For Official Use **DSTI/ICCP(2013)13**

Organisation de Coopération et de Développement Économiques
Organisation for Economic Co-operation and Development **27-Nov-2013**

English - Or. English

DIRECTORATE FOR SCIENCE, TECHNOLOGY AND INDUSTRY
COMMITTEE FOR INFORMATION, COMPUTER AND COMMUNICATIONS POLICY

DSTI/ICCP(2013)13
For Official Use

Cancels & replaces the same document of 22 November 2013

NEW SOURCES OF GROWTH: KNOWLEDGE-BASED CAPITAL (II): DATA PILLAR

Introduction to Data and Analytics (Module 1): Taxonomy, Data Governance Issues and Implications for further Work

9-13 December 2013

This paper contributes to the first module of the Data Pillar of the OECD project on New Sources of Growth: Knowledge-Based Capital II (KBC2:DATA).

The paper develops the foundation for KBC2:DATA by introducing the basic concepts and a data taxonomy. It also highlights the key challenges related to data governance that should be considered by governments, but also businesses, when designing data-related policies. It then concludes with the implications for further work under the KBC2:DATA project.

The paper was circulated for consultation to the OECD Secretariat members respectively in charge of all involved Committees including (besides the ICCP) the Committee on Consumer Policy (CCP), the Committee for Scientific and Technological Policy (CSTP), the Health Committee (HC), and the Public Governance Committee (PGC).

Delegates will be invited to discuss this paper and in particular to provide feedback on whether the proposed list of key data governance issues is complete.

Christian Reimsbach-Kounatze; christian.reimsbach-kounatze@oecd.org; Tel.: +33 1 45 24 76 16
Anne Carblanc; anne.carblanc@oecd.org; Tel.: +33 1 45 24 93 34

JT03349379

English - Or. English

TABLE OF CONTENTS

SUMMARY

1. This paper contributes to Phase II of the OECD project on *New Sources of Growth: Knowledge-Based Capital*, in particular to its Pillar on data (KBC2: DATA). The paper introduces the basic concepts and a taxonomy for data-related policies which will provide the basis for KBC2: DATA. It also highlights the key challenges related to data governance that should be considered when designing data-related policies, and then discusses the implications for KBC2: DATA.

2. The paper starts with an analysis of the concepts of data, information, and knowledge. This analysis helps to clarify the mechanisms leading to the transformation of data into information (through data analytics), and information into knowledge (through learning). The paper then analyses and defines **key data categories, which form the data taxonomy** highlighted in the Glossary below. They include: (i) public vs. private sector data, (ii) open vs. closed data, (iii) personal vs. non-personal data, (iv) volunteered vs. observed vs. inferred data, (v) user created vs. machine generated data, (vi) micro- vs. sectorial vs. macro data, (vii) structured vs. unstructured data, (viii) linked data vs. data silos, (ix) meta-data vs. primary data, and (x) real- or near-time vs. static data. The paper then identifies **key challenges related to data governance** including: (i) data value and pricing, (ii) data access and sharing, (iii) data quality and curation, (iv) data ownership and control, and (v) data privacy and security.

3. One of the key conclusions of the analysis in this paper is that assessing **the value of information and data** *ex ante* (before its use) is almost impossible, because **information is context dependent**. As a result, the value of data is not only a function of the data itself, but also a function of the (analytic) capacity of the data controller to link the data to other data sets and to extract insights. This capacity is not only determined by available analytic techniques and technologies. More importantly, it is a function of pre-existing knowledge and skills. This means that there are a number of factors beyond the data itself which determine data-driven value creation.

4. **As with other intangible assets, data has a non-rivalrous nature**. The use of data does not affect in principle its potential to meet the demands of others. In contrast to physical goods or resources, such as oil, which is depleted once extracted, transformed and burned during production processes, data can in principle be used and re-used infinitely at (almost) no marginal costs. Data can thus ease the constraints placed on growth by scarcity of other physical capital goods and resources. **What is scarce, however, is the capacity to use data in meaningful ways** as highlighted above.

5. **These two properties (context-dependency and non-rivalry) have profound implications for growth and well-being, and data-related policies such as affecting privacy and security as well as ownership and control.** These implications are pertinent for the KBC2: DATA project and are as follows:

i. A holistic view is needed to understand data-driven value creation. Since data has no intrinsic value, the information extracted from the data, the downstream creation of knowledge, and its use for decision-making all matter. Understanding the whole process from data generation and collection to data analysis and data-driven decision-making is therefore crucial. This **calls for a data value chain or data value cycle approach**.

ii. The non-rivalrous nature of data suggests that **non-discriminatory access to data (i.e. open data) can maximize the economic and social value of data**, leading to increasing (social) returns to scale. Since data is a non-rivalrous good, maximizing its value can be achieved through promoting a maximum number of uses though open data. At the same time the scope of legal regimes and policies promoting access to data (e.g. PSI frameworks) has decreased with the shift

from public to private sector data with the privatisation of data-intensive sectors including, but not limited to, telecommunication services, financial services, transportation, and utilities.

iii. **Open data can be a sustainable private and public strategy for value creation in complex environments characterized by high uncertainty and for open innovation** as discussed in the OECD (2010) *Innovation Strategy* and the OECD (2013d) project on *Knowledge Networks and Markets*. In particular, open data can: (i) facilitate joint production or co-operation with suppliers, customers or even competitors, (ii) support and encourage user-driven innovation including value-creating activities by users (incl. consumers and citizens), (iii) maximize the option value of data when data investments are irreversible and there is high uncertainty regarding sources of future market value, and, last but not least, (iv) effectively (cross-) subsidise the production of public and social goods without the need to rely on either the market or the government to 'pick winners'.

iv. **Openness is not a binary quality, but a continuum** ranging from *closed data* (access only by the data controller), to (i) restricted access to individual data stakeholders who can affect or are affected by the use of the data, to (ii) non-discriminatory access granted to members of specific communities (see OECD *Recommendation on Principles and Guidelines for Access to Research Data from Public Funding* focussing on the "international research community"), and last but not least, to (iii) non-discriminatory access to the public. Various factors affect the degree of openness, including legal and cultural issues some of which typically relate to confidentiality and privacy issues. But three **key factors affect the degree of openness in particular**: (i) technological design (incl. data availability on the web, machine readability, and linkability), (ii) intellectual property rights (IPRs) (incl. legal regimes such as copyright as well as other IPRs applicable to databases and trade secrets), and (iii) pricing.

v. **Among the three key factors affecting openness, data pricing schemes may be the most complex** because the value of data is context dependent. Pricing schemes based on the cost structure, including marginal and average cost pricing, are an often recommended approach. The OECD (2005) *Recommendation on Principles and Guidelines for Access to Research Data from Public Funding* and the OECD (2008a) *Council Recommendation on Enhanced Access and More Effective Use of Public Sector Information* (PSI) encourage marginal cost pricing. Data pricing schemes are often part of more complex revenue models including subscription fees, freemium[1], voluntary donations; sometimes combined with (cross-) subsidies.

vi. **Context dependency** also **explains why data linkage (*linked data*) enables "super-additive" insights**, leading to increasing 'returns to scope'. Linked data is a means to contextualise data and thus a source for insights and value that are greater than the sum of its isolated parts (*data silos*). The "super-additive" nature of linked data, **however,** is also the source for additional challenges. In particular**, linked data sets can undermine confidentiality and privacy protection** measures such as anonymisation and pseudonymisation.

vii. **Linked data** thus **challenges the concept of "personal data"**, which is defined by the OECD (2013c) *Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data* as "any information relating to an identified or identifiable individual (data subject)". This is because seemingly non-personal data can be related to an identified or identifiable individual thanks to data linkage, suddenly "transforming" non-personal into personal data. To classify data as "personal" *ex ante* thus assumes that personal information can always be extracted. Equivalently, the reverse is assumed for "non-personal" data. These assumptions, however, may not be valid in some cases where data sets are linked. This make (personal) data a weak point of

control for privacy protection and calls for the identification of, and a focus on, alternative points of control for privacy protection.

viii. **Unlinkability is both a protection from and an obstacle to the extraction of information, including confidential and personal information**. Various means to achieve unlinkability include data collection limitation, the distribution and separation of information systems as well as cryptography, mis- /dis-information, and noise addition techniques related to the most identifying data attributes. Reversely, in an era of big data, where the cost for data collection and storage has declined to negligible level, data collection limitation is challenged.

ix. **Context dependency also points to a number of other attributes affecting the quality and value of data**. These attributes depend on the use of the data and include: (i) relevance, (ii) accuracy, (iii) credibility, (iv) timeliness, (v) accessibility, (vi) interpretability, (vii) coherence, as well as (viii) completeness of the data as defined by the OECD (2011) *Quality Framework and Guidelines for OECD Statistical Activities* and the data quality principle of the OECD (2013c) *Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data* respectively.

x. **Open and linked data challenge the applicability of concepts such as data ownership**, which is already an often misunderstood and/or misused concept. In organisations, data ownership is often used in the sense of "data stewardship" where data users must consider the consequences of making changes over 'their' data. In contrast to other intangibles that can be protected through IPRs, **data typically involve complex assignments of different rights across different data stakeholders**; making it challenging to assign exclusive IPRs to a single data stakeholder. The situation is even more complex in cases where "personal data" is involved, since certain control rights of the data subject cannot be taken away, such as stated in the Individual Participation Principle of the OECD (2013c) *Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data*.

6. **Increasing human efforts is being dedicated to make machines (empowered by software) "understand" the world's growing volume of data and to enhance the capacity to externalise knowledge** thanks to progress in data collection and analytics. This can lead to data-driven automation, where machines can perform tasks that are more knowledge and labour intensive. The proliferation of **data-driven automation can boost productivity growth, but also has potential implications on employment and income inequalities**. Some of these implications relate to *capital-biased technological change*, which tends to shift the distribution of income away from workers to the owners of capital.

7. Finally, **"big data" as simply understood as 'data with huge volume' is not the only source for data-driven innovation. It is the exploitation and combination of data sources in the form of open data, linked data, (un-) structured data and real-time data, to name a few, that is at the core of data-driven innovation**. If big data is perceived as the umbrella concept covering and unifying all these data concepts as suggested for example by the three Vs (volume, velocity and variety) definition, then big data can indeed be perceived as the major source for data-driven innovation together with data analytics as the key enablers.

INTRODUCTION TO DATA AND ANALYTICS

TAXONOMY, DATA GOVERNANCE ISSUES, AND IMPLICATIONS FOR FURTHER WORK

Where is the Life we have lost in living?
Where is the wisdom we have lost in knowledge?
Where is the knowledge we have lost in information?

Eliot, T. S. (1934), *The Rock*, Faber & Faber

Introduction

8. Data has increasingly become an important source for value creation and for the formation of knowledge-based capital (KBC)[2]. More and more organisations collect, store, and process data today to expand their future production capacities (OECD, 2013b). The productivity improvements are dramatic, and striking. Utilities, for example, increasingly use smart meter data to identify overall consumption patterns in order to forecast future demand and to automatically adjust production capacities and pricing mechanisms to this demand. Governments can increase the transparency of their processes and performance by publishing relevant data online and sharing it with the public. Other popular examples include Internet firms such as Google, which collects (crawls) web content and other related data to feed and improve their web-based and mobile services including search engines, personal navigation systems, and recommendation systems.

