proteins that they have "folded". The game records the structure and the moves that the players make, and scientists can capture the data that are then used to improve the problem-solving process in every aspect, from the quality of the scientific results to how long people play the introductory levels meant to teach the game.[19]

Another example is Zooniverse, which enables researchers to design crowdsourcing platforms that take their data and present them in a format that will let the crowd help

them to achieve their objectives. Zooniverse has a community of over 850 000 people, who have taken part in more than 20 citizen science projects over the years. These initiatives support a form of "scientific democracy", where data can be shared among and utilised by investigators in public and private sectors, policy makers, and the public. Crowdsourcing platforms for research and health research in particular are discussed in more detail in Chapter 7 and 8.

Data-driven entrepreneurs

Data-driven entrepreneurs are using data analytics to various ends, ranging from cost saving through financial monitoring to revenue growth through new marketing strategies and product development. As a study by Brynjolfsson and McAffee (2012) points out, these goals strongly depend on the maturity of an organisation in terms of its ability to deploy data analytics and related technologies. As companies gain more advanced data analytics experience, the balance between cost saving and revenue growth will shift. Deploying data and analytics for marketing and sales becomes more important, as does, to a lesser extent, product research and strategy development purposes. More disruptive innovations that upend current business practices, or create new ones, require more experience, greater commitment, and a more solid belief in the potential of leveraging data. Still, the use of data and analytics for incremental changes can be a helpful precursor for more radical disruptive DDI, in which (networks of) organisations rethink products, business models or even whole value chains (Lavalle, 2010).

Although some companies have indeed shifted their data-related priorities from cost efficiency to revenue growth to innovating for competitive differentiation, this has not yet resulted in a grand-scale proliferation of more disruptive DDI since most organisations deploy data analytics to enhance their existing business models. However, examples of such kinds of disruptive innovation are increasing in number; they are realised by start-up companies as well as traditional (non-ICT) companies. These companies base their innovative business models on the deployment of applications that use data generated through the Internet including the Internet of Things (IoT – see Chapter 3). They thus build their products (goods and services) on top of existing data, using that data as an input to provide their innovative goods and services. The US-based start-up BrightScope, for example, extracts public data from the Department of Labor and processes it to bring transparency to opaque markets. Through the use of cloud-based software, the company aims to drive better decision-making in the areas of retirement plans and wealth management.

Two interesting examples of traditional (non-ICT) companies are Nike and the Dutch IJkdijk, which have redesigned some of their traditional products as "data products". Nike introduced the online Nike+ platform, the Nike+ sensor that can be clipped on running shoes, an app that tracks runs and more recently the FuelBand, a wristband that tracks activities and calories burned during the day. Although its core value proposition – supporting people to be physically active and healthy – has not changed, Nike is now more and more providing this proposition by using data that enables users to set their goals, track their progress and include social elements. It has also created an API that allows third parties to develop apps based on this data-driven platform. The IJkdijk is the result of a research program in which a dike in the north of the Netherlands was equipped with sensors. The collected data are analysed and visualised to improve dike monitoring and water management. Both examples illustrate the potential of sensor data and M2M for DDI. Another example is autonomous self-driving cars discussed in Chapter 3. The development of autonomous and smart cars is in line with a bigger transition towards

smart cities in which organisations are deploying data and data analytics to realise innovations in a complex and dynamic environment (see Chapter 9).

Building on their own experience and expertise and their accumulated assets including data and analytics, many data-driven entrepreneurs may become data and analytic service providers for others as well. In this case they are not solely consumers of data and analytic products; they also contribute with their data and software development activities for the benefit of other organisations that can reuse the data and the data analytic solutions for very different purposes. The example of Amazon as a big data powerhouse was already given above (Assay, 2013). Walmart, as another example, is developing its own data analytic services via its subsidiary Walmart Labs,[20] which is also actively contributing to the co-development of open source analytics. Another example is John Deere, which is transforming itself from a manufacturer of tractors to a highly advanced business intelligence service provider for farmers. Finally, there are businesses that open up their data; an example is the Dutch energy network service provider Alliander, which recently organised a workshop with partners and stakeholders to explore the potential of open data.

These examples illustrate how even organisations for which data and analytics originally were not part of their primary business model can become actors for different steps of the value creation process in the data ecosystem. This phenomenon has been described by Rao (2013), who wrote an article about non-tech corporations "eating" tech-start-ups as they try to position themselves since datafication is affecting their market:

> *It's no longer Google, Facebook and Yahoo that are competing to acquire the best and the brightest start-ups in Silicon Valley. There are plenty of corporations in retail, health, agriculture, financial services and other industries that are sending their corp-dev talent to scout out possible acquisitions in the Bay Area and beyond (Rao, 2013).*

Based on Hartmann et al. (2014), two major types of data-driven organisations can be distinguished: i) those that provide goods and services based on the collection of data available on the web and via data brokers (i.e. data explorers), and ii) those that provide goods and services to generate data that are used to enhance user experience and to empower additional services (i.e. data-generating platforms).

Data explorers

Data-driven entrepreneurs that act as data explorers are closely related to data brokers in the sense that they collect available data either by crawling the web, tapping into social media sites, or even purchasing data from brokers. However, in contrast to data brokers – that have as a primary business objective the provision of data and/or of value added insights – data explorers have a well-defined business objective that addresses particular business or consumer needs (other than the need for data or insights). And unlike the data-generating platforms discussed below, data explorers do not deploy the means to generate data themselves. An example of a data explorer is Gild, which helps companies recruit software developers by automatically evaluating the software source code these developers have published on open source software sites such as GitHub and Google Code, and their contributions to popular Internet forums on software development such as Stack Overflow. An expertise score is computed to rank a developer's ability to code, while another score, the demand score, assesses how competitive it will be to recruit the candidate (Gild, 2014).

The revenue model of data explorers depends on whether they are targeting the B2B or B2C market. In the case of B2B, their revenue model is similar to that of online analytic service providers: B2B data explorers tend to rely primarily on freemium (with subscription based premium) revenue models. In contrast B2C data explorer tend to rely more on revenue models-based on advertisement and commission fees, but sometime also in combination with freemium and subscription based revenue models (Hartmann, et al., 2014). For example, DealAngel, founded in 2010 in Moscow, Russia, provides consumers with a list of hotels with the best deals free of charge. Its revenues are generated based on commission fees from the booking websites that consumers are directed to when actually booking a hotel (Ha, 2012; Hartmann et al., 2014).

Data generating platforms

Data generating platforms include a wide range of companies ranging from small, low-tech SMEs to highly data-intensive companies such as Apple and Google, including traditional (non-ICT) companies such as Nike and TomTom. They typically include data-driven service providers (i.e. service platforms) from which data are generated as a by-product of their actual business activity to support the sales of goods and services: this contrasts with data explorers or data brokers, for which the reuse of existing data is at the core of their business models. The service platform providers also include businesses that sell mobile applications (apps) or sensor equipped smart devices that are interconnected via machine-to-machine communication (M2M) in the IoT (see Chapter 3). Companies such as Monsanto, John Deere and DuPont Pioneer are, for example, taking advantage of the "Industrial Internet" by integrating sensors with their latest equipment "to help farmers manage their fleet and to decrease downtime of their tractors as well as save on fuel" (Big Data Startups, 2013). The same sensor data are then linked with historical and real-time data on e.g. weather prediction, soil conditions, fertiliser usage and crop features to optimise and predict agricultural production. In the case of John Deere, some of the data and analysis results are presented to farmers via the MyJohnDeere.com platform (and its related apps) to empower farmers to optimise the selection of crops, and of where and when to plant and plough the crops (Big Data Startups, 2013).

The fact that service platform providers produce data as a by-product does not prevent them selling their data to third parties. For instance, service platforms may share their data with business partners, or may provide a platform (online) that allows the exchange of several information services to clients in a range of domains. Data collected on agriculture platforms such as provided by Monsanto, John Deere and DuPont Pioneer, for example, are being considered as an important data source for biotech companies to optimise genetically modified crops (GMC). Reuse of the data is also being considered by crop insurance companies and traders on commodity markets, which has led to controversial discussions on the potential harm to farmers from discrimination and financial exploitation (Bunge, 2014; *The Economist*, 2014).

The main characteristic of service platforms is that they benefit from data enabling multi-sided markets, where activities on one side of the market go hand in hand with the collection of data, which is exploited and used on the other side of the market (see Chapter 4 of this volume). These markets are also taking advantage of network effects emerging on at least one side. For data explorers and data brokers, in contrast, the characteristics of multi-sided markets are less applicable, but economies of scale, in particular due to network effects, are more relevant. The revenue model of data generating platforms therefore relies heavily on the combination of network effects that typically affect all sides of the market of the service platform provider. As the utility for

users on all sides of the market increases with the increase in their numbers, users are more willing to pay for access to a bigger network and/or to contribute with their own data. Combined with the increasing returns to scale and scope the data enable, these network effects can lead to huge profit margins for platform providers (see Chapter 4).

Cross subsidies are therefore often used by service platform providers: a service is offered for free or at a low price on one side of the market (often the B2C market), but subsidised with revenues generated on the other sides of the market (often the B2B market) (Bonina, 2013). For instance, service platform providers often use the *freemium model* on one or more sides of their market, where the cost of the free service is subsidised by premium customers across all sides of their markets. Online dating portals, for example, operate with a freemium and premium subscription based model on both sides of their market. But often more complex revenue models are used. For example, the freemium revenue model can be used on one side of the market (e.g. the consumer market), sometimes in combination with an advertisement revenue model. In the case where a physical device is required (e.g. navigation system hardware such as provided by TomTom), an asset sale model may be used instead or in addition (Hartmann et al., 2014). For the other side of the market, service platform providers can use the same models as described above for data brokers, namely freemium (with subscription-based premium) model or service contracts. Platforms such as Facebook are financed by (e.g.) advertisers on the side of their market that uses data provided by individuals on the other side of the market. Advertisers can thereby better target potential consumers to increase sales, and individuals have access to social network services that are provided to them free of charge in exchange for the free use of their personal information by Facebook.

2.2. Interactions in the data ecosystem

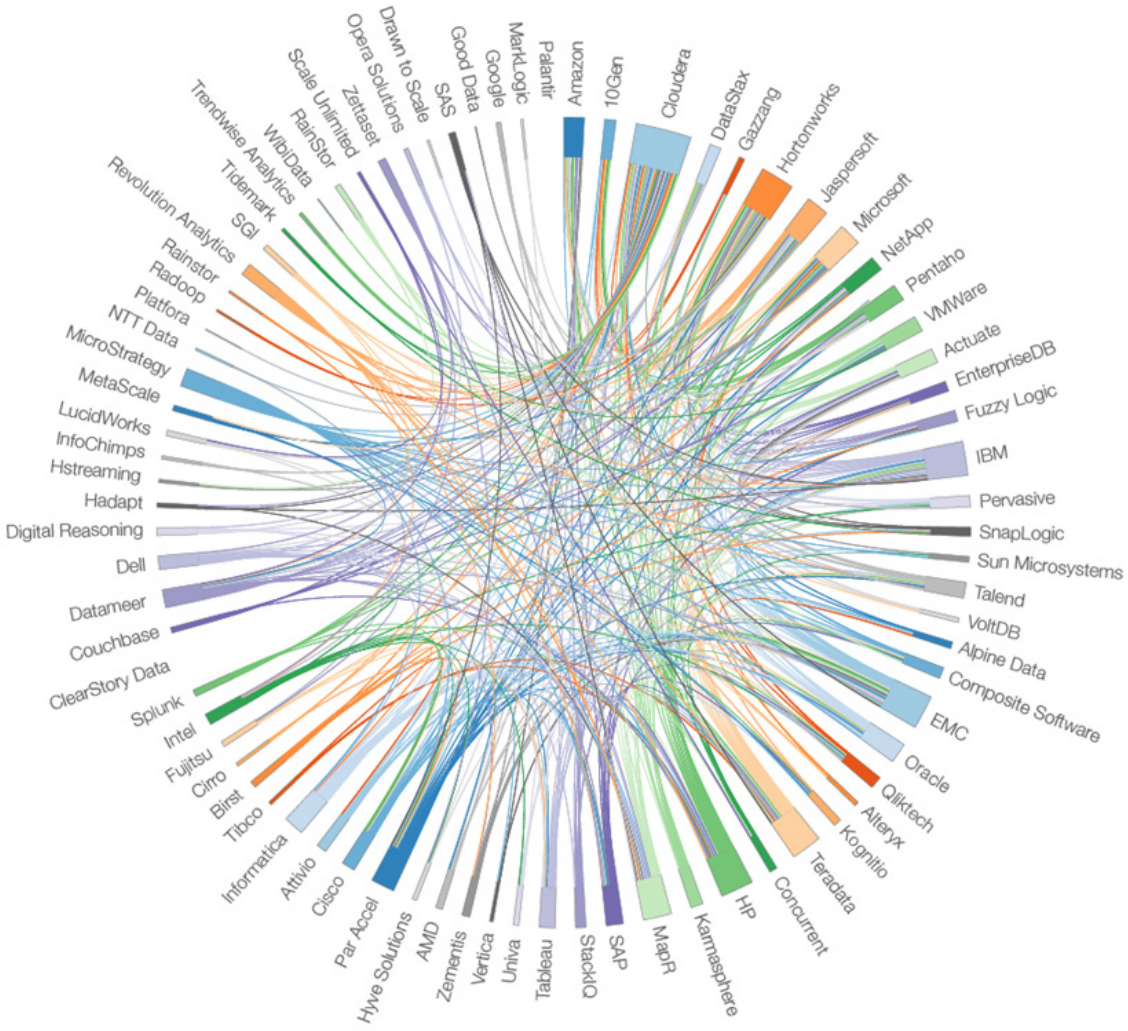
This section discusses the interaction among actors that structure the data ecosystem. It analyses in particular the relation between competition and collaboration for DDI, and how this translates into horizontal and vertical movements by the various actors.