9. The OECD horizontal project on *New Sources of Growth: Knowledge-Based Capital (KBC)* assesses this new role of data as a driver of value creation and a new source for economic growth and well-being.[3] Assessing this role, however, requires a solid understanding of data including the different types of data, their economic properties, and the specific issues associated to data. As the use of data becomes an increasingly important economic and social phenomenon, economists and policy analysts are trying to capture the phenomenon with existing concepts and theories. Metaphors like: "data is the new currency" (Schwartz, 2000 cited in IPC, 2000; Zax, 2011; Dumbill, 2011; Deloitte, 2013) or more recently "data is the new oil" (Kroes, 2012; Rotella, 2012; Arthur, 2013) are often used as rhetorical means to make this emerging phenomenon better understandable to policy and decision makers. Although at first helpful to highlight the (new) economic value of data, these metaphors often fall short and sometimes are even misleading, and therefore should be used with caution (see for example Thorp, 2012; Bracy, 2013; and Glanz, 2013). For example, data is *not a rivalrous good*, nor is it a primary resource, such as oil, which is depleted once extracted, transformed and burned during production processes. In contrast to oil, the use of data does not affect in principle its potential to meet the demands of others.[4]

10. Assessing the role of data for economic growth and well-being also calls for an understanding of the theoretical and conceptual links between data, information and knowledge. This is because, as highlighted in OECD (2013b) the economic value of data lies in its potential to generate knowledge and to enrich knowledge-intensive ("smart") applications. Furthermore, as concepts such as "knowledge economy", "information economy", and now "data-driven economy" are often used in media and literature, though in most cases without clear definitions or explanations of their interrelationships, it becomes even more necessary to define the three underlying concepts "data", "information", and "knowledge", and to clarify the theoretical and conceptual links between them.[5] This exercise promises to

turn helpful when, for example, assessing the relationship between *public sector information* (PSI) as addressed by the *OECD (2008a) Recommendation of the Council for Enhanced Access and More Effective Use of Public Sector Information* and "open data" as promoted through various *open government data* (OGD) initiatives.

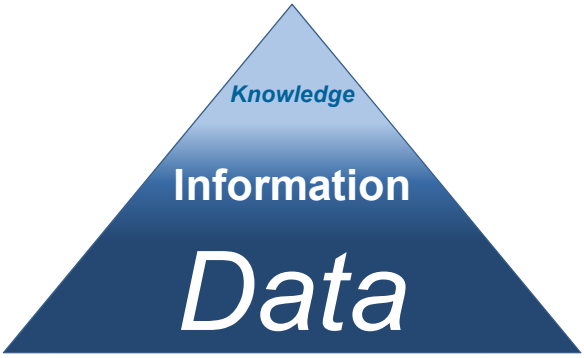
11. This paper contributes to Phase II of the OECD project on *New Sources of Growth: Knowledge-Based Capital*, in particular to the first module of its Pillar on data (KBC2: DATA). The paper introduces the basic concepts "knowledge", "information", and "data" and presents a taxonomy that distinguishes among and defines a number of data categories, many of which are at the core of KBC2:DATA. Such a differentiation is needed as different data categories may have different implications for individuals, businesses, and policy makers. For example, whether data is classified as personal or not has implications on the applicability of privacy legislation; whether data classifies as "open data" or not comes with a number of implications on competition, consumer protection and empowerment, and can affect the applicability of certain intellectual property rights (IPRs). Furthermore, the paper present key challenges related to data governance that should be considered by governments, as well as businesses, when designing data-related policies. Finally, the paper ends with some concluding remarks with implications for the following working paper.

From data to information to knowledge and back

12. Data, information and knowledge are different but interrelated concepts. Information is often conveyed through data, while knowledge is typically gained through the assimilation of information. The boundaries between data, information and knowledge may seem fuzzy, which explains why these concepts are often used as synonyms in media and literature.[6] However, as will be highlighted in this paper, separating these concepts is important to better understand data-driven value creation and the policy implications. A clearer distinction can, for example, help understand why one can have a lot of data, but not be able to extract value from it when not equipped with the appropriate analytic capacities (OECD, 2013b; Ubaldi, 2013). This is for example the case with "unstructured data" such as web logs and media content, for which information extraction tends to be more costly compared to "structured" data.[7] Similarly, one can have a lot of information, but not be able to gain knowledge from it. The last line of the citation by Eliot (1934)'s *The Rock* presented at the beginning of this paper ("Where is the knowledge we have lost in information?") hints at a phenomenon nowadays better known as "information overload" and which Nobel prize-winning economist Herbert Simon described with the words: "a wealth of information creates a poverty of attention" (Shapiro and Varian; 1999).[8] The fact that information overload has become a well-known management issue for modern organisations, while we hear "nobody ever complaining about 'knowledge overload'" (Hey, 2004), points to a difference between information and knowledge.[9]

13. The following sections present the concepts of "knowledge", "information", and "data" in more detail and how they are interrelated. It is based on discussions in disciplines such as computer science and cybernetics, but most importantly information science and knowledge management (see Hoshovsky and Massey, 1968 cited in Hawkins, 2001; Ackoff, 1989 cited in Hey, 2004; Cleveland, 1982; Lucky, 1989; Luhmann, 1996; Hjørland, 2000; Alavi and Leidner, 2011; Zins, 2007b).[10] In these disciplines, data, information, and knowledge are often regarded as parts of a sequential, sometimes even hierarchical order, with data being an input for information, and information an input for knowledge.[11] The hierarchical order is often illustrated in the form of the "information pyramid" presented in Figure 1.[12] This pyramid does not only reflect the notion of sequential order for the creation of knowledge from information and data, but it also reflects the notion that this upstream transformation from data to information to knowledge can be seen as a "distillation" process. As Hey (2007) explains: This "implies that a large amount of data can be turned into a smaller 'conceptual amount' of information". For example, it can take a significant amount of data about a phenomenon such as company's sales to finally get to the information that the company has increased its annual revenue by *x* percentage points in year *t*.

Figure 1. The information pyramid



Note: The information pyramid is sometime also known as the knowledge pyramid or the DIKW (Data, Information, Knowledge, Wisdom) pyramid in the knowledge management domain, often with an additional fourth step or concept "wisdom".

Knowledge

14. No single agreed upon definition of knowledge exists, although there are numerous theories that try to explain it.[13] In economics, knowledge is for example recognized as central to innovation, economic growth and development. According to new growth theory, "knowledge can raise the returns on investment, which can in turn contribute to the accumulation of knowledge" (OECD, 1996). This is done by enabling innovation including more efficient methods of organising production as well as new and improved goods and services. Furthermore, knowledge can easily spill over from one organisation or industry to another, as new ideas can be used repeatedly at little or no marginal costs.[14] These spillovers can than ease the constraints placed on growth by scarcity of other capital goods and resources.

15. Most theories agree that knowledge is embodied in human beings. According to economists, for example, knowledge is embodied in the form of *"human capital"*. Knowledge is often defined as "what someone knows"[15] including his or her belief system. It is "what is understood"[16] and "grasped"[17] by an individual through the "internalisation"[18], "assimilation"[19], or "appropriation"[20] of information and experience "in the process of learning, acting, interpreting"[21]. This internalisation or assimilation process modifies individuals' mental state or "mental stock of information"[22], including "their beliefs, values, procedures, actions, etc."[23] (Zins, 2007a). Furthermore, most definitions highlight that knowledge has an ultimate goal, namely to be used for decision-making.[24] In this respect, knowledge is considered to be subjective, a "human construct"[25] that does not have an existence outside the human "mind"[26], "brain"[27] and/or "body" (Hey, 2004; Le Coadic, 2004; Zins, 2007a). According to some, this does not exclude knowledge from existing at a social level as well, in a kind of "collective memory"[28]. As Thomas A. Childers explains (cited in Zins, 2007a): "Knowledge […] is by definition subjective, even when aggregated to the level of social, or public, knowledge – which is the sum, in a sense, of individual 'knowings'"[29]. At the collective level, knowledge is often "embedded in particular social environments such as academic disciplines, practitioner communities, or expert groups" (Oborn, et al., 2010 cited in Fazekas and Burns, 2012).

16. Most scientists also agree that knowledge can be externalised and embedded in tangible and intangible products, including books, standard procedures and, last but not least, KBCs such as patents and software.[30] To what degree, however, knowledge can be externalised depends on whether it is explicit or implicit (tacit):

- **Explicit knowledge** is knowledge that can be codified and thus cost effectively transformed into information (*i.e.* externalisation). This externalised knowledge can then be stored, communicated about and embedded in tangible and intangible products, which then can be the object of economic transactions. It is this kind of knowledge that Nonaka and Takeuchi (1995) refer to as "the stock" (while information is considered the flow) (Hey, 2007). And it is also here where the boundaries between knowledge and information are mostly blurred (see "know-what" and "know-why" in Box 1).

- In contrast, **tacit knowledge** is knowledge that is too hard to codify, formalise and communicate (see Polanyi, 1962; 1967; von Hippel 1988; Nonaka and Takeuchi, 1995). "It is *ephemeral* and *transitory* [and thus] cannot be *resolved into* information or *itemized* in the manner characteristic of information" (Hey, 2004). It emerges through experience combined with internalized information that is "meaningfully integrated into a unifying frame of experience among other information […] interiorized in the same way, the complex of which reflects subjective understanding of the environment" (Michael Lorenz cited in Zins, 2007a).

Box 1. Different kinds of knowledge: know-what, know-why, know-how and know-who

In order to facilitate economic analysis, OECD (1996) distinguishes between the following different kinds of knowledge which are important in the knowledge-based economy:

- **Know-what** refers to knowledge about "facts". How many people live in New York? What are the ingredients in pancakes? And when was the battle of Waterloo? These are examples of this kind of knowledge. Here, knowledge is close to what is normally called information – it can be broken down into bits. In some complex areas, experts must have a lot of this kind of knowledge in order to fulfill their jobs. Practitioners of law and medicine belong to this category.

- **Know-why** refers to scientific knowledge of the principles and laws of nature. This kind of knowledge underlies technological development and product and process advances in most industries. The production and reproduction of know-why is often organised in specialised organisations, such as research laboratories and universities. To get access to this kind of knowledge, firms have to interact with these organisations either through recruiting scientifically-trained labour or directly through contacts and joint activities.

- **Know-how** refers to skills or the capability to do something. Businessmen judging market prospects for a new product or a personnel manager selecting and training staff have to use their know-how. The same is true for the skilled worker operating complicated machine tools. Know-how is typically a kind of knowledge developed and kept within the border of an individual firm. One of the most important reasons for the formation of industrial networks is the need for firms to be able to share and combine elements of know-how.

- This is why **know-who** becomes increasingly important. Know-who involves information about who knows what and who knows how to do what. It involves the formation of special social relationships which make it possible to get access to experts and use their knowledge efficiently. It is significant in economies where skills are widely dispersed because of a highly developed division of labour among organisations and experts. For the modern manager and organisation, it is important to use this kind of knowledge in response to the acceleration in the rate of change. The know-who kind of knowledge is internal to the organisation to a higher degree than any other kind of knowledge.

Source: OECD (1996)

Information

17. Information is a key input for the formation of knowledge. As is the case for knowledge, no single agreed upon definition exists for information.[31] In information science, information is often seen as "meaning"[32] or "meaningful content"[33] that is exchanged in a communication process between a sender and a receiver (Zins, 2007a). This communication process is motivated by the "intension"[34] of the sender to create knowledge or to "modify the knowledge state of [the] interpreter or receiver"[35]. This is coherent with the origin of the word "information", which is the Latin verb "informare", meaning "to put into form" or "to give form to" (the mind or ideas).