Co-opetition: Competition and collaboration

Competition, but also collaboration, or "co-opetition" is key to leveraging the potential of the multidisciplinary field of DDI (see Woo, 2013). The multidisciplinary characteristics of "big data" challenges and opportunities are not confined to the complex stack of infrastructural elements, analytical techniques and visualisation tools. They also include organisational and HR-expertise and extensive domain knowledge in the specific areas where data are applied. In that respect, the functioning of the data ecosystem fits the general description of innovation ecosystems, in which collaboration among individual companies allows them to create value that no single company can deliver on its own (Adner, 2006).

It is difficult to properly assess what companies are currently dominating the data ecosystem. Exact numbers of market share are hard to come by and would be difficult to interpret as the ecosystem comprises many different kinds of intertwined services. However, if collaboration is a necessity in the data ecosystem, the number of partnerships could serve as a potential indicator of market activities. Figure 2.4 provides an overview of more than 50 companies with the highest number of partnerships with other organisations in the Hadoop ecosystem (O'Brien, 2013).

Figure 2.4. **Partnerships in the Hadoop ecosystem, January 2013**



Note: The larger the source bar, the greater will be the number of a company's partnerships. For example, Cloudera has by far the highest number of partnerships, followed by Hortonworks, IBM, and EMC.

Sources: O'Brien, 2013, based on Datameer, 2013.

The Hadoop ecosystem includes many actors from the different layers presented in Figure 2.2. The IT infrastructure providers Cloudera, Hortonworks and MapR, which act as independent Hadoop distribution providers, are especially well connected, but so are more traditional IT vendors such as IBM, EMC, HP, Microsoft, Oracle, SAP, VMware, Cisco and Intel. Data-driven entrepreneur Amazon and IT infrastructure provider Dell – which did not make the list in 2012 – have both improved their networks considerably in 2013 and are now among the top contributors to the Hadoop ecosystem. As noted before, many actors within the global data ecosystem are active in different layers, or steps in the value creation process. This includes in particular the biggest actors such as Microsoft, Google, Amazon, Oracle, SAP, SAS and VMware.

Looking at the biggest actors in more detail including their economic performance (Tables 2.1-2.3), it is worth highlighting a number of findings:

1. The large providers participating in the Hadoop ecosystem, are mainly companies registered in the United States with the exception of Yahoo Japan, NTT Data, and Fujitsu (Japan), SAP (Germany), Persistent System (India) and Acer (Chinese Taipei). That said, it should be also noted that the number of top providers has been reduced due to merger and acquisitions (M&A).
2. It comes as no surprise that most of the top firms participating as providers in the data ecosystem are Internet and software firm. However some hardware firms and in particular IT equipment firms are heavily involved as well. Semiconductor firm Intel and AMD (Advanced Micro Devises) are the exceptions. In terms of M&A these companies have been remarkably active, which explains their increasing involvement in the data ecosystem.

Table 2.1. **Performance of the top Internet firms involved in the Hadoop ecosystem, 2013**

USD million (except employment numbers)

Internet firms	Country of registration	Revenue 2013	Employment 2013	R&D 2013	Income 2013
Amazon.com Inc	United States	74 452	117 300	6 565	274
Google Inc	United States	56 168	53 861	5 467	14 655
Facebook	United States	7 872	6 818	1 415	1 500
Yahoo! Inc	United States	4 987	11 700	886	3 946
Netflix Inc	United States	4 375	2 045	379	112
Yahoo Japan Corp	Japan	3 795	5 780		1 229
Concurrent Computer Corp	United States	63	229		4

Source: OECD Information Technology database, compiled from annual reports, SEC filings and market financials, July 2014.

Table 2.2. **Performance of the top ICT service and software firms involved in the Hadoop ecosystem, 2013**

USD million (except employment numbers)

ICT service and software firms	Country of registration	Revenue 2013	Employment 2013	R&D 2013	Income 2013
International Business Machines Corp	United States	99 751	434 246	6 226	16 483
Microsoft Corp	United States	77 849	99 000	10 411	21 863
Oracle Corp	United States	37 920	120 000	5 149	10 806
SAP AG	Germany	22 858	66 061	3 102	4 521
Computer Sciences Corp	United States	13 544	87 000		1 501
SYNNEX Corp	United States	10 845	12 500		152
VMware Inc	United States	5 207	14 300	1 082	1 014
Teradata Corp	United States	2 743	10 200	179	395
Informatica Corporation	United States	948	3 234	166	86
Microstrategy	United States	576	3 221	98	83
Splunk Inc	United States	303	1 000	76	- 79
Persistent System	India	284	6 970		42
Tableau Software Inc	United States	232	1 360	61	7
Pervasive Software Inc	United States	49	255		2
NTT Data Intramart Corp	Japan	42	257		1

Source: OECD Information Technology database, compiled from annual reports, SEC filings and market financials, July 2014.

Table 2.3. **Performance of the top ICT hardware firms involved in the Hadoop ecosystem, 2013**

USD million (except employment numbers)

ICT hardware firms	Country of registration	Revenue 2013	Employment 2013	R&D 2013	Income 2013
Hewlett-Packard	United States	112 298	317 500	3 135	5 113
Dell Inc	United States	56 940	108 800	1 072	2 372
Intel Corp	United States	52 708	107 200	10 611	9 620
Cisco Systems Inc	United States	48 607	66 639	5 942	9 983
Fujitsu Ltd	Japan	43 046	168 733		478
EMC Corp	United States	22 787	60 000	2 689	2 557
Acer Incorporated	Chinese Taipei	11 967	7 967	103	- 90
NetApp Inc	United States	6 368	13 060	922	769
Advanced Micro Devices Inc	United States	5 299	10 340	1 201	- 83
Silicon Graphics International Corp	United States	767	1 400	61	- 3

Source: OECD Information Technology database, compiled from annual reports, SEC filings and market financials, July 2014.

Mergers and acquisitions, and vertical integration

As the data ecosystem evolves, many new companies emerge. Subsequently, larger companies try to strengthen their position. Not only will they develop new products and forge partnerships, but they will also acquire promising start-ups to improve and augment their propositions with analytics platforms, visualisations and applications (ESG, 2012). Infochimps CEO Nick Ducoff provides an explanation for this dynamic between the specialised nature of many big data start-ups and the more generic platforms they build on (Watters, 2011b):

> *If you are best at the presentation layer, you don't want to spend your time futzing around with databases […]. What we're seeing is start-ups focusing on pieces of the stack. Over time the big cloud providers will buy these companies to integrate into their stacks. (Watters, 2011b)*

There is a tendency of consolidation in the IT service industry that could also especially affect IT infrastructure providers in the data ecosystem. At the European Data Forum 2013, Siemens manager in charge of the big data initiatives, Gerhard Kress, emphasised the importance of research into vertically integrated algorithms. In an analysis of the big data market ESG,2012, an IT market research and advisory firm noted how data service companies try to obtain dominant positions in certain vertical industries: "[…] where whomever has 'the most data scientists with a vertical bent' may win".

According to a report from Orrick (2012) on emerging big data companies, based on deals and investments mainly in the United States, big data financing activity has increased significantly since 2008 (see Figure 1.4 in Chapter 1 of this volume). Recent years have also seen the take-off of the first IPOs of big data companies. The number of mergers and acquisitions (M&A) has increased rapidly from 55 deals in 2008 to almost 164 deals in 2012, with almost USD 5 billion being invested over that period. In the first half of 2013 alone, big data companies raised already almost USD 1.25 billion across 127 deals. IBM was the most active acquirer of big data companies in 2012, followed by Oracle.

The evolution in value creation with data seems to be reflected in the above described trends on M&A. In the past five years, in terms of both deals and (especially) investments, the focus has shifted from big data infrastructure to big data analytics and applications. Whereas in 2008 infrastructure accounted for 46% of big data investments, this share decreased to 31% in 2012. These numbers also illustrate that the analytics, visualisation and application layer, the "last mile of big data" is where most of the value of data is generated and where true differentiating quality resides as the commoditisation of data analytics continues (ESG, 2012).

Combining internal and external data sets – the emergence of data markets

Most organisations initially apply analytics to their own internal data sets, possibly combining several databases from various departments and processes. But the value of data analytics also lies in the combination of both internal and external data (Redman, 2008). As highlighted in Chapter 4, the value of data is highly context-dependent and "multiplies" when it can be shared and linked with other data sets. As the data are put in a larger context they can reveal additional insights that otherwise would not be possible to glean.[21] A white paper of the European Technology Platform NESSI (2012) stresses how important it is to integrate private data with external data to enhance existing products and services. As O'Reilly's Ed Dumbill (2012a) notes:

> *Mixing external data, such as geographical or social, with your own, can generate revealing insights. […] Your own data can become that much more potent when mixed with other datasets.*

Pointing out that "critical information often resides outside companies", Biesdorf, Court and Willmott (2013) from McKinsey & Company highlight what integrating external data sources involves:

> *Making this information a useful and long-lived asset will often require a large investment in new data capabilities. Plans may highlight a need for the massive reorganisation of data architectures over time: sifting through tangled repositories (separating transactions from analytical reports), creating unambiguous golden-source data, and implementing data-governance standards that systematically maintain accuracy. (Biesdorf, Court and Willmott, 2013)*

In addition to using data from external sources to create value, it could also be valuable to open up proprietary data sets to others. As Rufus Pollock stated at the OECD Technology Foresight Forum in October 2012:[22] "The best thing to do with your data will be thought of by someone else", referring to the open data movement. Chapter 4 highlights a number of reasons why open data can be an optimal strategy from a private and public sector perspective. Some organisations offer their data for free via their website or specific online portals – especially NGOs and governments as highlighted in Chapter 10. Other organisations sell their data. The example of French mobile ISP Orange with its Floating Mobile Data (FMD) technology was already given above. Other well-known examples are Internet firms such Facebook and Google, whose vast collections of personal data are a valuable resource for advertisers. In some cases social media companies work with third parties such as analytic service providers that commercially exploit the social data; examples are Gnip and Datasift. These companies have access to the so-called Twitter Firehose and other social media data, which they prepare and manage to make more accessible and useful to their customers by adding all kinds of filters that fit users' specific needs.

In addition to the data sources and intermediaries mentioned above, including in particular data brokers, data are also exchanged through online services (i.e. data markets) that host data from various publishers and offer the (possibly enhanced) data to interested parties (Dumbill, 2012b). The most established data markets are provided by Infochimp, Datamarket, Factual and Microsoft's Azure, although there are several more (Big Data Startups, 2014). Some data markets try to offer all the data they can, such as Infochimp. Others focus on specific kinds of data, such as Factual, which originally started with location data and is now branching out to a few new specific verticals. Another type of specialisation is to choose a specific target group, such as Figshare, a data market for researchers. The boundaries between data brokers and data market providers are blurred, in particular because both provide the following useful value-added service according to Dumbill (2012b):

1. they provide a point of discoverability and comparison for data, along with indicators of quality and scope
2. they handle the cleaning and formatting of the data, so they are ready for use
3. they provide an economic model for broad access to data that would otherwise prove difficult to either publish or consume.

However, it is interesting to note that despite the growth of data intermediaries, as yet there is no established data marketplace where organisations and individuals can sell or exchange data directly with each other. Some platforms provide some of these functionalities, but they are tailored to specific, tightly integrated value chains that are heavily dependent on each other, for example in mobility, logistics or agriculture (e.g. the "smart dairy project" from TNO and several Dutch companies in the field of dairy farming) (TNO, 2013).

These three propositions illustrate how data marketplaces and data brokers can facilitate finding the right kind of data and fulfil even some additional steps in the value creation process, such as data preparation, to ease further data integration. However, one important distinguishing factor between data brokers and data market providers is that data brokers are actively engaged in the collection of additional data, while data market providers are intermediaries through which data controllers (including data brokers) can offer their data sets. Furthermore, some marketplaces allow their customers to explore data and to mix them together with their own or other available data sets to create new value. Although most marketplaces are focused on developers as their main users, Dumbill (2012b) notes that some data marketplaces try to target less IT-savvy users as well. Microsoft's Azure, for instance, has aligned its data sets not only with its other big data products, but also with its business tools such as Excel. This makes it easier for smaller organisations (and even individual users) to download and combine different (internal and external) data sets. Furthermore, data marketplaces enable a new economic model for data use and sharing, which enhances the overall value of the data provided. As Factual's CEO Gil Elbaz explained at the Strata 2011 conference:

> *Another dimension that is relevant to Factual's current model: data as a currency. Some of our most interesting partnerships are based on an open exchange of information. Partners access our data and also contribute back streams of edits and other bulk data into our ecosystem. (Watters, 2011a)*

The data ecosystem and its global value chains

The data ecosystem involves global value chains (GVCs), formed by companies increasingly dividing up their production processes and locating productive activities in many countries. As highlighted above, the data ecosystem relies on technologies and resources that are distributed around the globe. The ICT infrastructures used to the perform data analytics including the data centres and the software will rarely be located within just one national boarder. They will instead be distributed around the globe to take advantage of factors including local work load, the environment (e.g. temperature and sun light) and labour costs.[23] Data can thus be collected from consumers or devices located in one country through devices and apps developed in another country. They can then be processed in a third country and used to improve marketing to the consumer in the first country and/or to other consumers around the globe.

Furthermore, as highlighted above, many data-driven services stand on the shoulders of giants who have made their innovative services (including their data) available via APIs – many of which are, as noted above, located in foreign countries. One example, which has become better known in developing economies, is Ushahidi, a non-profit software company based in Nairobi, Kenya. Ushahidi develops free and open source software for data collection, visualisation, and interactive mapping based on available APIs provided by Internet firms such as Google and Twitter. One of its first products was created in the aftermath of Kenya's disputed 2007 presidential election to collect eyewitness reports of violence via email and text messages to be visualised on Google Maps. Since then, Ushahidi's data-driven services have been used in particular during crises around the world – for example, in aftermath of the 2010 earthquake in Haiti and the 2010 earthquake in Chile, respectively, where it was used to locate the wounded.