18. In this respect, information requires, at "the end points"[36] of communication, intelligence enabling the receiver to extract, "understand"[37], and "interpret"[38] the meaning of the communicated "message"[39]. In most cases, human beings are therefore assumed to be at these end points.[40] However, significant efforts are being made to make "smart" machines better identify and "interpret"[41] relevant information in order to increase their capacity to autonomously process information (Box 2). Many of these efforts are based on concepts and techniques such as semantic networks (Sowa, 1992), pattern classification (Duda et al. 2000; Russel and Norvig, 2009), but also statistics (Hastie et al., 2011; James et al., 2013). These concepts and techniques suggest that information is not only reflected in its human understanding, but also in "structure"[42] or "organisation"[43] as embedded in the underlying data. This view is consistent with definitions according to which information is not only "data with meaning"[44] or "interpreted data", but also "structured data" (see Cleveland, 1982; Zins, 2007a).

Box 2. Making machines better "understand" Internet content: The Semantic Web

Modern history of information is significantly characterized by human efforts to make machines (empowered by software) "understand" the world's growing volume of information. The notion of "understanding" should be taken as a metaphor here, since there are still a number of limitations in computing, including in the field of artificial intelligence (AI), which makes it very difficult to ascribe any human-like capacity of understanding to machines (Alesso and Smith, 2008; Sopek, 2011). However, one cannot deny that software-empowered applications are becoming "smarter", in the sense that they have improving abilities to learn, to autonomously assimilate information, and as a result, to be more user-oriented, and to solve more complex, dynamic problems autonomously.

The "Semantic Web" is an example of human efforts to make machines better "understand" information, in this case information as stored on the Web. The concept of the Semantic Web was presented by Berners-Lee et al. (2001). The main idea behind the concept is "letting software use the vast collective genius embedded in [the web including] its published pages" (Schwarz, 2013). Web applications that interact with each other or with databases via application programming interfaces (APIs) require a certain level of "understanding" of the information and services that they provide and require. Examples include intelligent personal assistant software that can autonomously identify and process event information such as on flight, hotel and restaurant reservations.[45]

Open web standards such as the eXtensible Markup Language (XML) and the Resource Description Framework (RDF) are key for this machine "understanding" of web content. While XML allows individual programmers to define a structure for content, it does not specify what the structure means. Developed to extend XML, RDF provides some rudimentary semantic capabilities so to make work easier for autonomous smart applications to better "understand" information (Alesso and Smith, 2008).

19. The information relevant structure can be explicit such as in the case of SQL (Structured Query Language) databases, where the structure is reflected in the tables and their predefined interrelationships. However, in many cases, the relevant structure in the data may be implicit and thus more costly to extract. This is, for example, the case with semi-structured data such as log files and unstructured data such as multimedia data, which by definition lack an explicit data model. In these cases, information is extracted, when the data is "mined", processed, or analysed with the goal to extract the information from data, the signal from the noise (Silver, 2012). The set of techniques and tools required to extract information is referred to as *data analytics* (Box 3).

Box 3. Data analytics: extracting information from data

Data analytics encompasses a set of techniques and tools used to extract information from data. It does so by revealing the "context"[46] in which the data is embedded (Hey, 2004; Zins, 2007a), its organization and structure, and thus its "manifold hidden relations (patterns), *e.g.* correlations among facts, interactions among entities, relations among concepts" (Merelli and Rasetti, 2013).[47] There are a number of terms that are used (as synonyms) to refer to data analytics. Some of these terms may include aspects that go beyond data analysis. They are therefore presented in more detail below:

Data mining refers to a set of techniques used to extract information patterns from datasets. It is often said to go beyond data analytics as it combines data analytic methods such as statistics and machine learning (e.g. cluster analysis, classification, and regression), with data management methods (e.g. SQL databases, distributed data management with tools such as Hadoop[48]), data pre-processing (data cleaning), as well as model and inference considerations. The key aspect here, however, is the discovery of information patterns. Data mining is thus often used as a synonym for another term used more frequently in the past, namely **knowledge discovery**.

Profiling refers to the use of data analytics for the construction of *profiles* and the classification of entities in specific profiles, both based on the attributes of these entities. The term "profiling" is often used in cases where the profiled entities are individuals from which personal identifying information (PII) have been collection. Credit scoring, price discrimination, and targeted advertisement are therefore typical activities that are cited as examples involving profiling. But the term "profiling" can also be used where non-personal related entities are being profiled (e.g. malware activities).

Business intelligence (BI) was a term coined by Luhn (1958). There BI is defined as "the ability to apprehend the interrelationships of presented facts in such a way as to guide action towards a desired goal" (Luhn, 1958).[49] Today, BI refers to tools and techniques used to process data that have been previously stored in a database or data warehouse. The objective in BI is the creation of standard reports on a periodic basis, or the display of real-time business-related information on "management dashboards" highlighting key operation metrics for business management. So in contrast to data mining, BI typically focusses on business data as stored in databases for business reporting and monitoring. However, the boundaries between BI and data mining are blurring as BI software vendors are increasingly offering products and services covering BI as well as data mining.

Machine or statistical learning (ML) is a subfield in computer science, and more specifically artificial intelligence (AI). It is concerned with the design, development and use of algorithms that allow computers to "learn", that is to perform certain tasks, while improving performance with every empirical data set it analyses. ML involves activities such as pattern classification, cluster analysis, and regression (Duda et al. 2000; Russel and Norvig, 2009, Hastie et al., 2011; James et al., 2013).

Visual analytics refers to techniques and tools used for data visualization. They are used for gaining insights at a glance including through interactive data exploration. They are also used for communicating these insights to others (Unwin et al., 2006; Janert, 2010).

20. The two aspects of information highlighted above, (i) information as "data with meaning" or "interpreted data", and (ii) information as "data with structure", point to an important property of information, namely that *information is always context dependent*:

1. Information depends on the meaning as extracted or interpreted by the receiver. The same data sets can thus lead to different information depending on the analytic capacities of the "receiver" including her or his skills and (pre-) knowledge, available techniques and technologies for data analysis.

2. Information depends on how the underlying data is organized and structured. In other words, the same data sets can lead to different information depending on their structure including their (inter-) linkages.

Data

21. Data is a key means through which information is conveyed. Etymologically, the word "data", comes from Latin and is the plural of "datum", meaning "what is given" (and accepted as fact).[50] This origin is consistent with many of today's definition of data. Most scientists agree that data is a "representation"[51] of "disconnected facts"[52] or "observations"[53] about a "phenomenon"[54] (Zins, 2013a). Data is "atomic"[55] in the sense that it is the "smallest collectable unit associated with a phenomenon"[56]. It can be stored and communicated through "qualitative or quantitative"[57] "symbols"[58] or "artefacts"[59] suitable for being "processed"[60], "interpreted"[61] and "analysed"[62] by "human beings or by automated systems"[63] (Wellisch, 1996 cited in Zins, 2013a).

22. In contrast to information, data is commonly assumed to have no inherent meaning.[64] Data is in itself unorganized without any inherent structure or relationship within itself. It is therefore often seen as a "raw material"[65] that, until processed and interpreted, appears seemingly random and useless. In other words, it is only when it is processed, structured, and put in context, that data reveals its conveyed information. Structured data, in contrast, as stored for example in tables of a SQL database, already convey information which is reflected in the structure of the tables and their inter-linkages. Furthermore, the column names of the tables provide meta-information, without which the meaning of the tables' content cannot be interpreted. This means, on the other hand, that there are a number of means that can prevent or at least increase the cost for transforming "raw" data into information (Box 4).

23. To view data as "raw material" is consistent with some characteristics that data shares with other resources and "manipulable objects" (Hey, 2004). In particular, data has an "objective existence", it is "*discrete*, it can *pile up,* be *recorded* and *manipulated,* or *captured* and *retrieved*" (Hey, 2004; Zins, 2007b). It can be stored in databases, can fill a repository, and can be transported over networks. As a result, data can be measured.[66] However, data is an intangible and thus also has some unique properties in contrast to physical goods and resources.

24. Data is typically gained from information when information is *encoded* so it can be stored or communicated. Real world phenomenon can also be directly captured and transformed into data, thanks to the increasing deployment and interconnection of sensors through mobile and fixed networks (*i.e.* sensor networks) (OECD, 2009; 2013b).[67] This process of transforming the world into processable and quantifiable data is sometimes referred to as "datafication", a portmanteau for "data" and "quantification" (Hey, 2004; Bertolucci, 2013; Mayer-Schonberger and Cukier, 2013). As Mayer-Schonberger and Cukier (2013) explain: "To datafy a phenomenon is to put it in a quantified format so it can be tabulated and analyzed".

Box 4. Barriers to information extraction

The following means can prevent or at least increase the cost for transforming "raw" data into information:

- **Collection limitation**: The OECD (2013c) *Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data* lists the Collection Limitation Principle as its first basic principle. The Collection Limitation Principle states that "there should be limits to the collection of personal data and any such data should be obtained by lawful and fair means and, where appropriate, with the knowledge or consent of the data subject". Collection limitation can be considered the strongest means for preventing information extraction. This is because: (i) where no data is collected, no information can be obviously extracted. But also, as Pfitzmann and Hanse (2010) have highlighted, data minimization – which includes collection limitation – "is the only generic strategy [misinformation or disinformation aside] to enable unlinkability, since all correct personal data provide some linkability". As described below, linkability can be a strong means for undermining measures such as anonymisation and pseudonymisation.

- **Cryptography** is a practice that embodies principles, means, and methods for the transformation of data in order to hide its information content, establish its authenticity, prevent its undetected modification, prevent its repudiation, and/or prevent its unauthorised use. It is one of the technological means to provide security for data in information and communications systems. Cryptography can be used to protect the confidentiality of data, such as financial or personal data, whether that data is in storage or in transit. Cryptography can also be used to verify the integrity of data by revealing whether data has been altered and identifying the person or device that sent it. The OECD (1997) *Council Recommendation concerning Guidelines for Cryptography Policy* set forth a number of Principles concerning cryptography policy such as (i) Trust in Cryptographic Methods; (ii) Choice of Cryptographic Methods; and (iii) Market Driven Development of Cryptographic Methods to name of few.[68]

- **Anonymisation** is a process in which an entity's identifying information (II) (incl. personal identifying information, PII) are excluded or masked so that the entity's identity cannot be reconstructed, and the entity identified (Pfitzmann and Hanse, 2010; Mivule, 2012). Researchers have shown, however, that anonymised data *linked* with additional data can be de-anonymised; that is, II can be reconstructed (Narayanan and Shmatikov, 2007; Ohm, 2009). Narayanan and Shmatikov (2007), for example, have used the "anonymous" dataset released as part of the first Netflix prize to demonstrate how they could correlate Netflix's list of movie rentals with reviews posted on the Internet Movie Database (IMDb). This let them identify some individuals, and gave them access to their complete rental histories (Warden, 2011).

- **Pseudonymisation:** Having complete anonymity would prevent any useful two-way communication and transaction. For many applications, some kinds of identifiers are needed. Pseudonymisation is the process by which the most identifying attributes (*i.e.* identifiers) within a data record are replaced by unique artificial identifiers (*i.e.* pseudonyms[69]). The purpose is to take away the data record's identifiers that would link the data to a real person, and to keep all other attributes.

- **Unlinkability and distribution**: refer two or more items of interest (IOIs) (e.g., subjects, messages, actions) for which the data processor cannot sufficiently distinguish whether these IOIs are related or not (Pfitzmann and Hanse, 2010). According to ISO (2009 cited in Pfitzmann and Hanse, 2010), unlinkability "ensures that a user may make multiple uses of resources or services without others being able to link these uses together. [...] Unlinkability requires that users and/or subjects are unable to determine whether the same user caused certain specific operations in the system". Anonymisation and pseudonymisation are meant to enable unlinkability. Other technical means include distribution (decentralization) and separation of information systems and networks.