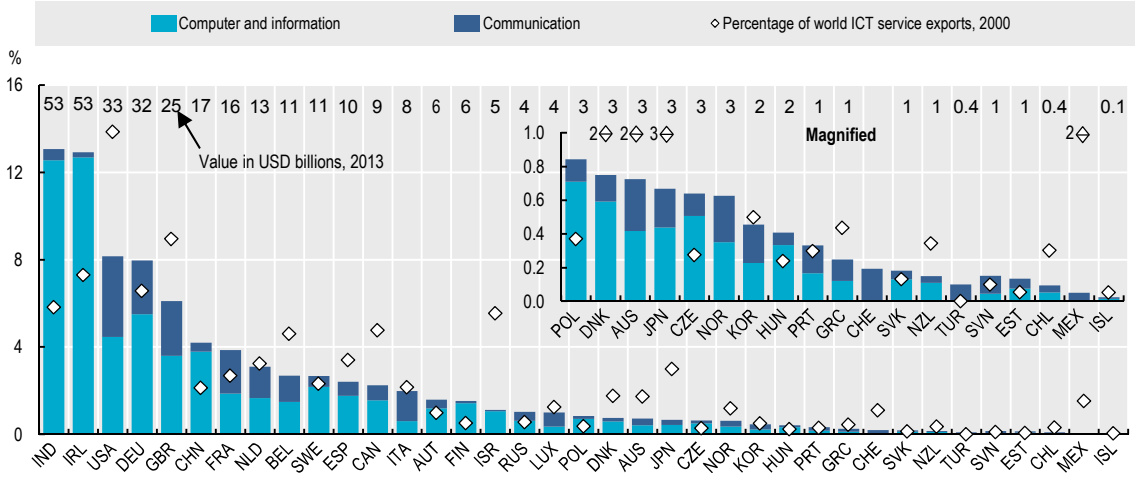
Figures on the distribution of data-driven services are not known. However, the distribution of Internet sites hosted by country code top-level domains (ccTLDs) as identified in the Alexa one million (a list of the top 1 million sites of the world) can provide an approximating picture of how data-driven services are distributed or actually concentrated in the world (OECD, 2014c).[24] The United States alone accounted for almost 60% of all top sites hosted in the OECD area in 2013, or more than 50% of all top sites hosted in the OECD area plus Brazil, People's Republic of China (hereafter 'China'), Colombia, Egypt, India, Indonesia, Russia, and South Africa taken together (see Figure 3.5 in Chapter 3 of this volume). Looking at all sites around the world and grouping the European and Asian countries into regions may give a better perspective. In 2013, the United States accounted for 42% of all top sites hosted, while Europe hosted 31% of the world's top sites and Asia 11% (Pingdom, 2012; 2013). The high concentration of top sites hosted in the United States, which is also reflected in the high number of co-location data centres there (see Figure **1.5** in Chapter 1), are most likely related to a backhaul market that function well in the United States (see Chapter 3). It also supports findings of the US digital data industry's relatively higher share of value-added highlighted in Box 2.1.

As the data ecosystem involves GVCs, many of its activities are captured in international trade. These include not only the trade in ICT services provided by actors in the IT infrastructure layer, but also trade in data-intensive services. As highlighted in Kommerskollegium (2014), even trade involving goods and services that are not data-intensive also typically involve data such as:

- corporate data (to coordinate among different parts of a company and to sell goods and services)
- end-customer data (B2C) (to sell goods and services, enable outsourcing, and provide [24/7] support, and for developing new products)
- human resources data (to co-ordinate among different parts of a company and to match skills, but also to enable outsourcing)
- merchant data (B2B) (to sell goods and services and provide [24/7] support, and for developing new products)
- technical data (to sell goods and services, upgrade software, monitor the operation of products, enable outsourcing and provide [24/7] support, and for developing new products).

There are no figures on data- and analytic-specific services. But taking as a proxy trends in trade in ICT related services, which obviously involve the exchange of data, one can assign a significant growth in cross-border (trade-related) data to the major exporters of ICT services between 2000 and 2012 (Figure 2.5). The largest exporters of ICT services in 2013 were India, Ireland, United States, Germany, the United Kingdom and China. These countries are estimated to be the largest destination of cross-border data. As a consequence, the leading OECD importers of ICT-related services are also the major sources of trade-related data, including in particular the United States and Germany.

Figure 2.5. **OECD and major exporters of ICT services, 2000 and 2013**



Source: OECD (2014d), Measuring the Digital Economy: A New Perspective, OECD Publishing, Paris, based on UNCTAD, UNCTADstat, June 2013, http://dx.doi.org/10.1787/888933148882.

2.3. Key challenges in the global data ecosystem

The globally distributed nature of the data ecosystem, its hyper interconnectedness, and the interdependencies of its actors and their technologies and resources raise a number of policy issues that are specific to the global data ecosystem. These challenges include: i) the difficulty of value attribution which challenges measurement but also taxation policies, ii) the exploitation of key points of control and the competition

implications, iii) the potential barriers to the free flow of data and the importance of the open Internet, and iv) interoperability and standard issues.

Attribution of value, and taxation[25]

The global distribution and interconnectedness of the data ecosystem makes it challenging to attribute the share of the overall value created to specific actors. This has implications for measurement (see Box 2.1), but also raises policy challenges related to taxation. In particular, some governments have expressed concerns that some of the characteristics of the global data ecosystem could create opportunities for *Base Erosion and Profit Shifting* (BEPS) through "aggressive tax planning by multinational enterprises making use of gaps in the interaction of different tax systems to artificially reduce taxable income or shift profits to low-tax jurisdictions in which little or no economic activity is performed" (OECD, 2014e). OECD work on *Addressing the Tax Challenges of the Digital Economy* (2014e) highlights a number of tax issues that the digital economy raises. Many of the issues discussed, however, are not necessarily specific to the global data ecosystem, such as business practices that take advantage of the cross border nature of the Internet to eliminate or reduce tax in a country or that exploit opportunities for BEPS with respect to VAT through e.g. the use of remote digital supplies to exempt businesses (OECD, 2014e).

This section briefly highlights potential BEPS issues discussed in OECD (2014e) that *are* specific to the data ecosystem. Many of these issues emerge due to the global distribution and interconnectedness of the data ecosystem in combination with the economic properties of data discussed in Chapter 4 of this volume. That combination raises a number of questions:

- whether data is being appropriately characterised and valued in corporate balance sheets for tax purposes
- whether any profits attributable to the remote gathering of data by an enterprise should be taxable in the State from which the data is gathered
- and whether current nexus rules continue to be appropriate.

At the core of the issues raised by these questions stands the challenge of attributing the value created in the data ecosystem to specific actors. Attribution is key for the current paradigm used by tax authorities to determine where tax-relevant economic activities are carried out and where value is created. The data ecosystem may challenge this paradigm – and with that, the foundation for taxation in most countries.

Measuring the monetary value of data

The value attribution challenge is most of all related to the challenge of measuring the monetary value of data (see Chapter 4). Most businesses still do not fully take into account the economic value of the data they control in their balance sheet, "although data purchased from another related or unrelated business would be treated as an asset in the hands of the buyer" (OECD, 2014e). As highlighted in Chapter 4, data can under some circumstances be considered a capital good (subject to depreciation). However, in many cases the context dependency of data challenges the applicability of market-based value attribution, since this assumes that markets can converge towards a price at which demand and offer meet. That is not always the case. As "Exploring the Economics of Personal Data: A Survey of Methodologies for Measuring Monetary Value" (OECD, 2013a) showed, the monetary valuation of the same data set can diverge significantly

among market participants.[26] Furthermore, where data are collected or generated and no market exists to set a price, businesses may have no means to objectively evaluate their data assets. In that particular context questions have emerged as to whether services provided in exchange for (personal) data can be considered free goods or barter transactions, and how they should be treated for accounting and tax purposes (OECD, 2014e).

Data ownership

Another factor making the attribution of value creation difficult is the challenge related to "data ownership", a concept has turned out to be impractical in many cases (see Chapter 4). In contrast to other intangible assets, data typically involve complex assignment of different rights across different data stakeholders, requiring "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 and no clear ownership can be assigned. Different stakeholders will typically have different degrees of rights depending on their role. In cases where the data are considered "personal" the situation is more complex, as privacy regimes typically tend to strengthen control rights of the individuals (see for example the Individual Participation Principle of the OECD [2013c] *Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data*).[27]

Global distribution and interconnectedness

The distribution and interconnectedness of data-driven services only increase the difficulty of attributing value and ascertaining ownership. As highlighted above, organisations are using analytics not only for their own internal data sets, but increasingly also for combinations of these and external data sets. As OECD,2014e, acknowledges, the attribution challenge "may be exacerbated by the fact that in practice a range of data may be gathered from different sources and for different purposes by businesses and combined in various ways to create value, making tracing the source of data challenging". And the re-combination of resource is not limited to "data mashups[28]". The development of data-driven services based on existing services provided via APIs is another common practice in the data ecosystem. The example of Ushahidi presented above is a good illustration.

That example also points to the global nature of the data ecosystem, which further increases the difficulty of value attribution because of the cross-border nature of the data flows involved. As highlighted above, the data ecosystem relies on cross-border transactions including data collection, processing and use around the world. Determining the functions (and countries) to which profit should be attributed continues to raise severe tax challenges (OECD, 2014e). Where multi-sided markets are used with the groups of customers from each side of the market spread around the globe, the attribution of profit becomes even more challenging, also given the use of different revenue models including freemium and the importance of cross-subsidies. Some have therefore raised the question of "whether the remote collection of data should give rise to nexus for tax purposes even in the absence of a physical presence" (physical establishment, PE), in which case non-resident enterprises could be taxed based mainly on their domestic activities involving data collection. It should be noted at this point that current tax treaties do not permit the taxation of business profits of non-resident enterprises in the absence of a PE to which these profits are attributable, raising further questions on the feasibility of data-based taxation (OECD, 2014e).

Exploitation of the key points of control, and competition

As highlighted above, actors can play different roles in the data ecosystem. Depending on their role and their market power, they will have more or less direct influence in shaping the ecosystem. The growing number of M&A activities related to big data businesses and the transformation of some businesses towards vertical integration described above are just two of the trends through which the data ecosystem is shaped. In addition, some dominant actors in the data ecosystem may have significant control and power over certain activities shaping the system. This section discusses the main points of control through which the data ecosystem can be influenced and eventually disrupted, focusing primarily on the main layers of the data ecosystem presented in Figure 2.2. The section builds on the control point analysis of the Internet developed by Clark (2012), which is a method for "determining which actors obtain power, economic or otherwise, by virtue of control over key components of the system".[29]

The data ecosystem contains a rich mix of control points that are distributed across the layers illustrated in Figure 2.2 and that may differ significantly across sectors.[30] The exploitation of these control points can raise serious competition and consumer protection concerns, as they can lead to the reduction of consumer choice and anticompetitive behaviour. Good governance of these points of control is therefore essential from a public policy perspective to assure that DDI leads to growth and the well-being of all members of society. Their identification is not, however, a simple task, especially given the rapidly evolving nature of the data ecosystem highlighted above.

Clark (2012) presents two criteria that can be used to identify points of control. These criteria assess the degree to which actors in the Internet ecosystem are controlled by others. They are:

- *The degree of choice*, which essentially tests whether users can "route around" a misbehaving actor. Where users have few possibilities to escape the control of the other actor, or where an actor has control over the choices of the users, a strong point of control can be assumed. A minimum level of choice is necessary for competition but also for trust in the data ecosystem as the ability to select among actors allows choosing those that are trustworthy. As Clark (2012) explains in the context of the Internet:

  If the user is to 'route around' a misbehaving actor, the design of the system must give the user that degree of choice. The tussle of control is often thus a tussle over who controls the choice. Examples … include which ISP to use, which DNS to use, which browser to use, and there are more subtle and complex choices that are embedded in the control picture. (Clark, 2012)

- *The degree of confidentiality and privacy*, which essentially tests "what options the actor has to observe what is being done" (Clark, 2012). The capacity to monitor and profile users, including in particular through the collection and analysis of personal data, creates a (soft) power of influence that enables actors to influence and perhaps limit the choices of users. As the monitoring and profiling activity becomes more exclusive, that power will grow commensurately.

Where users have few possibilities to escape the control of an actor, or where users have no real possibility to escape from the observation of an actor, a strong point of control can be assumed, at least at first. Alternatively, increasing the degree of choice and the confidentiality and privacy of users can mitigate the potential misuse of points of control in the data ecosystem. Possible measures to increase choice include those that

enhance interoperability through open standards, and data portability (see further below).[31] Possible measures to enhance privacy and confidentiality include privacy enhancing technologies including cryptography and privacy regulation (see Chapter 5).

The key points of control discussed in the sub-sections that follow focus on the main layers of the data ecosystem. While the Internet provides "a range of design principles that different actors use to 'blunt the instruments of control' by other actors" (Clark, 2012) (e.g. multi-homing, user-selected routes), this facility is less available in the case of the data ecosystem: lock-in and lack of interoperability are still common in some layers, notably in the IT infrastructure layer and the entrepreneur layer. In these layers users may have greatly narrowed choice once engaged with an actor, suggesting that these two layers are likely the strongest points of control. In addition, the layered structure of the data ecosystem, in which actors rely on services provided by the underlying layers, suggests that the power of control may be asymmetric in favour of the actors *in* the underlying layers. In that respect, ISPs have the strongest potential influence on the data ecosystem, as "they exercise ultimate control: if they do not forward packets, the operation fails" (Clark, 2012).

Internet access[32]

ISPs are the regional gatekeepers that provide access to the Internet through the physical transport infrastructure. While some ISPs are going further up the value chain of the data ecosystem by providing IT infrastructure, data and analytic services, most, if not all, still rely on their traditional business models, which consist of granting subscribers (businesses and consumers) access to the Internet. Having realised that they have "ultimate control" on the data flows on which the data ecosystem relies, some ISPs are looking into means for taking advantage of their position to generate more revenues, for example by, differentiating between classes of Internet traffic (e.g. gold, silver bronze) or by dedicating broadband capacity to certain applications including real-time applications requiring timely data transmission and guaranteed delivering times (i.e. quality of service, see OECD, 2014a).

The reorientation of ISPs' business models towards traffic prioritisation and discrimination of applications has raised a number of concerns among other actors in the data ecosystem. These concerns, which some have framed using the term "net neutrality"[33] are not however specific to the data ecosystem. The same concerns have been raised for example in the context of smart applications, such as connected television that expand and place additional capacity demands on the Internet (OECD, 2014a). That said, it is also true that as DDI becomes a new source of growth, the control of data flows become more and more critical.

Some have suggested that traffic prioritisation and the discrimination of applications could transform the business model of ISPs into a two-sided market (OECD, 2014a). In such a market, ISPs could impose charges on content or application providers in addition to end users. However, as OECD, 2014a, highlights, existing offers by ISPs to ensure fast delivery of content have not sufficiently attracted content providers so far. There are several reasons for this, one being that content providers have incurred other costs in order to improve the quality of their service, principally by building or contracting for Content Distribution Networks (CDNs). By caching content at multiple sites, the content provider can shorten the path that content must travel to reach an end user, thus increasing quality and reducing the resources needed for transport of the content over the Internet. As OECD,2014a, highlights, CDNs may be a way to balance the concerns of

policy makers that, on the one hand, content providers should have tools available to increase the quality of their service, but on the other investment in new applications should be encouraged so as not to put new content providers at a disadvantage relative to incumbents with respect to delivery of their application over the Internet.

There is one other aspect of the relationship of ISPs with other actors of the data ecosystem, including content/application providers. As stressed by OECD,2014a, to the extent that an ISP is vertically integrated into content/application provision – and so moves higher up the value chain, as mentioned above – it is important to remain alert to the possibility that it may have the incentive as well as the ability to behave anticompetitively with respect to independent content/application providers.

Application programming interfaces

Looking more closely at the IT infrastructure layer, proprietary solutions (including APIs) are strong potential points of control that could be exploited through vendor lock-in[34] and other anticompetitive measures. In the case of cloud computing, for example, recent surveys among potential cloud users have highlighted a lack of standards and of widespread adoption of existing open standards as one of the biggest barriers to the use of cloud computing (OECD, 2014b). Fear of potential vendor lock-in is often indicated as the reason. The lack of open standards is a key problem especially when it comes to the model of "platform as a service" (PaaS). In this service model, APIs are generally proprietary. Applications developed for one platform typically cannot easily be migrated to another cloud host. While data or infrastructure components that enable cloud computing (e.g. virtual machines) can currently be ported from selected providers to other providers, the process requires an interim step of manually moving the data, software and components to a non-cloud platform and/or conversion from one proprietary format to another. Consequently, once an organisation has chosen a PaaS cloud provider, it is – at least at the current stage – locked in (OECD, 2014b). Some customers have raised the concern that it will be difficult to extract data from particular cloud services that prevent some companies or government agencies from moving to the cloud. Another concern linked to this is that users can become extremely vulnerable to providers' price increases. This is the more relevant as some IT infrastructure providers may be able to observe and profile their users to apply price discrimination to maximise profit (see Chapter 5 of this volume).[35]

APIs, highlighted in this chapter as the "cytoplasm" lying between the layers of the data ecosystem, could thus be exploited as strategic point of control, for example by limiting users' choice in the applications used on top of a service provided over an API (see the example of Twitter in Box 2.4). Trends towards more closed APIs are therefore raising concerns among some actors that rely on open API for their innovative services. This is particularly relevant in view of the recent debate on the ability for legal entities to copyright APIs. This debate has gained significant momentum after a recent petition by the Electronic Frontier Foundation (EFF, 2014) to the United States Supreme Court in November 2014. The petition follows a court finding earlier in May 2012 that Google had infringed on Oracle's copyright on Java APIs in Android, "but the jury could not agree on whether it constituted fair use" (Duckett, 2014).

Box 2.4. **Competitive effects of Twitter's vertical integration**

Twitter's application programming interface (API) allows outside developers to build apps that can pull in information directly from Twitter to display in their own apps. The availability and openness of proprietary API's have been instrumental for the rapid expansion of apps and the growth of platforms such as Twitter.

Twitter has been pursuing a vertical integration strategy by acquiring and building a portfolio of apps. The company purchased apps such as TweetDeck (2011), Tweetie (2010) and Summize (2008) intending to later transform them into brand extensions that serve different platforms and services, e.g. search engines.

The result of this integration is that Twitter wants developers to start building apps that use Twitter, rather than Twitter apps. Twitter has been discouraging developers from using their APIs to make apps that compete directly with their platform, by rejecting apps that rely on tweet feeds via its API and by revoking API access. The risk of such an approach for Twitter or other growing platforms is that the uncertainty of future access to the API will stifle investment and innovation.

In August 2012, Twitter restricted the number of individual user tokens for an app that could access their APIs to 100 000. This essentially means that app developers are limited to 100 000 app installs on users' devices without special permission from Twitter to increase the number. Some developers were forced to require all members to re-login to free up unused keys for new users.

Source: OECD, 2013d, based on Musil, 2011; Mashable;[36] Twitter, 2012; and Yahoo News, 2013.

Intellectual property rights

The issue of API copyright highlighted above directly points to the role of *intellectual property rights (IPRs)*, which is often used strategically in the IT infrastructure layer as a key point of control (see OECD, 2015). This remains true despite the increasing use of open source software (OSS) applications, which have eased some of the constraints that IT infrastructure users have faced in the past (see Chapter 3). For example, some have expressed concerns that the patent US 7650331 B1 on MapReduce awarded to Google could put at risk companies that rely on the open source implementations of MapReduce such as Hadoop and CouchDB (Chapter 3). Such concerns may be justified, but given that Hadoop is widely used today – including by large companies such as IBM, Oracle and others, as well as by Google – expectations are that Google "obtained the patent for 'defensive' purposes" (Paul, 2010).[37] By granting a licence to (open source) Apache Hadoop under the Apache Contributor License Agreement (CLA), Google has officially eased fears of legal action against the Hadoop and CouchDB projects (Metz, 2010).

Data

Access to data can become a critical point of control in this ecosystem, where value creation and competitive advantage are directly related to the capacity to extract insights from (observed) data. The analysis of data can have a significant impact on confidentiality and the privacy of other actors, to the extent that these actors can be influenced and their choice eventually limited. Chapter 5 of this volume discusses in detail the risk of price discrimination, which is one possible way of exploiting data as a strategic point of control.

As actors across the data ecosystem acquire and control massive (proprietary) data sets, there is an increasing risk that "we're kind of heading toward data as a source of monopoly power", as Tim O'Reilly highlights in an interview with Bruner (2012). The risk of "monopoly power", however, must be assessed carefully on a case-by-case basis,

as it will typically depend on the extent to which data can be exploited as a control point. This in turns depends on factors such as the market (segment) under consideration, in particular its rate of technological change;[38] the data sources used; the degree of detriment to consumer welfare; the potential barriers to entry, including the level of investments required for building comparable data sets; and last but not least, other control points such as APIs and IPRs used sometimes in combination with data. Furthermore, it may also depend on the available means to escape the control of the dominant actor, including in particular the availability of open standards and data portability.

For example, access to points of sale (including to consumers' personal data), which is controlled by a single dominant data-driven enterprise, can become a strategic point of control that, if abused, could raise consumer protection and competition issues. In 2011 the Financial Times (FT), for example, pulled its iPad and iPhone apps from Apple's App Store after several month of negotiation. The primary rational for FT's reaction was not because 30% of revenue had to be shared with Apple, but to "keep control of customer data obtained through subscriptions" (Reuters, 2011). By switching its app to the open standard *HTML5* (see Box 2.5), the FT was finally able to bypass Apple's control, and to directly interact with iPad and iPhone users and so gain access to their data. As a consequence, the FT was able to gain more insights into its customers and increase the number of its digital subscribers by 14% within a year (Miller, 2013).

Box 2.5. **HTML5: An open standard for browsers, apps and operating systems**

HTML5 is an update of the HTML standard that dictates how content is displayed on the web. It will affect three key areas of the app ecosystem: i) mobile browsers, ii) mobile apps, and iii) mobile operating systems.

- *Browsers*: HTML5 is the next iteration of HTML, the web mark-up language that tells browsers how to display web pages. HTML5 is a significant evolution of the standard, in that it introduces richer functionality that allows websites in a browser to mimic the functionality of standalone apps.[39] *Strategy Analytics* (2011) estimates there were 336 million HTML5-capable smartphones sold in 2011, and predicts the number of HTML5-compatible phones sold in 2013 will reach 1 billion. One of the key benefits for app developers using HTML5 in a browser is that they are not tied to an app store that may require a share of app revenues. Despite advancements in the HTML standard, native apps (built specifically for one platform) often can make better use of specific hardware features of phones to deliver content, and often run faster than HTML5 content because they are tailored to a specific device or operating system.
- *Apps*: HTML5 can also be used as the core of standalone apps that can be written once and work across different mobile operating systems. The HTML5 can be viewed directly via a browser or "wrapped" into an app that is specific to a mobile operating system, so that the app that can take full advantage of the hardware potential of devices. These new hybrid solutions are emerging from companies such as PhoneGap and Marmalade. With the open-source PhoneGap, developers can write applications using HTML5, JavaScript and CSS, and then compile native apps using PhoneGap to take advantage of APIs for accelerometers, the camera, compass, etc.
- *Operating systems*: HTML5 content is available across platforms via HTML5-compliant browsers, but the emergence of new browser-based operating systems such as Chrome OS and Firefox OS that run apps could further promote use of the standard. In particular, Firefox OS from the Mozilla Corporation will only run HTML5 apps.

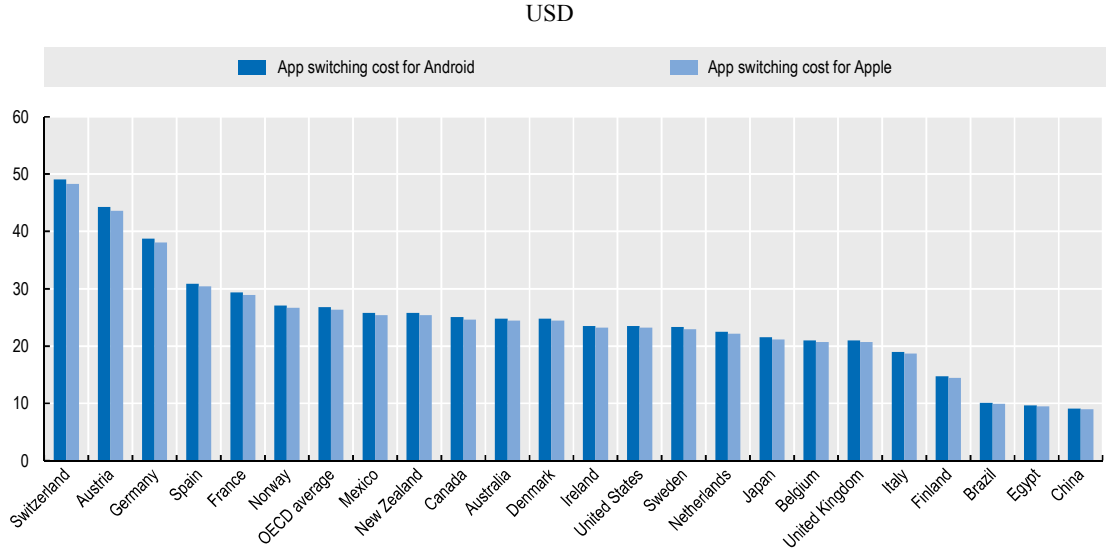
Source: OECD (2013d).

Walled garden

When it comes to multi-sided markets, the situation is more complex: points of control, including data but also other control points such as APIs and IPRs, can be exploited to command multiple sides of the market and thus create a "walled garden" (i.e. closed proprietary platforms). In the case of mobile application (app) platforms (e.g. Apple's App Store and Google Play), for example, consumers may be locked in due to upfront investments in both the hardware and software needed to access the platform, and barriers to data portability that prevent them from reusing their data in other applications. Developers may also be locked in due to upfront investments needed to develop the applications (including in particular the skills and competencies needed). Platform providers can consequently exploit the "stickiness" and lock-in effects of their platforms to reinforce their positions on all sides of their markets. It is important to note that these risks are not restricted to IT services, but this will increasingly also involve physical infrastructures and hardware, as the IoT becomes the dominant network.

Analysis (OECD, 2013d) of the average value of paid apps across countries shows that investments in paid apps could lead to consumer stickiness when it comes to switching platforms (Figure 2.6). Swiss users have an average of nearly USD 50 of paid apps on their phones. The OECD area average is estimated to be roughly USD 26 and these investments would be lost when switching platforms. It is important to highlight though that these sums do not cover all switching costs. They are only a relatively small percentage of the overall purchase price of many smartphones, which will also need to be replaced. Furthermore, as OECD, 2013b highlights, consumer stickiness to a specific platform will also be highly influenced by the amount of digital content that has been purchased on the platform and locked by digital rights management (DRM), as well as the data that have been generated over time.[40] According to Goldman Sachs, the explicit switching cost comes to an average of USD 122 to USD 301 per Apple's iOS device.[41]

Figure 2.6. **App switching costs by platform and by country, 2012**



Note: Data on the average number of apps on Apple's iOS platform were not available so the average number of paid Android apps was used with iOS prices to compute the Apple component.

Source: OECD, 2013d, based on *Think with Google* survey, "Our mobile planet" (2012), www.thinkwithgoogle.com/mobileplanet/en/downloads/.