- **Noise addition, misinformation, and disinformation**: Misinformation is false or inaccurate information that is spread unintentionally, in contrast to disinformation, which is false or inaccurate information spread intentionally to mislead. Both are "noise" that can lead to false results if not filtered through data analytics, and they are considered by some individuals and researchers as a means to protect information including PII. Noise addition techniques, in particular, are being discussed as a solution to help protect privacy and confidentiality in databases, while keeping all data sets "statistically close" to the original data sets. Noise addition could allow analysis based on the complete data set to remain significant while masking sensitive data attributes. However, achieving optimal data privacy while not shrinking data utility through noise addition techniques is still a challenge (Mivule, 2012).

25. Datafication should not be confounded with digitization, which refers to the process of transforming information into *digits* that can be represented by binary numbers and thus processed by computers. For example, the digitization of books and photos through scanning generates digital content as images, which can be stored and transmitted as zeros and ones of binary (digital) code. Although digital, the resulting data cannot be directly indexed, searched, or otherwise analysed by software. Because "the text hadn't been datafied […] only humans could transform [it] into useful information – by reading" (Mayer-Schonberger and Cukier, 2013). It is only after the image has been processed through optical character recognition (OCR) and transformed into machine-encoded text that it can be used by software for further analytical operations, including, but not limited to, natural language processing (NLP).[70] Although datafication does not require digitization (e.g. recording the number of shop visitors via pen and paper), it is true that digitization has significantly accelerated the capacity to datify and to analyse the resulting data.

26. Last, but not least, it is important to note that data is always "theory laden"[71], that is data can only be understood within the context of a specific theory or domain. This is why some have argued that there is no raw data *per se* since every data has already been affected by knowledge (Tuomi, 2009 cited in Ubaldi, 2013). Data is "theory laden" already at the level of its representation, which is typically based on pre-defined conventions for *encoding information*. For example, the American Standard Code for Information Interchange (ASCII), which is used to represent text in computers and communications equipment, defines a scheme to translate (encode) 128 specified characters into 7-bit binary integers. Data is also "theory laden" at the semantic level. For example, in order to understand a dataset on broadband penetration as provided by the *OECD Broadband Portal*[72], users need to understand a number of issues related to the underlying theory, such as (i) how broadband penetration is defined, in particular what technologies are included, and (2) how growth rates are calculated, in particular whether the rates describe year-on-year or quarter-on-quarter growth. In order to understand the data correctly, that is to extract the right information, users thus need to have complementary knowledge. This knowledge is typically gained from *meta-information* or *meta-data* provided with the data sets. The *OECD Broadband Portal*, for example, provides a list of "broadband criteria (definitions)"[73] required by non-experts to understand the broadband data.

Data taxonomy

27. There are a number of data categories that come with different implications for individuals, businesses, and policy makers. For example, open data is an increasingly relevant concept, in particular in the context of public sector data and information. Whether data is classified or not as "open data" depends on factors such as intellectual property rights (IPR) and technological standards, and it will typically have an impact on issues such as open innovation, competition, and consumer protection and empowerment. This section develops a taxonomy that distinguishes among and defines a number of data categories, many of which are at the core of KBC2:DATA . Table 1 lists those categories that are discussed in more detail below. These categories can be regrouped in different higher level categories depending on (i) their economic and social as well as (ii) technical dimensions of data. These categories are also presented in this paper as different aspects of a larger phenomenon known as "big data", which will be discussed at the end of this section.

Table 1. Data taxonomy and related issues

Economic and social dimensions	Technical dimension
• **Public vs. private sector data**	
• **Open vs. closed data**	• **Structured vs. unstructured data**
• **Personal vs. non-personal data**	• **Linked data vs. data silos**
• **Volunteered vs. observed vs. inferred data**	• **Meta-data vs. primary data**
• **User created vs. machine generated data**	• **Real- or near-time vs. static data**
• **Micro vs. sectorial vs. macro data**	
Big data vs. small data?	

Public vs. private sector data

28. The OECD (2008a) *Recommendation for Enhanced Access and More Effective Use of Public Sector Information* (PSI) refers to PSI as "information, including information products and services, generated, created, collected, processed, preserved, maintained, disseminated, or funded by or for the Government or public institution". The Recommendation thus covers all products that convey public sector information including:[74]

- ***Public sector content*** which is digital content typically held by cultural establishments and has characteristics of being: (i) static (*i.e.* it is an established record), (ii) held by the public sector rather than being generated by it (*e.g.* cultural archives, artistic works where third-party rights may be important), (iii) not directly associated with the day-to-day functioning of government, and (iv) not necessarily associated with commercial uses but having public good characteristics (*e.g.,* culture, education); and

- ***Public sector data*** which are characterised as being: (i) dynamic and continually generated, (ii) directly produced by the public sector, (iii) associated with the functioning of the public sector (*e.g.* meteorological data, geo-spatial data, business statistics), and (iv) often readily useable in commercial applications with relatively little transformation, as well as being the basis of extensive elaboration.

29. According to the *OECD Glossary of Statistical Terms*, the public sector is defined as covering the (central and local) government administration as well as all "public corporations including the central bank"[75], which are engaged more or less in commercial and/or public service delivery. PSI, and in particular public sector data, can thus be broken down further into: (i) *public corporate data*: public sector data produced and provided by public corporations, and (ii) *government data*: public sector data produced and provided by (central and local) government administration. Very often, however, the term "government data" is used more broadly as a synonym for public sector data and in some cases even for PSI (including public sector content) (Ubaldi, 2013).[76]

30. *Private sector data*, in contrast, is expected by many to refer to data produced and/or owned by the private sector, which comprises "private corporations, households and non-profit institutions serving households (NPISHs)" according to the *OECD Glossary of Statistical Terms*.[77] Such a definition of *private sector data*, however, would conflict with the one presented above on public sector data. This is not only

because the concept of data ownership is an often misunderstood and misused concept (see section below), but also because the OECD (2008a) *Recommendation for Enhanced Access and More Effective Use of Public Sector Information* lists two conditions in which data qualifies as public sector data even though it may have been produced by the private sector: namely when the data has been "funded *by* or *for* the Government or public institution" (emphasis added). This covers a wide range of data sets, many of which can be produced or controlled by the private sector, but still be public sector data. For example, meteorological data collected and maintained by a private company but funded by and for the government qualifies as public sector data according to the OECD (2008a) *Recommendation for Enhanced Access and More Effective Use of Public Sector Information*. It is therefore suggested to define private sector data as a complement to public sector data, namely as data that is *not* generated, created, collected, processed, preserved, maintained, disseminated, or funded by or for the Government or public institutions, or positively as *data that is generated, created, collected, processed, preserved, maintained, disseminated, and funded only by the private sector*.

Box 5. A shift from public to private sector data

It is interesting to note that there has been a historically important transformation from public sector data to private sector data, primarily caused by the privatisation of data-intensive sectors including, but not limited to, telecommunication services, financial services, transportation, and utilities. This had implications for the scope of legal regimes and policies related to access to data such as those promoted, for example, by PSI frameworks, but also by "Freedom of Information" laws, which guarantee a public right of access to information from a human rights perspective (Siraj, 2010; Ubaldi, 2013). The social and economic implications of this transformation which has accelerated as the economy becomes more data intensive remain underexplored. This is important to note from a public policy perspective because the private sector is increasingly performing data-intensive public services that were traditionally performed by the government, while at the same time the relative scope of current data access regimes is narrowing compared to the range of data-intensive public services.

Open vs. closed data

31. The term "open data" is increasingly used in many different contexts and thus may refer to different concepts, which however may share a number of commonalities. Open data for governments, for example, often refers to initiatives such as data.gov (United States), data.gov.uk (United Kingdom), or data.gov.fr (France), which enhance access to public sector information (PSI), including public sector data, as encouraged by the OECD (2008a) *Council Recommendation on Enhanced Access and More Effective Use of Public Sector Information* (see section above, see also Ubaldi, 2013).[78] The term "open data" in the scientific community refers to open access to scientific data as promoted for example by the OECD (2004) *Declaration on Access to Research Data from Public Funding* and the OECD (2005) *Principles and Guidelines for Access to Research Data from Public Funding*. All these OECD instruments highlight openness as the first key principle. Last, but not least, open data is also often associated with movements such as the open source movement, which became particularly popular in the context of open source software (OSS) such as Linux (see Box 6 illustrating "openness").[79]

Box 6. Illustrating "openness"

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

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

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

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

Source: OECD (2013d)

32. It is important to note that the concept of open data is not limited to the public sector. UN Global Pulse (2012), for example, introduced the concept of "data philanthropy", whereby the private sector shares data to support more timely and targeted policy action, and to highlight the public interest in shared data. In this context two ideas are debated: *i)* the "data commons" where some data are shared publicly after adequate anonymisation and aggregation; and *ii)* the "digital smoke signals" where sensitive data are analysed by companies but results are shared with governments. The Open Data Institute (ODI), a not-for-profit organisation based in the United Kingdom, is also promoting the release of open data in the private sectors including, but not limited to, finance and health care.

33. Most definitions for open data point to a number of criteria or "principles". According to the OECD (2005) *Recommendation on Principles and Guidelines for Access to Research Data from Public Funding*, for example, "openness means access on equal terms for the international research community at the lowest possible cost, preferably at no more than the marginal cost of dissemination". Open data is characterized here by (i) access that should be granted on equal or non-discriminatory terms and by (ii) access costs that should not exceed the marginal cost of dissemination. As another example, at a meeting of open data advocates in 2007, participants agreed on "8 principles of Open Government"[80]. These principals include:

1. *Complete*: All public data is made available. Public data is data that is not subject to valid privacy, security or privilege limitations.

2. *Primary*: Data is as collected at the source, with the highest possible level of granularity, not in aggregate or modified forms.

3. *Timely*: Data is made available as quickly as necessary to preserve the value of the data.

4. *Accessible*: Data is available to the widest range of users for the widest range of purposes.

5. *Machine processable*: Data is reasonably structured to allow automated processing.

6. *Non-discriminatory*: Data is available to anyone, with no requirement of registration.

7. *Non-proprietary*: Data is available in a format over which no entity has exclusive control.

8. *License-free*: Data is not subject to any copyright, patent, trademark or trade secret regulation. Reasonable privacy, security and privilege restrictions may be allowed.

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

35. Among the criteria listed in the definitions presented so far, *non-discriminatory access* (or "access on equal terms" as stated in the OECD (2005) *Recommendation on Principles and Guidelines for Access to Research Data from Public Funding*) is the criterion that is always highlighted as central for open data. Non-discriminatory access is about "terms that do not depend on the users' identity or intended use" (Frischmann, 2013; see also United Kingdom Government, 2012). In other words, open data is data for which access is not limited based on users' identity or intended use of the data. As is highlighted in Box 7, identity- and intent-independent access to resources can be crucial for maximizing the value of data across the society as it keeps the range of opportunities as wide as possible.

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

1. **Technological design** is a broad concept including all technical aspects affecting the (re-) use and distribution of data. These factors were presented in Tim Berners-Lee's proposed *5 Star Deployment Scheme for Open Data*. They include: (i) "make your stuff available on the Web (whatever format) [under an open license]", (ii) make it available as structured data (e.g., Excel instead of image scan of a table), (iii) "use non-proprietary formats (e.g., CSV instead of Excel)", (iv) "use URIs [Uniform Resource Identifiers] to identify things, so that people can point at your stuff", (v) "link your data to other data to provide context". In essence, the *5 Star Deployment Scheme for Open Data* points to the following key technological factors affecting the degree of data openness: (i) data availability (online in the best case), (ii) machine readability (of structured data), and (iii) data linkability. It should be noted that the factor (i) is required for factor (ii), which in turn is a requirement for factor (iii).