Competition implications[42]

As highlighted in Chapter 4, the accumulation of data can lead to significant improvements in data-driven services that in turn can attract more users, leading to even more data that can be collected (positive feedback). For example, the more people use services such as Google search, recommendation engines such as provided by Amazon, or a navigation system by TomTom, the better services will be as they become more accurate in delivering requested sites and products, and in providing traffic information. As a result, the service can attract more users. Where data linkage is possible, the diversification of services can lead to further positive feedback. This feedback, which is also characteristic of markets with network effects, finally reinforces the market position of the service provider and has a tendency to lead to its market dominance, or at least to higher market concentration. As Shapiro and Varian (1999) highlighted: "positive feedback makes the strong get stronger and the weak get weaker, leading to extreme outcomes".[43]

Case-by-case analysis of the situation may be required because the degree to which competition issues emerge will typically depend on a number of factors, as highlighted above. However, there are a number of factors that make this analysis particularly difficult, and these may challenge the traditional approach used by competition authorities for assessing potential abuses and harms of market dominance and mergers. The following three types of challenges are highlighted: challenges in i) defining the relevant market, ii) assessing the degree of market power, and iii) assessing potential consumer detriment. As will become clear in the following sections, the factors behind these challenges are basically common to those questioning the attribution of value discussed above.

Challenges in defining the relevant market

Competition authorities rely on a definition of the relevant market as "one of the most fundamental concepts underpinning essentially all competition policy issues, from mergers, through dominance/monopolisation to agreements" (OECD, 2012b). It is the "analytical framework for the ultimate inquiry of whether a particular conduct or transaction is likely to produce anticompetitive effects" (OECD, 2012b). Defining the relevant market is necessary for assessing the effective competition level, including whether an incumbent with significant market power is vulnerable to new competition. Factors in the market definition process will typically include consideration of the goods and services which are perceived by consumers as substitutable, the geographic market, and a time dimension reflecting technological change and changes in consumer behaviour. Given the particular properties of the global data ecosystem, however, establishing a proper market definition can be particularly difficult for the following reason.

Multi-sided markets, such as enabled by data, challenge the traditional market definition, which generally focuses on one side of the market. That approach would tend to define the relevant market too narrowly in a multi-sided market case. As Filistrucchi et al. (2014) argue in the case of two-sided markets: "only in the case of a two-sided non-transactional market, and only when on side does not exert an externality on the other side, can one proceed to define the relevant market on the first side irrespective of the presence of the other side". In many cases however, multi-sided platforms must co-ordinate demand among the interdependent customer groups, and price changes on one side of the market will have "positive feedbacks on the other sides of the market"

(OECD, 2009). In the particular case of data-generating platforms, as has been highlighted, the motivation for the creation of multi-sided markets enabled by data is in many cases founded on exactly these positive feedbacks and externalities that data enable. As a result, focusing on one side of market will rarely lead to a proper market definition. An extended market definition is also justified due to the cross-subsidies often used across multiple sides of the platform. In other words, a proper market definition would have to include all sides involved in the cross-subsidy. Overall, as OECD (2009) highlights, it cannot be assumed that market power and abuse are any less prevalent in multi-sided markets than in traditional markets.

Challenges in assessing market power

Only once the relevant market has been properly defined, can the market power of the market participants be assessed. "[M]arket power can be thought of as the ability […] to sustain prices above competitive levels or restrict output or quality below competitive levels" (OFT, 2004).[44] However, a large share of data-driven products are provided for "free" in exchange for access to personal data, and/or in addition to an offer of a premium version as in the case of the freemium revenue model. In these cases information on prices for the single product will rarely be available, rendering it difficult to assess the degree of market power if applying the narrow market definition discussed above. Other mechanisms therefore have to be used, including in particular a proper market definition: as the data provided will typically be used for different purposes across multi-sided markets, market power will need to be assessed in most cases across all sides of the market as well. As Evans (2011) explains:

> *The existence of a free good signals that there is a companion good, that firms consider both products simultaneously in maximising profit, and that commonly used methods of antitrust analysis, including market definition, probably need to be adjusted to properly analyse two inextricably linked products. (Evans, 2011)*

Nor will assessing market value through the economic value of the collected (personal) data be helpful in most cases; as data have no intrinsic value, as already highlighted above. Admittedly, as possession of that data is necessary for a business to succeed, it can be assumed that the data have economic value. However, the monetary valuation of the same data set can diverge significantly among market participants and uses. This implies that focusing on the ability to sustain prices above competitive levels, is a less practical approach for assessing market power. The restriction of output or quality (to below competitive levels) should be more strongly considered by competition authorities. However, the dimensions that should be included as quality criteria are still not clear – in particular in regard to privacy, which some have argued should be considered when assessing the anticompetitive effects of a particular conduct or transaction (see next section).

Challenges in assessing potential consumer detriment

Anticompetitive behaviour and mergers are often assessed based on the consumer detriment or reduction in consumer welfare they could induce. However, in the particular case where data-driven services rely on personal data, privacy harms are still not fully acknowledged by competition authorities, which will tend to direct the specific privacy issues to the privacy protection authorities; the latter, however, have no authority over competition issues.

The degree to which privacy harms should be considered when assessing anticompetitive behaviour and mergers is therefore still an ongoing debate. That debate was triggered most notably by the former Commissioner Pamela Jones Harbour's dissent in the Federal Trade Commission decision (FTC, 2007) to clear the Google/DoubleClick merger. The dissent was based inter alia on concerns that "the network effects from combining the parties' data would risk depriving consumers of meaningful privacy choices" (Cooper, 2013). Harbour and Koslov (2010) therefore called for competition authorities to consider whether "achieving a dominant market position might change the firm's incentives to compete on privacy dimensions" and thus to promote development of innovative privacy-enhancing technologies and services.

This underscores the need for further dialogue among competition, privacy and consumer protection authorities on potential detriment due to DDI. A preliminary EDPS (2014) opinion confirms that:

> *There is currently little dialogue between policy makers and experts in these fields. […] It is essential that synergies in the enforcement of rules controlling anti-competitive practices, mergers, the marketing of so-called "free" on-line services and the legitimacy of data processing are explored. This will help to enforce competition and consumer rules more effectively and also stimulate the market for privacy-enhancing services. (EDPS, 2014)*

At this point it is important to emphasise that the competition issues discussed above should not be neglected by competition authorities, even when they are engaged in a dialogue with privacy and consumer protection authorities. DDI does not always involve personal data, and the competition issues raised above may still occur in the case of non-personal data; in those cases privacy and consumer protection authorities may have no jurisdiction. The accumulation and control of M2M and sensor data, for example, may raise a number of competition issues in the near future, as data and analytics are increasingly used in areas such as manufacturing and agriculture where non-personal data may become a strategic point of control as well.

Furthermore, it should be highlighted that most competition jurisdictions only enable their authorities to block or challenge anticompetitive practices in which the consequent lessening of competition leads to detriment. If no competition issues are raised, competition authorities will have no jurisdiction. For example, a merger between two companies that do not in any way compete, but whose data sets when combined create links that harm the privacy of consumers, would generate a detriment, but no loss of competition. It that case, competition authorities may not have the right to take action, but consumer and/or privacy protection authorities would.

The free flow of data, and the open Internet

The free flow of information and data is not only a condition for information and knowledge exchange, but also a vital condition for the globally distributed data ecosystem as it enables access to GVCs and markets. As stated already in the *OECD (1985) Declaration on Transborder Data Flows*, "these flows acquire an international dimension, known as Transborder Data Flows", which also favour trade between countries and global competition among actors in the data ecosystem (see Annex of this chapter). In other words, barriers to the free flow of data can limit the effects of DDI by limiting trade and competition, for example.

Some of the barriers to the free flow of data are the intended or unintended results of measures affecting the openness of the Internet (see Chapter 3). These include technical means such as IP package filtering, used *inter alia* to optimise the flow of data for specific purposes, or "data localisation" efforts, either through territorial routing or legal obligations to locate servers in local markets. The social and economic effects of limiting the openness of the Internet are still unknown, although a number of studies have tried to assess the economic costs of barriers. A 2014 working paper by the European Centre for International Political Economy (ECIPE, 2014) aims to quantify the losses resulting from data localisation requirements and related privacy/security laws in seven jurisdictions. According to these estimates, data localisation requirements may result in considerable GDP losses if economy-wide requirements were to be introduced on top of existing privacy/security legislation.[45] However, this study conflates data localisation requirements and privacy and security legal requirements. A more comprehensive analysis is therefore needed that would separate out the economic effects of data localisation requirements from privacy, security and IPR regulations.

There is common interest among countries in finding consensus on how to maintain a vibrant and open Internet and in exchanging views on better practices. The OECD's High-Level Meeting on the Internet Economy on 28-29 June 2011 discussed the openness of the Internet and how best to ensure the continued growth and innovation of the Internet economy. The resulting draft communiqué, which led to the OECD (2011b) *Council Recommendation on Principles for Internet Policy Making*, contains a number of basic principles whose goal is to help ensure that the Internet remains open and dynamic, that it "allows people to give voice to their democratic aspirations, and that any policy-making associated with it must promote openness and be grounded in respect for human rights and the rule of law". The following first five principles are highlighted here as highly relevant for the use of data. This is not to say that other principles are less important to DDI overall:

1. promote and protect the global free flow of information
2. promote the open, distributed and interconnected nature of the Internet
3. promote investment and competition in high speed networks and services
4. promote and enable the cross-border delivery of services
5. encourage multi-stakeholder cooperation in policy development processes.

Interoperability and (open) standards

Barriers to the free flow of data are an issue not only across borders but also across sectors and organisations, including between organisations and individuals (consumers and citizens). Many actors in the data ecosystem still face barriers to data interoperability and portability. Despite the widely agreed benefits, there are still significant (non-legal) issues limiting data exchange and interoperability. This is in particular the case in sectors that require significant investment with a high threshold for new entrants, and more especially capital-intensive industries. The datafication of agriculture, for instance, enables new services from start-ups like Crop-R, but it is driven by innovations from incumbents like John Deere, Monsanto and Lely that enhance their machines and tools with sensors and connectivity to capture and use the data. Interoperability and standards enabling data exchange across the different incumbents' services can be crucial for start-ups like Crop-R.

Interoperability

Open Internet standards such as TCP/IP and HTML5 are crucial for the global data ecosystem – which, as highlighted above, heavily relies on the open Internet for its functioning. In addition, the reuse of data and of data-driven services underlines the importance of (open) standards related to APIs and data formats (including the metadata). However, the lack of appropriate standards that results in potential vendor lock-in and vendors' exploitation of control points in the data ecosystem is still an issue for many users. Vendor lock-in in the cloud computing industry was mentioned above; attempts have been made in that industry to extend general programming models with cloud capabilities in order to enhance interoperability, in particular for PaaS (Schubert et. al., 2010). However, these attempts have not met with success. Promoting open standards for APIs and further work on interoperability are therefore seen as the appropriate response to this problem. As a result many initiatives are under way, covering the full spectrum from infrastructure standards – such as virtualisation formats and open APIs for management – to standards for web applications and services, security, identity management, trust, privacy, and linked data.[46]

But even if data can be extracted, reusability will typically be limited if data are not machine readable and cannot be reused across IT systems (i.e. data interoperability, see Box 2.6). Data are rarely harmonised across sectors or organisations as individual units collects and/or produce their own set of data using different metadata, formats and standards. This means that even if access to data is provided, the data cannot be reused in a different context. This can make it difficult to reuse data for new applications in particular. Unresolved interoperability issues are therefore still high on the e-government agendas of many OECD countries (see Chapter 10). For instance, interoperability of data catalogues, or the creation of a pan-European data catalogue, is a major challenge EU policy makers are facing at the moment.

Box 2.6. **The role of standards for data interoperability**

Reusability of data typically requires that data are machine readable and can be reused across IT systems (i.e. interoperability). Some data formats that are considered machine readable are based on open standards such RDF (Resource Description Framework), XML (eXtensible Markup Language), and more recently JSON (JavaScript Object Notation). But other standards include file formats such as CSV (comma-separated values) and proprietary file formats such as the Microsoft Excel file formats.

To further enable data linkage, meta-data are often needed. They provide the context without which primary data cannot be accessed, linked, or fully understood. As data become abundant and data analytics increasingly automated, finding and making sense of data often requires meta-data. As Cukier (2010) illustrates, meta-data make (primary) data "useable and meaningful as a large library is useless without a card-catalogue system to organise and find the books". Meta-data can be categorised in several types depending on their purpose (see NISO, 2004). Some metadata are provided as open standards, such as the Dublin Core Metadata Term*s*, which defines 15 meta-data elements for describing (web and physical) resources.[1]

1. The Dublin Core Metadata Terms were endorsed in IETF (Internet Engineering Task Force) RFC 5013 and ISO (International Organization for Standardization) Standard 15836-2009.

Data portability

An important development in the area of data portability and interoperability is the increasing role of consumers in the data ecosystem. As highlighted above, consumers play an important role in promoting the free flow of their own personal data across organisations. This role is strengthened by the Individual Participation Principle of the OECD (2013c) *Recommendation of the Council concerning Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data* (OECD Privacy Guidelines) (see Chapter 5 of this volume). Government initiatives are promoting *data portability* and are thus contributing to the promotion of 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,[47] 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,[48] a US initiative launched in 2006 that provides a model for Vendor Relationship Management by individual consumers. Finally, the right to data portability suggested by the EC in the current proposal for reform of their 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).

Data portability may involve significant costs to those that need, want, or must implement portability in their (existing) data-driven services. These include, costs both for developing and maintaining the mechanisms for enhanced data access, and for complying with relevant regulations (see Chapter 5). This may raise questions about who should bear the costs for developing and maintaining these mechanisms.[49]

2.4. Key findings and policy conclusions

A global data ecosystem is emerging in which, more than ever before, data and analytic services are traded and used across sectors and across national borders. For the ICT industry alone, this represents a USD 17 billion business opportunity that is growing at more than 40% on average every year since 2010. As a result, top ICT companies are strengthening their position through acquisitions of young start-ups specialised in big data technologies and services and/or through collaboration with potential competitors (co-opetition) in open source projects such as Hadoop. IBM was the most active acquirer of big data companies in 2012, followed by Oracle.