2. **Intellectual property rights (IPRs)**: Data can be subject to legal regimes such as copyright as well as other IPRs applicable to databases and trade secrets, which need to be respected as highlighted in the OECD (2008a) *Council Recommendation on Enhanced Access and More Effective Use of Public Sector Information*. These rights can in some cases limit or prevent the (re-) use and distribution of open data. Some open data initiatives therefore explicitly state that open data should be free of any IPRs (see the *8 Principles of Open Government*). In other cases, innovative IP regimes are used and even promoted, through open data regimes as long as they do not restrict the rights of users to reuse and sometimes redistribute the data. In 2010, for example, the United Kingdom created the *Open Government Licence*[81] to release public sector information

(including data) for free without restricting (re-) use and distribution with the only requirement being attribution. This new licence scheme was based on the *Creative Commons (CC)* licences, another licence scheme widely used for open data (see data.australia.gov.au, data.gv.at, and *Google Ngram Viewer*[82]). Another example of open licence schemes used for data includes the *Open Data Commons Open Database License* (ODbL), which is for example used for *OpenStreetMap* data (for further discussion on IPRs see the IPR-pillar of the OECD KBC project).[83]

3. **Pricing**: Although the impact of pricing can be considered less heavy on the degree of openness compared to technological design and IPRs, pricing can be one of the most challenging factors, because optimal pricing can be hard to determine. Many governments, for example, wish to engage in cost recovery, partly for budgetary reasons and partly on the principle that those who benefit should pay. But the calculation of benefits can be problematic due to significant spill-over effects through the creation of public and social goods based on open data. Furthermore, as Stiglitz *et al.* (2000) have argued, if government provision of a data-related service is a valid role, then generating revenue from that service is not. Many open data initiatives therefore encourage the provision of data "at the lowest possible cost, preferably at no more than the marginal cost" as stated in the OECD (2005) *Recommendation on Principles and Guidelines for Access to Research Data from Public Funding*. The OECD (2008a) *Council Recommendation on Enhanced Access and More Effective Use of Public Sector Information* further specify that "where possible, costs charged to any user should not exceed marginal costs of maintenance and distribution, and in special cases extra costs for example of digitisation". While marginal cost pricing is often considered the best option for the public sector, this option is, however, seen as unattractive for the private sector for which at least cost recovery is a necessity. This can lead to average cost pricing as an alternative pricing model or can even require complex revenue models including subscription fees, freemium[84], voluntary donations, in combination with cross-subsidies as highlighted in Box 7.

37. The three factors presented above (technological design, IPRs, and pricing) determine the degree of openness, which can range from *closed* (access only by the data controller) to *open to the public* at its two extremes. In between, access may be restricted to (i) individual stakeholders who can affect or are affected by the use of the data, with access typically being granted on discriminatory bases, and to (ii) specific communities (see OECD *Recommendation on Principles and Guidelines for Access to Research Data from Public Funding* with access being restricted to the "international research community"). This leads to *a three level definition of open access* as illustrated in Figure 2.

Figure 2. Open data continuum

38. Overall, open data can be an optimal (private and social) strategy for maximizing the benefits of data, in particular in environments characterized by high uncertainty, complexity, and dynamic changes such as climate change, urban development, and health care research (Box 7). These complex systems are often characterized by complementary elements, and non-discriminatory access can be a means to internalize these complementary effects by encouraging "experimentation and innovation among complementary applications" (Frischmann, 2013).

Box 7. Non-discriminatory access: maximizing the economic and social value of data

Data as no intrinsic value as its value depends on the context of its use. The value-in-use, however, is difficult or almost impossible to fully assess *ex ante*, given the non-rivalrousness nature of data and the resulting wide range of downstream production of private, public, and social goods that data can enable. As a consequence, there can be significant (social) opportunity costs in closing access to some data. In other words: open (closed) data enables (restricts) user opportunities and degrees of freedom in the downstream production of private, public, and social goods, many of which having by nature significant spill-over effects (Frischmann, 2013). In particular, in environments characterized by high uncertainty, complexity, and dynamic changes open data can be an optimal (private and social) strategy for maximizing the benefits of data.

Based on Frischmann (2013), one can distinguish among the following factors for adopting non-discriminatory access. It should be highlighted at this point that the first two factors are closely associated to the concept of *open innovation*[85] as discussed in the OECD (2010) *Innovation Strategy* and the OECD (2013d) project on *Knowledge Networks and Markets*:

(i) Facilitating joint production or co-operation with suppliers, customers or even competitors is not a new phenomenon. Joint research ventures or patent pools are well known examples, where firms share resources under non-discriminatory access regimes to facilitate joint production. This is "because independent research efforts are inhibited by complexity, expense, strategic concerns, transaction costs, or other impediments" (Frischmann, 2013). Data sharing agreements are very often an important part of these collaboration efforts. In these cases data does not need to be open to the public, but openness is limited to the partners (level 2 open access, see Figure 11) who share their data as commons to "overcome collective action problems, sometimes mere co-ordination problems and sometimes more difficult prisoner's dilemma problems" (Frischmann, 2013).

(ii) Supporting and encouraging value-creating activities by users (incl. consumers) (user/consumer-driven innovation) can be enabled thanks to open access. Open access is an optimal strategy for organizations, "when they recognize that users may be best positioned to create value" (Frischmann, 2013). In its weakest form where users are granted access only to their own personal data (level 1 open access, see Figure 2), users and consumers are given e.g. "better visibility into their own consumption, often revealing information that can lead to changes in behavior" (MGI, 2013). An example for non-discriminatory access of community members (level 2 open access) is Kaggle[86], a crowd-sourcing platform on which governments, firms and individuals all over the world can post their data and let data scientists registered on Kaggle compete to produce the best analytic results. In its most extreme form, where access is granted to the public (level 3 open access), users (including consumers and citizens) are empowered to "provide input to improve the quality of goods and services" (MGI, 2013). This includes improving public services as well as the quality of data itself.[87]

(iii) Maximizing the option value of data when there is high uncertainty regarding sources of future market value. In contrast with the case described above, where organisations know that users are best placed to create future value, here organisations "are uncertain about the future sources of the value […] what unforeseen uses may emerge, what people will want, how much people will be willing to pay, what complementary goods and services may arise in the future, and so on" (Frischmann, 2013). They adopt open access strategies taking "advantage of the increased value of experimentation by users, the increased range of potential value-creating services, market selection of the best services that eventually emerge, and learning over time about user preferences and possible paths for continued development." (Frischmann, 2013). The advantage for the organisation is that it "maintains flexibility and avoids premature optimization or lock-in to a particular development path or narrow range of paths". In doing so it maximizes the option value of its data.

(iv) (Cross-)Subsidizing the production of public and social goods requires to pick winners (users or applications) by assessing (social) demand for such goods based on the (social) value they create (Frischmann, 2013). Governments can support the production of public goods (i) by directly producing these goods, or (ii) by supporting private firms' production of public and social goods through e.g. research grants, procurement programs, contracted research, and tax incentives. All these strategies raise a number of issues including, but not limited to, difficulties in picking winners and losers, and the fact that resources are limited. Open access regimes can be a more efficient and politically attractive "indirect intervention" to support the production of public and social goods. As Frishmann (2013) highlights "commons management is not a direct subsidy to […] users who produce public or social goods, but it effectively creates cross-subsidies and eliminates the need to rely on either the market or the government to 'pick winners' – that is, to prioritize or rank […] users worthy of access and support".

Personal vs. non-personal data

39. The OECD (2013c) *Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data* define "personal data" as "any information relating to an identified or identifiable individual (data subject)". Any data that are not related to an identified or identifiable individual is therefore non-personal. However, data analytics have made it easier to relate seemingly non-personal data to an identified or identifiable individual (Narayanan and Shmatikov, 2007; Ohm, 2010). These developments are blurring the distinction between personal and non-personal data and as a result they challenge any regulatory approach that determines the applicability of rights, restrictions and obligations on the basis of the "personal" nature of the data involved (OECD, 2013b).

40. There is however a conceptual issue with the definition of personal data as defined above, which explains the challenges faced by any regulatory approach based on the concept of "personal data": Data is not information and therefore cannot be personal *per se*. To recall: Data is a key means through which information is conveyed. In contrast to information, however, data has no inherent meaning. Only once it is processed, structured, and put in context, can data reveal its conveyed information. This information, however, is *context dependent*: It is a function of how data is processed, structured, and put in context. Whether the data conveys any personal information thus depends on a number of factors including, but not limited to, the data itself.

41. The use of Radio Frequency Identification (RFID) technologies[88] as discussed in OECD (2008b) illustrates this problem: "In some cases, RFID data is personal without ambiguity (e.g. in many access control applications). In other cases, RFID data may become personal data when it is possible to relate it to an identifiable individual. For example, when RFID is used in supply chain systems, the unique number stored on an RFID chip attached for example to a box of medicine to identify and track it, is not personal data. But the same RFID data can become personal data if it is collected or processed in such a manner as to enable a party to associate it with another set of information relating to an individual, i.e. by a nurse to track which patient has been provided with which medicine or by a drugstore to provide assistance services to patients".

42. To classify data as "personal" *ex ante* thus assumes that personal information can always be extracted. Equivalently, the reverse is assumed for "non-personal" data. These assumptions may be valid in many cases, but they cannot be generalised since a significant number of cases exist for which these assumptions are not true. Two types of cases can be identified:

- Those cases where the data conveys personal information, but no personal information can be extracted because the identification of an individual is too costly, or because of technical and skills limitations of the data controller, or because the identification of an individual is prohibited according to law, ethical norms and codes of conduct. A concrete example would be the applications of anonymisation techniques, which increase the cost of (re-)identification and thus require high-level skills and/or additional linked data sets for (re-) identification.

- The other cases are those where the data does not convey any personal "individualizable" information, and thus cannot not be used alone to extract personal information, but if linked with other data sets, suddenly can reveal information relating to an identified or identifiable individual. A concrete example would be when linking IP addresses to Internet Service Providers (ISP) information and their covered region together with the history of web search queries and websites visited (Ohm, 2010).

43. The cases presented above show that non-personal data can potentially convey information relating to an identified or identifiable individual or can at least enrich such information. When the

capacity to extract information is low, as in the past, the scale and scope of data that could be used to extract personal information was limited and could be well defined as "personal data". In such a world, protecting personal information could be achieved by protecting that well defined personal data. However, in a "big data" world, where the availability of exploitable data is high as well as the capacity to extract information out of it, the concept of personal data becomes inoperative. In other words, the scale and scope of data that can convey personal information or contribute to its enrichment is so large, that it becomes almost impossible to consistently apply the concept of "personal data" in practice. For these reasons, (personal) data can hardly be used today as effective point of control for privacy protection, and new points of control may have to be identified. What these points of control could be merits a discussion that goes beyond the scope of this paper.

Volunteered vs. observed vs. inferred data

44. Data can be distinguished by the way it is created. Data can be:

- **Volunteered** or surrendered by individuals when they explicitly share information about themselves or others. Examples include creating a social network profile and entering credit card information for online purchases.

- **Observed**, when captured by recording activities. In contrast to volunteered data where the data subject is actively and purposefully sharing its data, the role of the data subject in case of observed data is passive and it is the data controller that plays the active role. Examples of observed data include location data of cellular mobile phones and data on web usage behaviour.

- **Inferred**, when based on data analytic results. In this case, it is the data processor that plays the active role in the data analysis phase. The data subject typically has no control over what is inferred about her or him. Examples of inferred data include credit scores calculated based on an individual's financial history. It is interesting to note that personal information can be "inferred" from several pieces of seemingly anonymous or non-personal data (see Narayanan and Shmatikov, 2010).

User created vs. machine generated data

45. User created data refers to data that has been made available by an individual as opposed to machine generated data. It should be noted that the definition of user created data applied here is much larger than the concept of user created content (UCC) as defined in OECD (2007). There, UCC is defined as (i) content made publicly available over the Internet, (ii) which reflects a certain amount of creative effort, and (iii) which is created outside of professional routines and practices. Data, however, does not need to be made public in order to qualify as user created data. It also does not require a certain amount of creative effort, nor does the context of data creation (whether inside or outside professional routines) matters.

Microdata vs. sectoral data vs. macrodata

46. "Microdata" is a term most frequently used in particular in the context of statistics. It refers to data that has been collected and stored at the level of individual respondents or business entities. It is sometimes seen as the "true wealth" of National Statistic Offices (NSOs) (Giovannini, 2012). For example, the 2010 Community Innovation Survey (CIS), which is part of the EU science and technology statistics, includes (non-anonymised) microdata about of business innovation activities in 22 countries. The Current Population Survey (CPS), as another example, is a statistical survey conducted by the United States Census Bureau to collect microdata on the employment situation on a monthly basis.

47. Microdata can be anonymised or not, but they are always disaggregated in contrast to *sectoral data* and *macrodata*, which summarize individual details to aggregates at the sectoral or regional level respectively. The aggregation typically results in information loss, which could otherwise provide important insights. In particular, microdata can be used for exploring relationships between two or more different data sets when data linkage is feasible.[89] For example, the OECD Directorate for Science, Technology and Industry (DSTI) has developed a Micro-Data Lab to integrate microdata sets from its *OECD Patent Database* (73 million record on patent applications), the O*rbis*© database (85 million records containing comprehensive data on companies worldwide), the *Scopus*© database (26 million records on scientific publications) with trademark data (6 million records) and design right data (almost 1 million records). The inter-linkage of these different data sources through the Micro-Data Lab enables STI to have deeper insights into the origins of innovative activities across the economy (see section on data linkage).

48. Access to microdata provides researchers with much more freedom to investigate complex interactions and perform detailed analysis. Microdata allow for example to better understand industry and macro dynamics, and can be use to better inform policy design and monitoring (de Panizza and de Prato, 2009). However, microdata also can raise issues on confidentiality and privacy given that microdata are collected at the level of individual respondents or business entities. NSOs have developed a number of strategies to give access to microdata, while protecting confidentiality and privacy. For example, Eurostat, the statistical body of the European Commission, provides access to the CIS microdata only to researchers that have successfully applied for access and the full CIS microdata sets can only be accessed in the *Safe Center* at Eurostat's premises in Luxembourg.[90] DSTI, as another example, has pioneered a "distributed" approach to empirical analysis which draws on confidential micro data (see Box 8).

Box 8. The use of micro data in the OECD Directorate for Science, Technology and Industry (DSTI)

DSTI is working with large datasets on patents, trademarks, and now design rights, to develop new policy-relevant indicators, e.g. on the technological value of patents and on broader innovation. It has therefore developed a micro data lab which compiles and links at the micro level large scale administrative datasets with company information to look at IP bundles in companies, and to analyse the impact of IP assets and firms' performance. The source datasets are not "confidential" as in the case of micro data from statistical offices, but, in some cases the data providers are private companies and licensing agreements are needed.

In the case of micro data from statistical offices, access to micro data is restricted by laws that protect confidentiality and privacy in all countries. As a consequence, micro data from different countries cannot be pooled. The results are typically not comparable across countries as country-specific analyses generally use different models and methodologies. DSTI has pioneered a "distributed" approach to empirical analysis which draws on confidential micro data. The OECD provides a common framework (experts meet and identify common research and policy questions, the indicators and the econometric modelling is decided and software routines are developed) and researchers with access to their own country's micro data compile results that are then compared and analysed by OECD or lead countries. Advantages of a distributed approach to the analysis include:

1. Exploit the potential of firm-level data to address policy relevant questions, e.g. go beyond "average explanations to look at the heterogeneity of firms' behaviour;
2. Pragmatic way of addressing issues of access to confidential data and provide policy-relevant messages to policy makers;
3. Improve the relevance and usability of official statistics. Comparative analysis serves as a test of the usefulness of questions in surveys and provides feedback on survey design;
4. Build the case for the development of linked micro data statistical infrastructures in countries (a cost effective way to provide evidence based analysis and reduce burden on respondents);
5. Build the case for improved access to micro-data by researchers in countries.

Source: OECD Science, Technology and Industry Scoreboard (http://www.oecd.org/sti/scoreboard.htm)

Structured vs. unstructured data

49. Data are considered *structured* if they are based on a predefined *data model* (i.e. an abstract representation of "real world" objects and phenomenons).[91] Such models can be explicit as in the case of a SQL database where the data model is reflected in the structure of the database's tables and their inter-linkages. The data model can also be implicit as in the case of structured web content or as in the case of web logs, where the underlying (implicit) model can be made explicit at relatively low cost. Although they do not have an explicit but implicit data model, these data sources are often referred to as *semi-structured data*. Semi-structured data can also refer to data without explicit a data model, but which contains semantic elements such as *tags* that highlight the structure within the data. In contrast, *unstructured data* are data that have no predefined data model and where such models cannot be cost effectively extracted. Typical examples include text-heavy datasets such as text documents and e-mails, as well as multimedia content such as videos, images and audio streams. These unstructured data sets provide one of the greatest opportunities for data-driven growth (Box 8).

Box 9. Unstructured data: underexploited source for data-driven growth?

Unstructured data is by far the most frequent type of data, and thus provides the greatest potential for data analytics today. According to a survey of data management professionals by Russom (2007), less than half of total data stored in businesses is structured. The remaining data are either unstructured (31%) or semi-structured (21%). The author however admits that the real share of unstructured (including semi-structured) data could be much higher as only data management professionals dealing mostly with structured data and rarely with unstructured data were surveyed. Older estimates suggest that the share of unstructured data could be as high as 80% to 85% (see Shilakes and Tylman, 1998). In health care, for example, health records and medical images are the dominant type of data and they are stored as unstructured data. Estimates suggest that in the United States alone 2.5 petabytes are stored away each year from mammograms. A recent study by IDC (2012) suggests that the global volume of data will grow by a factor of 300, from 130 exabytes in 2005 to 40 000 exabytes, or 40 trillion gigabytes (more than 5 200 gigabytes for every person on earth) in 2020. Today, however, not even 5% of this "digital universe" is tagged and thus can be considered structured or semi-structured data.

It should be noted, however, that the differences between structured, semi-structured, and unstructured data are becoming less important in the long run. With growing computing capacities, data analytics are increasingly able to automatically extract some structure embedded in unstructured data. For example, optical character recognition (OCR) can transform images of text into machine-encoded text, which then can be indexed for search services such as via Google Books, and then used for data analytics, in particular natural language processing (NLP), for tagging or for extracting relevant communication patterns and even emotional patterns. Twitter, for example, has be discussed as a potential (unstructured) data source for analysing and even predicting the "emotional roller coaster" and its impact on the ups and downs of stock markets (Grossman, 2010; MIT Technology Review, 2010). Other examples include applications based on face recognition, which are powered by machine learning algorithms, which can recognize individuals from images and video streams. Facebook, for example, uses face recognition algorithms to automatically identify and tag its users out of user provided images (Andrade et al. 2013).

50. The concept of "structured data" is closely associated with that of "machine-readable data", which refers to data that can be interpreted by machines, or more precisely by software. The concept does not go as far as to require data to be semantically enriched as discussed in Box 2. All that is required is for the data to be structured or at least semi-structured. 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.

Linked data vs. data silos

51. As highlighted in the first section, the value of data is highly context-dependent. In particular, the value of data multiplies when it can be shared and linked with other data sets, as the data is put in a larger context which can reveal additional insights that otherwise were not possible to glean. Linked data is thus a major creator of (super-additive) value. In other words, the value of linked data is greater than the sum of its parts (data silos). 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).[92]

52. The term "linked data" sometimes also refers to structured data that are published so that they can be interlinked. The concept is also related to open data, which is sometimes seen as a requirement for linked data. This is because the full benefits of open data can be achieved only if open data sets can be interlinked. Linked data is also seen as a specific instance of the Semantic Web, known as the "Web of Data". In the Web of Data, data are cost effectively interlinked (including with *meta-data*) based on open standards such as the *Dublin Core metadata terms*. This standard defines 15 metadata elements for describing (web and physical) resources.[93]

53. The "super-additive" nature of linked data can explain why data linkage can undermine privacy protection measures such as anonymisation and pseudonymisation. To recall, researchers have shown that anonymised data *linked* with additional data, can be de-anonymised, that is: the personal identifying information (PII) can be reconstructed (Narayanan and Shmatikov, 2007; Ohm, 2009). Narayanan and Shmatikov (2007), for example, have used the "anonymous" dataset released as part of the first Netflix prize, and demonstrated how they could correlate Netflix's list of movie rentals with reviews posted on the public Internet Movie Database (IMDb). This let them identify some individuals, and gave them access to their complete rental histories (Warden, 2011). Unlinkability of two or more data sets is therefore seen as an important barrier to the extraction of information (see Box 4).

Meta-data vs. primary data

54. Meta-data is data about entities including, but not limited to, (primary) data. They provide the necessary 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 are needed to make (primary) data "useable and meaningful as a large library is useless without a card-catalogue system to organize and find the books". Meta-data are linked data in the sense that metadata are linked to the primary data. Meta-data are therefore also information enriching. Metadata can be categorised in several types depending on their purpose (see NISO, 2004). Metadata can be:

- **Descriptive**: that is based on attributes that can be used to search and find an entity such as title, author, subjects, keywords, or publisher.

- **Structural**: when used to describe the structure and organisation of an entity such as databases including the database tables, columns, and so on; and

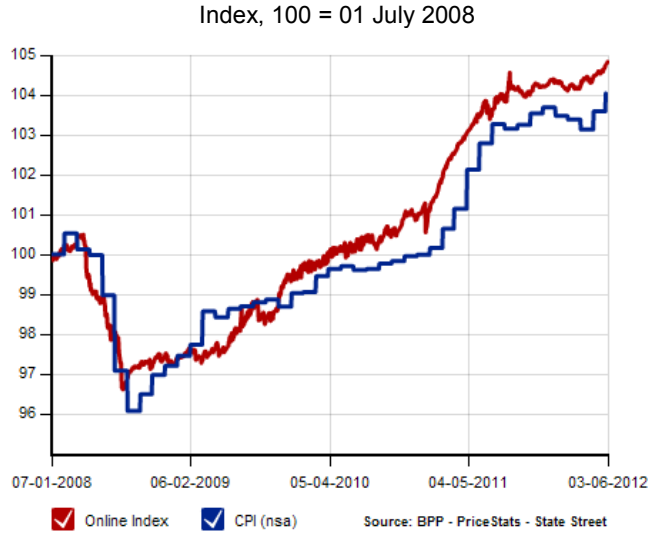
- **Administrative**: providing "information to help manage a resource, such as when and how it was created, file type and other technical information, and who can access it" (NISO, 2004).

Real- and near-time data vs. static data

55. Real-time or near-time data is data made available for processing immediately or almost immediately after collection. It is a dynamic stream of data, and therefore sometimes seen in contrast to **static data**. Real-time data is often used in situations where delays in data-driven decision making cannot be tolerated. Typical use cases include: (i) tracking and (ii) control systems (e.g. navigation systems).