The top ICT firms contributing to the Hadoop ecosystem are to a large extent companies registered in the United States, with the exception of Yahoo Japan, NTT Data and Fujitsu (Japan), SAP (Germany), Persistent Systems (India) and Acer (Chinese Taipei). Most of the top ICT firms in the Hadoop ecosystem are Internet and software firms. Nevertheless, some hardware firms, in particular IT equipment firms, are heavily involved in big data-related technologies as well. Semiconductor firms, such as Intel and AMD, are the exceptions.

But the economic impact of the global data ecosystem goes far beyond the market prospects of the ICT industry, which mainly supplies goods and services for data collection, processing, and analysis. The data ecosystem involves a wide range of different types of actors with different business models and technologies. Besides ISPs and IT infrastructure providers, this includes in particular data service providers, analytic service providers, and data-driven entrepreneurs, many of these are start-ups. Many also act as users and producers of data and analytics, which suggests that the data ecosystem is a logical continuation of Web 2.0.

The global data ecosystem involves global value chains (GVCs), in which companies increasingly divide up their data related processes and locate productive activities in many countries. Figures on the distribution of data-driven services are not known. However, analysis of the world's top Internet sites suggests that data-driven services may be concentrated in the United States, which alone accounted for almost 60% of all top sites hosted in the OECD area in 2013, or more than 50% of all top sites hosted in OECD area plus Brazil, China, Colombia, Egypt, India, Indonesia, Russia, and South Africa taken together. The concentration of sites in the United States is most likely related to its well-functioning co-location and backhaul market, which reinforce the flourishing data ecosystem in that country. Statistics on trade in ICT-related services suggest further that the largest exporters of ICT service in 2013 – India, Ireland, the United States, Germany, the United Kingdom, and China – are more likely to be the largest destinations of cross-border data flows. As a consequence, the leading OECD importers of ICT-related services are also the major sources for trade-related data, and they include in particular the United States and Germany.

The characteristics of the global data ecosystem could create opportunities for BEPS through aggressive tax planning by multinational enterprises; this involves making use of gaps in the interaction of different tax systems to artificially reduce taxable income or shift profits to low-tax jurisdictions. What makes such action possible is the data ecosystem's ability to challenge the current paradigm used by tax authorities to determine where tax-relevant economic activities are carried out and value is created. Therein lies the difficulty in i) measuring the monetary value of data, ii) determining data ownership, and iii) acquiring a clear picture of the global distribution and interconnectedness of data-driven services..

The data ecosystem contains a rich mix of points of control that are distributed across all its layers, which however differ significantly across sectors. The exploitation of these points of control can raise serious competition and consumer protection concerns when they lead to the reduction of consumer choice, and anticompetitive behaviour. Lack of interoperability and vendor lock-in are two major risks through which points of controls can be exploited. In the area of cloud computing, the lack of open standards is still a huge problem, in particular in the area of platform as a service (PaaS). But points of control in the entrepreneurial layer also exist. These include, for example, data and walled gardens (i.e. closed proprietary platforms) based on multi-sided markets.

Analysis of points of control underlines the importance of (open) standards related to APIs and data formats. Lack of interoperability is among the most challenging barriers to the reuse of data and data-driven services in the data ecosystem. This is especially the case where data are not provided in a machine readable format and thus cannot be reused across IT systems. Individuals (consumers) also play an important role in promoting the free flow of their personal data across organisations if data portability is possible. Government and private sector initiatives promoting data portability are therefore

contributing to the free flow of data across organisations, and in so doing are strengthening the participation of individuals in DDI processes.

Characteristics of the global data ecosystem may also challenge the traditional approach employed by competition authorities to assess potential abuses and harms of market dominance and mergers. Challenges include: i) defining the relevant market, ii) assessing the degree of market concentration, and iii) ascertaining potential consumer detriment. Policy makers should encourage dialogue between competition, privacy and also consumer protection authorities, so that i) potential consumer harms due to DDI are taken into account, ii) synergies in enforcing rules controlling privacy violations, anticompetitive practices and mergers are unleashed, and iii) firms' incentives to compete with privacy-enhancing goods and services are increased.

Barriers to the open Internet, whether legitimate or not, can limit the effects of DDI. Some of these barriers may be technical, such as IP package filtering, or regulatory, such as "data localisation" requirements. They may result from business practices or government policies. Some of these have a legal basis, such as privacy and security (see Chapter 5), as well as the protection of trade secrets and copyright (OECD, 2015). However, these barriers can have an adverse impact on DDI – for example, if they limit trade and competition. Governments looking to promote DDI in their countries should consider further the OECD (2011b) *Council Recommendation on Principles for Internet Policy Making* as well as ongoing OECD work to develop better understanding of the characteristics and the social and economic impacts of an open Internet.

Annex – OECD (1985) Declaration on Transborder Data Flows

(Adopted by the Governments of OECD Member countries on 11th April 1985)

Rapid technological developments in the field of information, computers and communications are leading to significant structural changes in the economies of Member countries. Flows of computerised data and information are an important consequence of technological advances and are playing an increasing role in national economies. With the growing economic interdependence of Member countries, these flows acquire an international dimension, known as Transborder Data Flows. It is therefore appropriate for the OECD to pay attention to policy issues connected with these transborder data flows.

This declaration is intended to make clear the general spirit in which Member countries will address these issues.

In view of the above, the GOVERNMENTS OF OECD MEMBER COUNTRIES:

- Acknowledging that computerised data and information now circulate, by and large, freely on an international scale;
- Considering the OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data and the significant progress that has been achieved in the area of privacy protection at national and international levels;
- Recognising the diversity of participants in transborder data flows, such as commercial and non-commercial organisations, individuals and governments, and recognising the wide variety of computerised data and information, traded or exchanged across national borders, such as data and information related to trading activities, intracorporate flows, computerised information services and scientific and technological exchanges;
- Recognising the growing importance of transborder data flows and the benefits that can be derived from transborder data flows; and recognising that the ability of Member countries to reap such benefits may vary;
- Recognising that investment and trade in this field cannot but benefit from transparency and stability of policies, regulations and practices;
- Recognising that national policies which affect transborder data flows reflect a range of social and economic goals, and that governments may adopt different means to achieve their policy goals;
- Aware of the social and economic benefits resulting from access to a variety of sources of information and of efficient and effective information services;
- Recognising that Member countries have a common interest in facilitating transborder data flows, and in reconciling different policy objectives in this field;

- Having due regard to their national laws, do hereby DECLARE THEIR INTENTION TO:

1. Promote access to data and information and related services, and avoid the creation of unjustified barriers to the international exchange of data and information;
2. Seek transparency in regulations and policies relating to information, computer and communications services affecting transborder data flows;
3. Develop common approaches for dealing with issues related to transborder data flows and, when appropriate, develop harmonised solutions;
4. Consider possible implications for other countries when dealing with issues related to transborder data flows.

Bearing in mind the intention expressed above, and taking into account the work being carried out in other international fora, the GOVERNMENTS OF OECD MEMBER COUNTRIES:

Agree that further work should be undertaken and that such work should concentrate at the outset on issues emerging from the following types of transborder data flows:

1. Flows of data accompanying international trade;
2. Marketed computer services and computerised information services; and
3. Intracorporate data flows.

The GOVERNMENTS OF OECD MEMBER COUNTRIES AGREED to co-operate and consult with each other in carrying out this important work, and in furthering the objectives of this Declaration.

Notes

1 This includes the market for technologies and services related to data storage, which is expected to be the fastest growing segment, followed by networking, and services.

2 This chapter is partly based on a follow-up study to TNO (2013), which was provided to the OECD as a contribution by the government of the Netherlands. To allow for an extensive investigation and detailed mapping of developments, TNO employed for the case studies a combination of top-down and bottom-up approaches. The case studies focus on a specific topic as a starting point of departure, and then the network exploration reaches beyond that initial domain in search of actors and markets that span the boundaries between sectors.

3 The notion of "entrepreneur" is to be understood here in a broader sense to include not only start-up entrepreneurs, but also civic entrepreneurs, who are engaged in social innovation, as well as public servants who are innovating in the public sector to give few examples. Ries (2011) discusses this broader notion of "entrepreneur" in more detail.

4 Data analytics is also used by ISPs for timely data transmission and for guaranteeing the delivering time of sensitive data even in crowded networks, through for example quality of service (QoS).

5 As described further below, many actors – including IT infrastructure providers, data providers, analytic service providers and data-driven entrepreneurs – are contributing to the development of open source software tools such as Hadoop and R, and are also generating, sharing or selling their data to third parties that can reuse the data for the development of new services.

6 The extent to which the data ecosystem could be referred to as the Web 4.0 (Web 3.0 being the Semantic Web) is left to the reader to decide.

7 Of the 100 randomly selected start-ups focusing on "big data" or "big data analytics" analysed by Hartmann et al. (2014), 70 businesses had a B2B business models, while 17 businesses built their businesses solely on a B2C model. The remaining 13 businesses used both models, B2B and B2C.

8 In January 2012 for example, Orange signed an agreement with Mediamobile, allowing it to use FMD data for its traffic information service V-Trafic – see www.traffictechnologytoday.com/news.php?NewsID=36182.

9 As Dumbill (2012c) explains: "Practical big data implementations don't in general fall neatly into either structured or unstructured data categories. You will invariably find Hadoop working as part of a system with a relational or MPP database."

10 "The MIT License is a permissive license that is short and to the point. It lets people do anything they want with your code as long as they provide attribution back to you and don't hold you liable. jQuery and Rails use the MIT License." (See http://choosealicense.com/).

11 The BSD License is "a permissive license that comes in two variants, the BSD 2-Clause and BSD 3-Clause. Both have very minute differences to the MIT license." (See http://choosealicense.com/licenses/).

12 "The Apache License is a permissive license similar to the MIT License, but also provides an express grant of patent rights from contributors to users. Apache, SVN, and NuGet use the Apache License." (See http://choosealicense.com/.)

13 "The GPL (V2 or V3) is a copyleft license that requires anyone who distributes your code or a derivative work to make the source available under the same terms. V3 is similar to V2, but further restricts use in hardware that forbids software alterations. Linux, Git, and WordPress use the GPL." (See http://choosealicense.com/.)

14 Clouds are sometimes also classified as private, public, or hybrid, according to their ownership and management control mechanisms.

15 These include e.g. Aristotle, LexisNexis, DocuSearch, Experian, Merlin Data, Pallorium. It is interesting to note that a combination of several data types – such as address, date of birth, social security number, credit record and military is estimated to cost around USD 55.

16 The public sector in the United States employed on average 1.6 database administrators per 1 000 employees in 2011.

17 On his first day in office, US President Obama announced his strategy for "open government", and the European Commission recently launched its Open Data Portal (Veenstra and en Broek, 2013).

18 These included Index Ventures and Khosla Ventures, SV Angel, Yuri Milner's Start Fund, Stanford Management Company, PayPal Founder Max Levchin; Google Chief Economist Hal Varian; and Applied Semantics' Co-Founder and Factual Chief Executive Officer Gil Elbaz.

19 The whole game is like an ongoing experiment. Foldit was successfully used to remodel the backbone of a computationally designed enzyme that catalyses the Diels-Alder reaction, which brings together two small molecules to form a particular kind of bond that the scientists were interested in making (see www.nature.com/nbt/journal/v30/n2/full/nbt.2109.html).

20 Walmart Labs is developing a number of (internal) solutions such as Social Genome, which allows Walmart to reach to potential customers, including friends of direct customers, who have mentioned specific products online, to provide discounts on these exact products. Social Genome builds on public data from the web (including social media data) as well as Walmart's proprietary data such as its customer purchasing and contact data. "This has resulted in a vast, constantly changing, up-to-date knowledge base with hundreds of millions of entities and relationships" (Big Data Startups, 2013).

21 For example, when population data from different sources are linked to health-sector data, some causes of illness can be better understood that could hardly be explained otherwise. An example is the analysis of environmental determinants of illnesses linked to nutrition, stress and mental health (OECD-NSF, 2011).

22 The 2012 Technology Foresight Forum (the Foresight Forum), held on 22 October 2012, highlighted the potential of big data analytics as a new source of growth. It put big data analytics in the context of key technological trends such as cloud computing,

smart ICT applications and the Internet of Things. It focused on the socioeconomic implications of harnessing data as a new source of growth and looked at specific areas: science and research (including public health), marketing (including competition) and public administration (see http://oe.cd/tff2012).

23 In a strategy referred to as "follow the moon", for example, companies such as Google are automatically and seamlessly shifting computing operations around the globe so computing heavy operations are done at night when the temperature is lower and cooling costs are cheaper (Higginbotham, 2009).

24 "For this analysis, the generic top-level domains were omitted from the list, as there is no reliable public data as to where the domains are registered. Out of the one million top sites, 948 00 were scanned, 474 000 were generic top-level domains, 40 000 had no identifiable host country, around 4 000 had no identifiable domain, just an IP-address. The remaining 429 000 domains were analysed and their hosting country identified. For each country the percentage of domains hosted in the country were [sic] identified" (OECD, 2014b; see also Pingdom, 2012).

25 This section is in part adapted from OECD, 2014d.

26 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 data brokers in the United States such as Pallorium and LexisNexis.

27 The Individual Participation Principle of the OECD (2013c) *Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data* (OECD Privacy Guidelines), for example, recommends that individuals should 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; […] and d) to challenge data relating to him". These rights of the data subject are far-reaching and limit any possibility of exclusive right to the storage and use of personal data [by the data controller].

28 Mashups are web applications that use and combine content from different sources, including but not limited to web documents such as web pages and multimedia content; data such as cartographic and geographic data; and applications converter, communication, and visualisation tools.

29 However, the analysis undertaken here does not go much into the details of all relevant interaction, as does that undertaken by Clark (2012), which would have been necessary for a comprehensive control point analysis. Further studies following this method are therefore recommended for the future.

30 In many fields, competitive differentiation is most likely gained via proprietary data, data capturing interfaces, vertical analytics and visualisation. When speed is a differentiating asset, the data infrastructure will be essential as well. This is the case in high-frequency trading where hundreds of millions of dollars are being invested by companies in proprietary fibre optic cables to gain a milliseconds edge on their stock trading competitors.

31 This is in particular relevant when the collection of historical data series increases the value of the data. When these historical data are not portable, there is a (soft) lock-in, because the user loses this value. This is the case for instance in the agricultural

sector, where building a historical database provides increasing value and users are locked in and cannot move their data to another provider.

32 This section is in part adapted from OECD, 2014a.

33 OECD,2014a does not refer to "net neutrality", but takes the position that "the actions of market players can be described more accurately, with other terms, when considered across multiple countries that may not always use this term or do so with different interpretations".

34 Lock-in is still an attractive business strategy to maximise profit. As Swire and Lagos (2013) have argued, "the ability to attract users to a software service, and keep them there in at least some instances, is an important incentive for innovation and new entrants". Depending on their market and their market power, businesses could however develop a critical point of control though their lock-in capacity, that if exploited could raise consumer protection and competition issues.

35 Netflix, for example, uses Amazon's Web Services (AWS) for computing and storage (over 1 petabyte). Almost all of Netflix's information technology services run on AWS. Additionally, Netflix uses the services from Aspera to manage its data in Amazon's cloud. Netflix relies heavily on Amazon's infrastructure and, in the process, is one of Amazon's biggest customers. Simultaneously, Amazon is also a competitor in the on-demand video market with its Amazon Prime services, and Netflix is supporting the development of "an ecosystem that could lead to more competition for Amazon in the long term." Coincidentally, the adoption of these technologies by other cloud infrastructure providers would make it easier for Netflix to migrate to a provider other than Amazon (see King, 2013).

36 http://mashable.com/2010/09/17/google-voice-app-store-return/.

37 As Paul (2010) explains: "Many companies in technical fields attempt to collect as many broad patents as they can so that they will have ammunition with which to retaliate when they are faced with patent infringement lawsuits." For more on IP strategies (see OECD, 2015).

38 Markets featuring a series of disruptive innovations can lead to patterns in which firms rise to positions of temporary monopoly power but are then displaced by a competitor with superior innovation.

39 For example, HTML5 has tags for showing video on a web page or allowing users to drag and drop elements within the browser window.

40 Music purchases are commonly free of DRM restrictions and can played on nearly any device, regardless of platform. Downloaded video content, however, is almost always tied to one platform and cannot be viewed on others (OECD, 2013b).

41 See www.iclarified.com/entry/comments.php?enid=22914&laid=33#commentsanchor, accessed 15 May 2015.

42 This section benefited from the OECD Competition Committee hearings on the digital economy. Two hearings were held, in October 2011 and February 2012. OECD, 2013e includes an executive summary, an issues note by the Secretariat, a summary of each hearing, papers from panellists Eric Brousseau and Tim Wu and written submissions from: France, Japan, Norway, Poland, Turkey and Russia.

43 This observation has been confirmed in the OECD (2013f) work on competition in the digital economy undertaken under the first phase of the OECD (2013g) horizontal project on New Sources of Growth: Knowledge-based Capital (KBC 1). The conclusion reached was that markets characterised by the economic properties described above (increasing returns to and economies of scale and scope, paired with multi-sided markets and network effects) can lead to a "winner takes all" outcome where monopoly is the nearly inevitable outcome of market success.

44 It should be noted that in 2014 the UK Office of Fair Trading (OFT) became part of the Competition and Markets Authority (CMA), following reform of the competition system in the United Kingdom.

45 The study estimates the following effects: Brazil (-0.8%), the EU (-1.1%), India (-0.8%), Indonesia (-0.7%), Korea (-1.1%).

46 As an example, the Swedish standardisation committee "DIPAT" – SIS/TK 542, run by the Swedish Standards Institute (SIS), launched an initiative to work on national and European-level standardisation issues, linking and aligning the initiative with global efforts run by Subcommittee 38 of the Joint Technical Committee 1 of the International Organization for Standardization and the International Electrotechnical Commission (ISO/IEC JTC 1/SC 38). The goal is to assist in the development of harmonised, sustainable and well-designed standards.

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

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

49 The question is, should the data controller who will have to implement the mechanism pay, or the customers who request data portability, or the government that promotes the free flow of data across organisations and individuals?

References

Adner, R. (2006), "Match your innovation strategy to your innovation ecosystem", *Harvard Business Review*, April, http://pds12.egloos.com/pds/200811/07/31/R0604Fp2.pdf, accessed 12 June 2015.

Amazon (2009), "Amazon Elastic MapReduce Developer Guide API", 30 November, http://s3.amazonaws.com/awsdocs/ElasticMapReduce/latest/emr-dg.pdf, accessed 12 June 2015.

Angwin, J. (2010), "The web's new gold mine: Your secrets", *Wall Street Journal*, 30 July, http://online.wsj.com/article/SB10001424052748703940904575395073512989404.html.

Arthur D. Little, (2013), "Cloud from Telcos: Business distraction or a key to growth?", Arthur D. Little, www.adlittle.com/downloads/tx_adlreports/2013_TIME_Report_Cloud_from_Telcos.pdf, accessed 25 May 2015.

Assay, M. (2013), "VMWare: If Amazon wins, we all lose", *Readwrite,* http://readwrite.com/2013/03/01/vmware-if-amazon-wins-we-all-lose#awesm=~ohpgw6RGoncktJ, accessed 22 May 2015.

Bakhshi, H. and J. Mateos-Garcia (2012), "Rise of the Datavores: How UK businesses can benefit from their data", *Nesta,* 28 November, www.nesta.org.uk/publications/rise-datavores-how-uk-businesses-can-benefit-their-data.

Biesdorf, S., D. Court and P. Wilmott (2013), "Big data: What's your plan?", *McKinsey Quarterly,* McKinsey & Company, www.mckinsey.com/insights/business_technology/big_data_whats_your_plan, accessed 22 May 2015.

Big Data Startups (2013), "Walmart makes big data part of its DNA", http://smartdatacollective.com/bigdatastartups/111681/walmart-makes-big-data-part-its-social-media, accessed 22 May 2015.

Bonina, C. (2013), "New business models and the value of open data: Definitions, challenges, opportunities", RCUK Digital Economy Theme, www.nemode.ac.uk/wp-content/uploads/2013/11/Bonina-Opendata-Report-FINAL.pdf, accessed 12 June 2014.

Bruner, J. (2012), "Will data monopolies paralyze the Internet?", *Forbes*, 12 April, www.forbes.com/sites/jonbruner/2012/04/12/will-data-monopolies-paralyze-the-internet/.

Brynjolfsson, B. and A. McAffee (2012), "Big data: The management revolution", *Harvard Business Review*, October, http://hbr.org/product/big-data-themanagement-revolution/an/R1210C-PDF-ENG, accessed 12 June 2015.

Bunge, J. (2014), "Big data comes to the farm, sowing mistrust: Seed makers barrel into technology business", *The Wall Street Journal*, 25 February.

Capgemini Consulting (2013), "The open data economy: Unlocking the economic value by opening government and public data", www.capgemini.com/resources/the-open-data-economy-unlockingeconomic-value-by-opening-government-and-public-data, accessed 12 June 2014.

Chang, F. et al. (2006), "Bigtable: A distributed storage system for structured data", Google, appeared in Seventh Symposium on Operating System Design and Implementation (OSDI'06), November, http://research.google.com/archive/bigtable.html, accessed 24 May 2015.

Chen, L. et al. (2012), "Business intelligence and analytics: From big data to big impact", *MIS Quarterly*, Vol. 36, No. 4, pp. 1165-88.

Christensen, C.M. (1997), *The Innovator's Dilemma*, Harvard Business School Press, Boston.

Clark, D. (2012), "Control point analysis", 2012 TRPC Conference, 10 September, Social Science Research Network (SSRN), http://ssrn.com/abstract=2032124.

Cooper, J.C. (2013), "Privacy and antitrust: Underpants gnomes, the First Amendment, and subjectivity", *George Mason Law Review*, Rev. 1129 (2013), http://www.law.gmu.edu/assets/files/publications/working_papers/1339PrivacyandAntitrust.pdf, accessed 24 May 2015.

Criscuolo, P., N. Nicolaou and A. Salter (2012), "The elixir (or burden) of youth? Exploring differences in innovation between start-ups and established firms", *Research Policy*, Vol. 41, No. 2, pp. 319-333.

Cukier, K. (2010), "Data, data everywhere", *The Economist Special Report*, 25 February, www.economist.com/node/15557443.

Datameer (2013), "Hadoop Ecosystem: Who has the most connections", Datameer blog, http://datameer2.datameer.com/blog/wp-content/uploads/2013/01/hadoop_ecosystem_full2.png, accessed 12 June 2015.

Dean, J. and S. Ghemawat (2004), "MapReduce: Simplified data processing on large clusters", in Sixth Symposium on Operating System Design and Implementation (OSDI'04), December, San Francisco, http://research.google.com/archive/mapreduce.html, accessed 12 June 2015.

Duckett, C. (2014), "Computing experts call for repeal of copyrightable API decision", ZDnet, 10 November, www.zdnet.com/computing-experts-call-for-repeal-of-copyrightable-api-decision-7000035590/.

Dumbill, E. (2012a), "Microsoft's plan for big data", *O'Reilly Planning for Big Data*, http://oreilly.com/data/radarreports/planning-for-big-data.csp, accessed 12 June 2014.

Dumbill, E. (2012b), "Big data market survey", *O'Reilly Planning for Big Data*, http://oreilly.com/data/radarreports/planning-for-big-data.csp, accessed 12 June 2015.

Dumbill, E. (2012c), "Big data market survey: Hadoop solutions", *O'Reilly Strata*, http://strata.oreilly.com/2012/01/big-data-ecosystem.html, accessed 12 June 2015.

Dumbill, E. (2011), "Five data predictions for 2012", *O'Reilly Strata*, http://strata.oreilly.com/2011/12/5-big-data-predictions-2012.html, accessed 12 June 2014.

Dumbill, E. (2010), "The SMAQ stack for big data: Storage, MapReduce and Query are ushering in data-driven products and services", *O'Reilly Radar*, 22 September, http://radar.oreilly.com/2010/09/the-smaq-stack-for-big-data.html.

Dwyer, J. (2010). "Four nerds and a cry to arms against Facebook", *New York Times*, 11 May, www.nytimes.com/2010/05/12/nyregion/12about.html?_r=1.

ECIPE (European Centre for International Political Economy) (2014), "A friendly fire on economic recovery: A Methodology to estimate the costs of data regulations", *ECIPE Working Paper*, No. 02/2014, www.ecipe.org/media/publication_pdfs/WP22014.pdf, accessed 12 June 2015.

EDPS (2014), "Privacy and competitiveness in the age of big data: The interplay between data protection, competition law and consumer protection in the Digital Economy", European Data Protection Supervisor, March, https://secure.edps.europa.eu/EDPSWEB/edps/Home/Consultation/OpinionsC, accessed 24 May 2015.

EFF (2014), "Brief of a Amici Curiae Computer Scientists in Support of Petitioner", Electronic Frontier Foundation, 7 November.

EC (2013), "Study on business models for linked open government data", European Commission, http://ec.europa.eu/isa/documents/study-on-business-models-open-government_en.pdf, accessed 24 May 2015.

EC (2012), "Commission proposes a comprehensive reform of the data protection rules", Data Protection – Newsroom, European Commission, http://ec.europa.eu/justice/newsroom/data-protection/news/120125_en.htm, accessed 24 May 2015.

ESG (2012), "Boiling the ocean of control points in the Hadoop big data market", Enterprise Strategy Group, www.esg-global.com/blogs/boiling-the-ocean-of-control-points-in-the-hadoop-big-data-market/, accessed 24 May 2015.

Evans, D.S. (2011), "Antitrust Economics of Free", *Competition Policy International*, Spring 2011, http://ssrn.com/abstract=1813193.

Filistrucchi, L. et al. (2014), "Market definition in two-sided markets: Theory and practice", *Journal of Competition Law and Economics*, Vol. 10, No. 2, pp. 293-339.

Forbes (2013), "HTML5 vs. native mobile apps: Myths and misconceptions", 23 January, www.forbes.com/sites/ciocentral/2013/01/23/html5-vs-native-mobile-apps-myths-and-misconceptions/, accessed 24 May 2015.

FTC (2014), "Data brokers: A call for transparency and accountability", Federal Trade Commission, May, Washington, DC, https://www.ftc.gov/system/files/documents/reports/data-brokers-call-transparency-accountability-report-federal-trade-commission-may-2014/140527databrokerreport.pdf, accessed 24 May 2015.

FTC (2007), "Federal Trade Commission Closes Google/DoubleClick Investigation", 20 December, Federal Trade Commission, www.ftc.gov/news-events/press-

releases/2007/12/federal-trade-commission-closes-googledoubleclick-investigation, accessed 24 May 2015.

Gild (2014), "Intelligent sourcing", Gild, https://www.gild.com/score-candidates/, accessed 30 October 2014.

Ha, A. (2012), "DealAngel helps you find the best hotel deals — Not just the cheapest", *TechCrunch*, 17 April, http://techcrunch.com/2012/04/17/dealangel-helps-you-find-the-best-hotel-deals-not-just-the-cheapest/.

Harbour, P.J. and T. Koslov (2010), "Section 2 in a Web 2.0 world: An expanded vision of relevant product markets", *76 Antitrust J.L.*, pp. 769-94.

Harris, D. (2012), "Five low-profile start-ups that could change the face of big data", GigaOm, 28 January, http://gigaom.com/2012/01/28/5-low-profile-startups-that-could-change-the-face-of-big-data/.

Harris, D. (2011a), "As big data takes off, Hadoop wars begin", GigaOm, 25 March, http://gigaom.com/2011/03/25/as-big-data-takes-off-the-hadoop-wars-begin.

Harris, D. (2011b), "Why big data start-ups should take a narrow view", GigaOm, 28 March, http://gigaom.com/2011/03/28/why-big-data-startups-should-take-a-narrow-view/.

Hartmann, P. et al. (2014), "Big data for big business? A taxonomy of data-driven business models used by start-up firms", 27 March, Cambridge Service Alliance working paper, www.cambridgeservicealliance.org/uploads/downloadfiles/2014_March_Data%20Driven%20Business%20Models.pdf.