56. Real-time data is also used to inform policy makers. For example, analysts have come to use readily available data to make real-time "nowcasts" ranging from purchases of automobiles to flu epidemics to employment/unemployment trends in order to improve the quality of policy and business decisions (Choi and Varian, 2009; Carrière-Swallow and Labbé, 2010). The Billion Price Project (BPP), launched at MIT and spun off to a firm called PriceStats, collects more than half a million prices on goods (not services) a day by "scraping the web". Its primary benefit is its capacity to provide real-time price statistics that are timelier than official statistics. In September 2008, for example, when Lehman Brothers collapsed, the BPP showed a decline in prices that was not picked up until November by the official Consumer Price Index (Surowiecki, 2011) (Figure 3). Real-time data is also used for security purposes, such as real-time monitoring of information systems and networks to identify malware and cyberattack patterns. The security company ipTrust, for instance, uses real-time data to assign reputation scores to IP addresses to identify traffic patterns from bot-infected machines in real time (Harris, 2011).

Figure 3. Daily online price index, United States, 2008-2012



Source: bpp.mit.edu.

Big data vs. small data?

57. Big data is often seen as a new paradigm in the data ecosystem (Autonomy, 2012; Zinow, 2012). However, the literature offers no clear definition of "big data". Initially the term "big data" referred to data sets for which volume became an issue in terms of data management and processing. This is consistent with many of today's definitions such as that suggested by Loukides (2010), who defines big data as data for which "the size of the data itself becomes part of the problem". The McKinsey Global Institute (MGI, 2011) similarly defines it as data for which the "size is beyond the ability of typical database software tools to capture, store, manage, and analyse".[94] The problem with such definitions is that they are in continuous flux, as they depend on the evolving performance of available storage technologies, which would suggest that big data depend on current computing capacities.

58. Furthermore, *volume* is not the only important characteristic highlighted by some experts as key to big data. Real-time data (*velocity*) is also sometimes highlighted as an important dimension of big data. In some other instances, big data is also defined by the capacity to analyse a *variety* of mostly unstructured data sets from sources as diverse as web logs, social media, mobile communications, sensors and financial transactions. This capacity is often associated with the capacity to link these diverse data sets (linked data). These three properties – volume, velocity and variety – are considered by many the three main characteristics of big data and are commonly referred to as the three Vs (Gartner, 2011).[95] Others have also suggested a fourth V, Value, which is related to the increasing social and economic value of data (OECD 2013b).

Box 10. Implications of "big data" for official statistics

Torrents of data streaming across public and private networks can improve the quality of statistics in an era of declining responses to national surveys and can create close to real-time evidence for policy making in areas such as prices, employment, economic output and development, and demographics. Some of the new sources of statistics are search engine data derived from keywords entered by users searching for web content. Google Insights for Search, for example, provides statistics on the regional and time-based popularity of specific keywords. Where keywords are related to specific topics such as unemployment, Google Insights can provide real-time indicators for measuring and predicting unemployment trends. Askitas and Zimmermann (2009), for example, analyse the predictive power of keywords such as "Arbeitsamt OR Arbeitsagentur" ("unemployment office or agency") for forecasting unemployment in Germany. The authors find that the forecast based on these keywords indicated changes in trends much earlier than official statistics (see also D'Amuri and Marcucci, 2010; Suhoy, 2010)

Other statistics are created by directly "scraping" the web. The Billion Price Project (BPP), for example, collects price information over the Internet to compute a daily online price index and estimate annual and monthly inflation. The online price index is basically an average of all individual price changes across all retailers and categories of goods. More than half a million prices for goods (not services) are collected daily by "scraping" the content of online retailers' websites such as Amazon.com. This is not only five times what the US government collects, it is also cheaper because the information is not collected by researchers who visit thousands of shops as they do for traditional inflation statistics. Furthermore, unlike official inflation numbers, which are published monthly with a lag of weeks, the online price index is updated daily with a lag of just three days. In addition, the BPP has a periodicity of days as opposed to months. This allows researchers and policy makers to identify major inflation trends before they appear in official statistics. For example, in September 2008, when Lehman Brothers collapsed, the online price index showed a decline in prices, a movement that was not picked up until November by the CPI (Figure 3; Surowiecki, 2011).

Currently, while methods to mine these new sources are still in their infancy and need rigorous scientific scrutiny, their rapid take-up by policy makers is a harbinger of a growing trend. Governments in the United States, the United Kingdom, Germany and France and in major non-OECD countries such as Brazil have established a partnership with PriceStats, which manages the BPP index, to contribute to and use the index. In another example, the Central Bank of Chile has explored the use of Google Insight for Search to predict present (i.e. to "nowcast") economic metrics related to retail good consumption (Carrière-Swallow and Labbé, 2010). For developing economies, in particular, where NSOs' capacity to sufficiently inform policy makers is often low, the exploitation of these new data sources provide a new opportunity to better inform public policy making for development (see UN Global Pulse, 2012).

Source: OECD (2012c).

Key data governance challenges

59. This section discusses the key challenges related to data governance. These are common challenges that individuals, businesses, and policy makers face in every domain in which data is used, irrespective of the type of the data used. Since many of these issues have been raised in the context of the specific data categories presented in the previous section, the discussion in the following section is kept short to avoid too much repetition. For example, the data governance issues related to "data access and sharing" were extensively discussed in the section on "open vs. close data", "data linkage and integration" in the section on "linked data vs. data silos", and "data protection and security" in the section on "personal vs. non-personal data" to name a few. This section discusses and highlights these specific data governance challenges separately to show that they are not necessarily bound to specific data categories and thus should be considered by governments, but also businesses, when designing data-related policies.

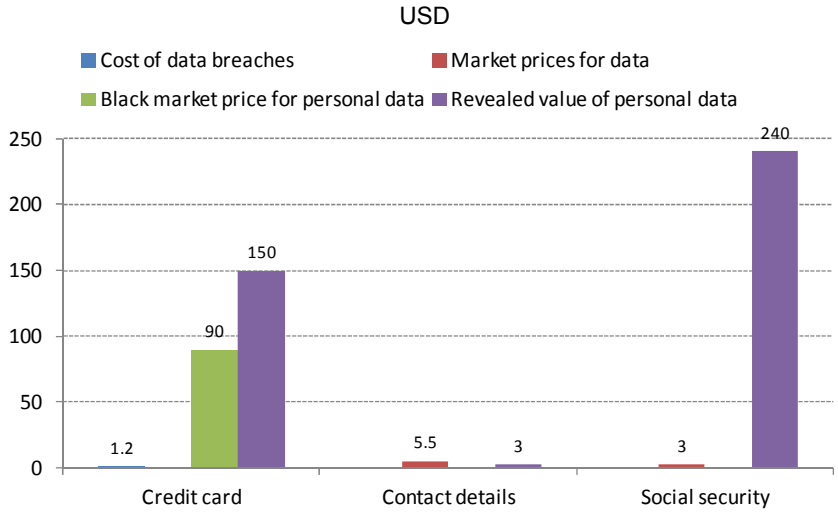
Data value and pricing

60. As highlighted in the previous section, data has no intrinsic value. Its value depends on the context of its use. This is partly because the information that can be derived from the data is context dependent. This also means, information more than any other good is an *experience good*, that is, a good that consumers must experience in order to value. "Virtually any new product is an experience good", however, "information is an experience good *every* time it's consumed" (Shapiro and Varian, 1999). The context dependency of data points to a number of other attributes affecting its value such as for example (i) the *accuracy* and (ii) *timeliness* of data: the more relevant and accurate data is for the particular context in which it is used, the more useful and thus valuable data is (see Oppenheim et al. 2004; cited in Engelsman, 2009). This implies on the other hand, that the value of data can perish over time depending on the use case (see Moody and Walsh, 1999; cited in Engelsman, 2009). These factors are related to data quality as discussed below.

61. Data pricing schemes are thus complex. In particular, the context dependency of data challenges the applicability of market-based pricing, since this assumes that markets can converge towards a price at which demand and offer meet. This, however, is not always the case. As a recent OECD (2012b) study "Exploring the Economics of Personal Data: A Survey of Methodologies for Measuring Monetary Value" showed, the monetary valuation of the same data set can diverge significantly among market participants. For example, while economic experiments and surveys in the United States indicate that individuals are willing to reveal their social security numbers for USD 240 on average, the same data sets can be obtained for less than USD 10 from U.S. data brokers such as Pallorium and LexisNexis (see Figure 4).

62. Data pricing schemes based on the cost structure seem to be a more common approach. The OECD (2005) *Recommendation on Principles and Guidelines for Access to Research Data from Public Funding* and the OECD (2008a) *Council Recommendation on Enhanced Access and More Effective Use of Public Sector Information*, both encourage the provision of data "at the lowest possible cost, preferably at no more than the marginal cost" which can include cost for "maintenance and distribution, and in special cases extra costs for example of digitisation". While marginal cost pricing is often considered the best option for the public sector, it is, however, seen as unattractive for the private sector for which at least cost recovery is a necessity. This calls for pricing schemes such as average cost pricing or more complex revenue models including subscription fees, freemium[96], voluntary donations, in combination with cross-subsidies as highlighted in Box 7.

Figure 4. Monetary value of selected data sets by means of valuation



Source: Based on OECD (2012b)

Data access and sharing

63. Access to data is a condition for the creation of the economic and social value of data. Data is a *non-rivalrous good*; that is, the use of data does not affect in principle its potential to meet the demands of others. There are therefore unlimited potential use cases in which data can be used to create value. Barriers to data access can inhibit data sharing and hinder collaboration, (open) innovation, and the downstream production of data-based goods and services, many of which having significant spill-over effects (see Box 7). As a consequence, there can be significant (social) opportunity costs due to barriers to access.

64. Some barriers to data access relate to the criteria affecting the degree of data openness: (i) technological design, (ii) IPRs, and (iii) pricing. Confidentiality and privacy consideration may be justified reasons for limiting data access in some cases as well. Furthermore, access problems and issues at the international level can emerge due to differences in culture and legislations. OECD (2013e) discusses the following factors in the particular context of science, but they are valid for other domains as well:

- **Legal and cultural barriers**: depending upon the perceived sensitivity of the data and/or the legal framework governing data-sharing arrangements some departmental 'gatekeepers' regulate access conditions tightly.

- **Public concerns**: to date there has been relatively little public engagement to explain the potential of data linkage, and the methods that are used to protect individual confidentiality when such linkages are made.

- **Technical barriers**: while various models for secure data access exist in some countries, the expertise, hardware and software to implement secure access is unevenly distributed among countries.

65. Finally, data access also requires the provision of meta-data to be accessed. As highlighted in the previous section, meta-data provide the necessary context without which the data cannot be accessed, linked, nor fully understood.

Data linkage and integration

66. The value of data is highly context-dependent – it explodes when it can be linked and integrated with other data sets. This is in particular true for microdata as the example of the Micro-Data Lab of the OECD Directorate for Science, Technology and Industry (DSTI) showed. As data is put in a larger context, it can reveal additional insights that otherwise were not possible to gain. Linked data thus creates super-additive value, which is greater than the sum of its parts (data silos).

67. There are various reasons why linking data across different data silos might be challenging. Some are obviously related to the legal, cultural, and technical barriers to data access and sharing as highlighted above. Others may be related skills barriers. As OECD (2013e) highlights: "even though techniques for record linkage are now well developed, and are used by numerous organisations regularly, the capacity with which to carry out successful linkages may be in short supply". It should be also highlighted that some of the barriers to data linkage are legitimate since data linkage cannot only be a source for great insights but also a serious means for undermining privacy protective measures such as anonymisation and pseudonymisation as highlighted in the previous section.