Heath, N. (2012), "Slow start for big data in Europe", *TechRepublic*, www.techrepublic.com/blog/european-technology/slow-start-for-big-data-in-europe/, accessed 24 May 2015.

Higginbotham, S. (2009), "Google gets shifty with its data center operations", GigaOm, 16 July, https://gigaom.com/2009/07/16/google-gets-shifty-with-its-data-center-operations/.

Kelly, J. (2013), "Hadoop pure-play business models explained", Wikibon, 17 December, http://wikibon.org/wiki/v/Hadoop_Pure-Play_Business_Models_Explained.

King, R. (2013), "Netflix brings Amazon web services closer", *The Wall Street Journal,* 28 July, http://blogs.wsj.com/cio/2013/07/28/netflix-brings-amazon-web-services-closer/.

Koehler, P., A. Anandasivam and M. Dan (2010), "Cloud services from a consumer perspective", *AMCIS 2010 Proceedings*, Paper 329, http://aisel.aisnet.org/amcis2010/329, accessed 24 May 2015.

Kommerskollegium (2014), "No transfer, no trade – The importance of cross-border data transfers for companies based in Sweden", January, www.kommers.se/Documents/dokumentarkiv/publikationer/2014/No_Transfer_No_Trade_webb.pdf, accessed 24 May 2015.

Koski, H. (2011), "Does marginal cost pricing of public sector information spur firm growth?", *ETLA Discussion Papers,* No. 1260, The Research Institute of the Finnish Economy, https://www.econstor.eu/dspace/bitstream/10419/87764/1/669255319.pdf, accessed 24 May 2015.

Lavalle, S. et al. (2010), "Analytics: The new path to value", *MIT Sloan Management Review*, p. 15, http://cci.uncc.edu/sites/cci.uncc.edu/files/media/pdf_files/MIT-SMR-IBM-Analytics-The-New-Path-to-Value-Fall-2010.pdf, accessed 24 May 2015.

Lomas, N. (2013), "Handshake is a personal data marketplace where users get paid to sell their own data", *Techcrunch*, 2 September, http://techcrunch.com/2013/09/02/handshake/.

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

Metz, C. (2010), "Google blesses Hadoop with MapReduce patent license", *The Register*, 27 April, www.theregister.co.uk/2010/04/27/google_licenses_mapreduce_patent_to_hadoop/.

Microsoft (2011), "Microsoft expands data platform with SQL Server 2012: New investments for managing any data, any size, anywhere", *Microsoft News Center*, 12 October, www.microsoft.com/en-us/news/press/2011/oct11/10-12PASS1PR.aspx.

Miller, P. (2013), "Visualization, the key that unlocks data's value?", http://cloudofdata.com/2013/04/visualisation-the-key-that-unlocks-datas-value, accessed 24 May 2015.

Moore, F. (1993), "Predators and prey: A new ecology of competition", *Harvard Business Review*, May-June, http://blogs.law.harvard.edu/jim/files/2010/04/Predators-and-Prey.pdf, accessed 24 May 2015.

Muenchen, R. (2014), "The popularity of data analysis software", r4stats.com, http://r4stats.com/articles/popularity/, accessed 14 November 2014.

Musil, Steven (2011), "Report: Twitter buys TweetDeck for $40 million", CNET News, May, http://news.cnet.com/8301-1023_3-20065533-93.html, accessed 23 June 2015.

NAICS (2002), US North American Industry Classification System 2002.

NESSI (2012), "Big Data: A New World of Opportunities", Networked European Software and Services Initiative, www.nessi-europe.com/Files/Private/NESSI_WhitePaper_BigData.pdf, accessed 25 May 2015.

NISO (2004), *Understanding Metadata*, NISO Press, National Information Standards Organization, www.niso.org/publications/press/UnderstandingMetadata.pdf, accessed 24 May 2015.

O'Brien, S.P. (2013), "Hadoop ecosystem as of January 2013 – Now an app!", Datameer, 15 January, www.datameer.com/blog/perspectives/hadoop-ecosystem-as-of-january-2013-now-an-app.html.

O'Dell, J. (2011), "In a world without tracking & cookies, can online commerce succeed?", mashable.com, 10 May, mashable.com/2011/05/10/buyosphere/.

OECD (2015), *Inquiries into Intellectual Property's Economic Impact?*, OECD Publishing, Paris, forthcoming.

OECD (2014a), "Connected televisions: Convergence and emerging business models", *OECD Digital Economy Papers*, No. 231, OECD Publishing, Paris, http://dx.doi.org/10.1787/5jzb36wjqkvg-en.

OECD (2014b), "Cloud computing: The concept, impacts and the role of government policy", *OECD Digital Economy Papers*, No. 240, OECD Publishing, Paris, http://dx.doi.org/10.1787/5jxzf4lcc7f5-en.

OECD (2014c), "International cables, gateways, backhaul and international exchange points", *OECD Digital Economy Papers*, No. 232, OECD Publishing, Paris, http://dx.doi.org/10.1787/5jz8m9jf3wkl-en.

OECD (2014d), *Measuring the Digital Economy: A New Perspective*, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264221796-en.

OECD (2014e), *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), "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 (2013b), "Exploring data-driven innovation as a new source of growth: Mapping the policy issues raised by 'big data'", in OECD, *Supporting Investment in Knowledge Capital, Growth and Innovation*, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264193307-12-en.

OECD (2013c), *Recommendation of the Council concerning Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data*, OECD Publishing, Paris.

OECD (2013d), "The app economy", *OECD Digital Economy Papers*, No. 230, OECD Publishing, Paris, http://dx.doi.org/10.1787/5k3ttftlv95k-en.

OECD (2012a), *OECD Internet Economy Outlook 2012*, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264086463-en.

OECD (2012b), "Market definition: Competition committee policy roundtable", DAF/COMP(2012)19, OECD Publishing, Paris, 11 October, www.oecd.org/daf/competition/Marketdefinition2012.pdf, accessed 24 May 2015.

OECD (2011a), "Internet intermediaries: Definitions, economic models and role in the value chain", in *The Role of Internet Intermediaries in Advancing Public Policy Objectives*, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264115644-4-en.

OECD (2011b), *Recommendation of the Council on Principles for Internet Policy Making,* OECD Publishing, Paris, www.oecd.org/daf/competition/44445730.pdf, accessed 24 May 2015.

OECD (2009), "Two-sided markets: Competition committee policy roundtable", DAF/COMP(2009)20, OECD Publishing, Paris, 17 December, www.oecd.org/daf/competition/Marketdefinition2012.pdf.

OECD (2008), *Recommendation of the Council for Enhanced Access and More Effective Use of Public Sector Information,* OECD Publishing, Paris, www.oecd.org/internet/ieconomy/40826024.pdf, accessed 24 May 2015.

OECD-NSF, (2011) "OECD-NSF Workshop: Building a smarter health and wellness future", Summary of key messages, 15-16 February, internal working document.

OFT (2004), *Assessment of Market Power: Understanding Competition Law*, Competition Law Guideline, Office of Fair Trading, www.gov.uk/government/uploads/system/uploads/attachment_data/file/284422/oft402.pdf, accessed 24 May 2015.

Orrick (2012), *The Big Data Report*, Orrick, http://www.slideshare.net/CBInsights/big-data-report-31586014, accessed 25 May 2015.

Paul, R. (2010), "Google's MapReduce patent: What does it mean for Hadoop?", Arstechnica.com, 20 January, http://arstechnica.com/information-technology/2010/01/googles-mapreduce-patent-what-does-it-mean-for-hadoop/, accessed 24 May 2015.

Pingdom (2013), "The top 100 web hosting countries", 14 March, http://royal.pingdom.com/2013/03/14/web-hosting-countries-2013, accessed 20 April 2014.

Rao, L. (2013), "As software eats the world, non-tech corporations are eating start-ups", *TechCrunch*, http://techcrunch.com/2013/12/14/as-software-eats-the-world-non-tech-corporations-are-eating-startups/, accessed 24 May 2015.

Rao, L. (2011), "Index and Khosla lead $11m round in Kaggle, a platform for data modeling competitions", *TechCrunch*, 2 November, http://techcrunch.com/2011/11/02/index-and-khosla-lead-11m-round-in-kaggle-a-platform-for-data-modeling-competitions/, accessed 24 May 2015.

Redman, T. (2008), *Data Driven*, Harvard Business School Publishing, Boston, p. 25.

Reuters (2011), "*Financial Times* pulls its apps from Apple store", 31 August, www.reuters.com/article/2011/08/31/us-apple-ft-idUSTRE77U1O020110831, accessed 24 May 2015.

Ries, E. (2011), *The Lean Startup: How Today's Entrepreneurs Use Continuous Innovation to Create Radically Successful Businesses*, 13 September, First edition, Crown Business.

Russom, P. (2011) "Big data analytics", TDWI Best Practices Report, p. 24, http://tdwi.org/research/2011/09/best-practices-report-q4-big-data-analytics.aspx, accessed 24 May 2015.

Schubert, L., K. Jefferey, and B. Neidecker-Lutz (2010), "The future of cloud computing: Opportunities for European cloud computing beyond 2010", public version 1.0, http://cordis.europa.eu/fp7/ict/ssai/docs/cloud-report-final.pdf, accessed 22 June 2014.

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

Stiglitz, J., P. Orszag and J. Orszag (2000), "Role of government in a digital age", Computer and Communications Industry Association, October.

Strategy Analytics (2011), "One billion HTML5 phones to be sold worldwide in 2013", 7 December, www.strategyanalytics.com/default.aspx?mod=pressreleaseviewer&a0=5145.

Suster, M. (2010), "Social networking: The future", TechCrunch, 5 December, http://techcrunch.com/2010/12/05/social-networking-future/.

Swire, P. and Y. Lagos (2013), "Why the right to data portability likely reduces consumer welfare: Antitrust and privacy critique", Maryland Law Review, Vol. 72.2, http://digitalcommons.law.umaryland.edu/mlr/vol72/iss2/1, accessed 24 May 2015.

Telefónica (2012), "Telefónica launches Telefónica dynamic insights – A new global big data business unit", Press release, Telefónica, 9 October, http://blog.digital.telefonica.com/?press-release=telefonica-launches-telefonica-dynamic-insights-a-new-global-big-data-business-unit.

The Economist (2014), "Digital disruption on the farm", 24 May, www.economist.com/news/business/21602757-managers-most-traditional-industries-distrust-promising-new-technology-digital.

The Economist (2012), "Know thyself", 15 December, www.economist.com/news/business/21568438-data-lockers-promise-help-people-profit-their-personal-information-know-thyself.

The Economist (2011), "Incentive prizes: Healthy competition", 10 April, www.economist.com/blogs/babbage/2011/04/incentive_prizes.

TNO (2013), "Thriving and surviving in a data-driven society", TNO, 24 September, http://publications.tno.nl/publication/34610048/xcv74S/TNO-2013-R11427.pdf.

Twitter (2012), "Changes coming in Version 1.1 of the Twitter API", 16 August, https://blog.twitter.com/2012/changes-coming-to-twitter-api, accessed 23 June 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.

Van der Berg, R. (2014), "The connected television debate in OECD countries", OECD Insights blog, OECD Publishing, Paris, 23 January, http://oecdinsights.org/2014/01/23/the-connected-television-debate-in-oecd-countries/.

van Veenstra, A.F. and T.A. van den Broek (2013), "Opening moves – Drivers, enablers and barriers in open data in a semi-public organization", in *M.A. Wimmer, M. Janssen and H.J. Scholl (eds.)*, EGOV 2013, LNCS 8074, pp. 50-61.

Watters, A. (2011a), "An iTunes model for data", O'Reilly Strata, http://strata.oreilly.com/2011/04/itunes-for-data.html. **[accessed when?]**

Watters, A. (2011b), "Scraping, cleaning and selling big data", O'Reilly Strata, http://strata.oreilly.com/2011/05/data-scraping-infochimps.html, accessed 24 May 2015.

Wireless Week (2011), "Native apps vs. HTML5-based apps: What does the future hold?", 7 September, www.wirelessweek.com/articles/2011/09/native-apps-vs-html5-based-apps-what-does-future-hold.

Woo, B. (2013), "A mind blowing big data experience: Notes from Strata", Forbes, 27 February, www.forbes.com/sites/bwoo/2013/02/27/a-mind-blowing-big-data-experience-notes-from-strata-2013/.

Yahoo News, (2013), "Twitter's new policies kill three more apps", 7 March.

Further reading

Brave, S. (2012), "We don't need more data scientists – Just make big data easier to use", GigaOm, https://gigaom.com/2012/12/22/we-dont-need-more-data-scientists-just-simpler-ways-to-use-big-data/, accessed 25 May 2015.

OECD (2013e), "The digital economy: Competition committee hearings", DAF/COMP(2012)22, OECD Publishing, Paris, 7 February, www.oecd.org/daf/competition/The-Digital-Economy-2012.pdf.

OECD (2013f), "Competition policy and knowledge-based capital", in OECD, *Supporting Investment in Knowledge Capital, Growth and Innovation*, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264193307-7-en.

OECD (2013g), *Supporting Investment in Knowledge Capital, Growth and Innovation*, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264193307-en.

OECD (2004), "518111 Internet Service Providers", Glossary of Terms, OECD Publishing, Paris.

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.

Taylor, R. (2012), "The Hadoop ecosystem, visualized in Datameer", Datameer blog, www.datameer.com/blog/uncategorized/the-hadoop-ecosystem-visualized-in-datameer.html, accessed 25 May 2015.

From:

Data-Driven Innovation

Big Data for Growth and Well-Being

Access the complete publication at:
https://doi.org/10.1787/9789264229358-en

Please cite this chapter as:

OECD (2015), "Mapping the global data ecosystem and its points of control", in *Data-Driven Innovation: Big Data for Growth and Well-Being*, OECD Publishing, Paris.

DOI: https://doi.org/10.1787/9789264229358-6-en