Data quality and curation

68. The information that can be extracted from data depends on the quality of the data. Because information is context dependent, data quality depends on the intended use: Data that is of good quality for certain applications can thus be of poor quality for other applications. The OECD (2011) *Quality Framework and Guidelines for OECD Statistical Activities* therefore defines data quality as "fitness for use" in terms of user needs. "If data is accurate, they cannot be said to be of good quality if they are produced too late to be useful, or cannot be easily accessed, or appear to conflict with other data" (OECD, 2011). Thus, data quality needs to be viewed as a multi-faceted concept. The OECD (2011) defines the following seven dimensions:

1. **Relevance**: "is characterised by the degree to which the data serves to address the purposes for which they are sought by users. It depends upon both the coverage of the required topics and the use of appropriate concepts";

2. **Accuracy**: is "the degree to which the data correctly estimate or describe the quantities or characteristics they are designed to measure";

3. **Credibility**: "the credibility of data products refers to the confidence that users place in those products based simply on their image of the data producer, i.e. the brand image. Confidence by users is built over time. One important aspect is trust in the objectivity of the data";

4. **Timeliness**: "reflects the length of time between their availability and the event or phenomenon they describe, but considered in the context of the time period that permits the information to be of value and still acted upon". Real-time data is data with a minimal timeliness";

5. **Accessibility**: "reflects how readily the data can be located and accessed" as discussed in the previous section on data access and sharing**;**

6. **Interpretability**: "reflects the ease with which the user may understand and properly use and analyse the data". The availability of meta-data plays an important role here as they provide for example "the definitions of concepts, target populations, variables and terminology, underlying the data, and information describing the limitations of the data, if any"; and

7. **Coherence**: "reflects the degree to which they are logically connected and mutually consistent. Coherence implies that the same term should not be used without explanation for different concepts or data items; that different terms should not be used without explanation for the same concept or data item; and that variations in methodology that might affect data values should not be made without explanation. Coherence in its loosest sense implies the data are 'at least reconcilable'".

69. The OECD (2013c) *Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data* also provide a number of criteria for data quality in the context of privacy protection. It states that "personal data should be relevant to the purposes for which they are to be used, and, to the extent necessary for those purposes, should be accurate, complete and kept up-to-date". So the data quality dimensions would have to include **completeness** as an eighth dimension according to the OECD Privacy Guidelines. Furthermore, the cost-efficiency with which data is collected could also be considered as a measure for data quality. "Whilst the OECD does not regard cost-efficiency as a dimension of quality, it is a factor that must be taken into account in any analysis of quality as it can affect quality in all dimensions" (OECD, 2011).

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

Data ownership and control

71. Data ownership is a concept that is often misunderstood and/or misused. In businesses, for example, data ownership is often used to assign responsibility and accountability for specific databases to "data owners". In this context it is perceived as a means to assure data quality and curation as well as data protection and security along the complete data life cycle. However, ownership is assigned without IPRs being granted to the "data owner" (see Scofield. 1998; Chrisholm, 2011). Scofield (1998) therefore suggests replacing the term "ownership" with "stewardship" as this implies the responsibility which organisations are looking to promote, where data users must consider the consequences of making changes over 'their' data.

72. The concept of ownership typically means "to have legal title and full property rights to something" (Chrisholm, 2011). Data is an intangible asset, as other information related goods, data can reproduced and transferred at almost zero marginal costs. So in contrast to the concept of ownership of physical goods, where the owner typically has exclusive rights and control over the good, including for example the freedom to destroy the good, this is not the case for intangibles such as data. For these types of goods IPRs are typically suggested as legal means to establish clear ownership. In the case of data in particular, legal regimes such as copyright as well as other IPRs applicable to databases and trade secrets can be used (see work stream on IPR in NSG:KBC).

73. However, in contrast to other intangibles, data typically involve complex assignments 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; cited in Department of Health and Human Services, 2013). So in many cases no single data stakeholder will have exclusive rights. Different stakeholders will typically have different power depending on their role. As Trotter (2012) highlights in the case of health patient data, all stakeholders

(including patient, doctor, and programmer) "have a unique set of privileges that do not line up exactly with any traditional notion of 'ownership'. Ironically, it is neither the patient nor the [doctor] who is closest to 'owning' the data. The programmer has the most complete access and the only role with the ability to avoid rules that are enforced automatically by electronic health record (EHR) software". Loshin (2002) identifies the following list of data stakeholders that could claim data ownership:

- **Creator**: "The party that creates or generate data";
- **Consumer**: "The party that uses the data owns the data";
- **Compiler**: "This is the entity that selects and compiles information from different information sources";
- **Enterprise**: "All data that enters the enterprise or is created within the enterprise is completely owned by the enterprise";
- **Funder**: "the user that commissions the data creation claims ownership";
- **Decoder**: "In environments where information is "locked" inside particular encoded formats, the party that can unlock the information becomes an owner of that information";
- **Packager**: "the party that collects information for a particular use and adds value through formatting the information for a particular market or set of consumers";
- **Reader as owner**: "the value of any data that can be read is subsumed by the reader and, therefore, the reader gains value through adding that information to an information repository";
- **Subject as owner**: "the subject of the data claims ownership of that data, mostly in reaction to another party claiming ownership of the same data"; and
- **Purchaser/Licenser as Owner**: "the individual or organization that buys or licenses data may stake a claim to ownership".

74. In cases where the data is considered "personal data" the situation is even more complex, since certain rights of the data subject cannot be taken away. For example, the Individual Participation Principle of the OECD (2013c) *Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data* recommends that individuals 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 for exclusive right on the storage and use of the data.

75. Overall, "the concept [of ownership] doesn't map well to the people and organizations that have relationships with that data" (Trotter, 2012). Data ownership can even be a poor starting point for data governance and can even be misleading. As Croll (2011) points out: "The important question isn't who owns the data. Ultimately, we all do. A better question is, who owns the means of analysis? Because that's how […] you get the right information in the right place. The digital divide isn't about who owns data — it's about who can put that data to work".

Data privacy and security

76. Data privacy refers to the protection of the privacy of individuals related to the collection and use of personal data. The OECD (2013c) *Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data* (Privacy Guidelines) defines eight Basic Principles of National Application which include:

1. **Collection Limitation Principle**: There should be limits to the collection of personal data and any such data should be obtained by lawful and fair means and, where appropriate, with the knowledge or consent of the data subject;

2. **Data Quality Principle**: Personal data should be relevant to the purposes for which they are to be used, and, to the extent necessary for those purposes, should be accurate, complete and kept up-to-date.

3. **Purpose Specification Principle**: The purposes for which personal data are collected should be specified not later than at the time of data collection and the subsequent use limited to the fulfilment of those purposes or such others as are not incompatible with those purposes and as are specified on each occasion of change of purpose.

4. **Use Limitation Principle**: Personal data should not be disclosed, made available or otherwise used for purposes other than those specified in accordance with Paragraph 9 except: a) with the consent of the data subject; or b) by the authority of law.

5. **Security Safeguards Principle**: Personal data should be protected by reasonable security safeguards against such risks as loss or unauthorised access, destruction, use, modification or disclosure of data.

6. **Openness Principle**: There should be a general policy of openness about developments, practices and policies with respect to personal data. Means should be readily available of establishing the existence and nature of personal data, and the main purposes of their use, as well as the identity and usual residence of the data controller.

7. **Individual Participation Principle**: An individual 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; at a charge, if any, that is not excessive; in a reasonable manner; and in a form that is readily intelligible to him; c) to be given reasons if a request made under subparagraphs(a) and (b) is denied, and to be able to challenge such denial; and d) to challenge data relating to him and, if the challenge is successful to have the data erased, rectified, completed or amended.

8. **Accountability Principle**: A data controller should be accountable for complying with measures which give effect to the principles stated above.

77. Data security is "traditionally" defined as the preservation of confidentiality, integrity and availability of information (see ISO, 2009b):

1. **Confidentiality** refers to the prevention of data disclosure to unauthorized individuals, entities or processes.

2. **Integrity** refers to the protection of data quality in terms of accuracy and completeness.

3. **Availability** refers to the accessibility and usability of data upon demand by an authorized entity.

78. It is interesting to note that data security is often seen as an integral part of data privacy. This is also underlined by the security safeguards principle of the OECD Privacy Guidelines. However, there are some challenges in implementing a "traditional" security and privacy approach which focusses on data in the context of a data-driven economy, which relies on an open and interconnected digital environment (see OECD, 2013f).

Conclusion

79. This paper, which contributes to KBC2:DATA, introduced the basic concepts and a taxonomy for data-related policies. These concepts, the taxonomy, and the identified key data governance issues will provide the basis for KBC2:DATA.

80. The paper highlights five major implications for further work under KBC2:DATA:

1. A holistic view is needed to understand data-driven value creation. This calls for a value chain/cycle approach.

2. The *non-rivalrous nature* of data and the *context dependency* of its value suggest that non-discriminatory access to data (*open data*) and data linkage (*linked data*) are two key concepts which are at the source of the opportunities and challenges for data-driven growth and well-being.

3. Data can ease the constraints placed on growth by scarcity of other physical capital goods and resources. However, what is scarce today is the capacity to use data in meaningful ways. This calls for a focus on data analytic skills and infrastructure including, e.g. cloud computing.

4. The increasing capacity to externalise knowledge, combined with the improving capacities of machines (empowered by software) to "understand" the world's growing volume of data, empowers machines to perform a wider range of tasks, many of which are considered labour- and knowledge-intensive. This comes with potential implications on employment and income inequalities that policy makers should be aware of.

5. "Big data" simply understood as 'data with huge volume' is not the only source for data-driven innovation. It turned out that pre-existing concepts such as open data, linked data, (un-) structured data and real-time data are at the core of data-driven innovation. If big data is perceived as the umbrella concept covering and unifying all these data concepts as suggested for example by the three Vs (volume, velocity and variety) definition, then big data can be perceived as the major source for data-driven innovation together with data analytics as the key enablers. In most cases, however, big data is still primarily understood in its narrow traditional sense as data for which the size is beyond the ability of typical database management software. But volume alone does not contribute to data-driven innovation, not to development, economic growth and well-being.

GLOSSARY

81. **Data** is introduced as the representation of facts stored or transmitted as qualified or quantified symbols. Data has no inherent meaning; however, it is theory-laden and can thus be domain specific.

82. **Information**[97] is the meaning resulting from the interpretation of facts as conveyed through data or other sources such as words. This meaning is reflected in the structure or organisation of the underlying source, including its hidden relationships and patterns of correlations, which can be revealed through data analytics. Information is therefore always context dependent: It depends on the capacity of its interpreter to extract meaning from the information source, including available data analytic techniques and technologies as well as skills and (pre-) knowledge.

83. **Knowledge** is information and experience internalised or assimilated in a process, commonly referred to as learning. It provides its user with the capacity to make effective decisions autonomously. Knowledge can be explicit, in which case it can be cost effectively externalised to be communicated about and embedded in tangible and intangible products. But it can also be tacit, based on an amalgam of information and experience, which is too costly to codify and thus to externalise.

84. The set of techniques and tools required to extract information from data is referred to as **data analytics**. There are a number of other terms frequently used. Some of these terms may include aspects that go beyond data analysis.

- **Data mining** and **knowledge discovery** typically include aspects such as data pre-processing (data cleaning), as well as model and inference considerations.
- **Profiling** is often used for describing the construction of *profiles* and the classification of entities in specific profiles.
- **Business intelligence**, a term coined by IBM researcher Luhn (1958), now refers to the analysis of business-related data as stored in databases (data warehouses) for business reporting and monitoring purposes.
- **Machine or statistical learning** is a subfield in computer science, and more specifically artificial intelligence (AI), concerned with the design, development and use of data analytic algorithms that allow computers to "learn", that is, to improve performance with every data set analysed.
- Last but not least, **visual analytics** refers to techniques and tools used for data visualization including (interactive) data exploration.

85. There are a number of means that can prevent or at least increase the cost for transforming "raw" data into information. They include:

- **Collection limitation** as called for in the OECD (2013c) *Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data*. It can be considered among the strongest means for preventing information extraction.

- **Cryptography** is a discipline that embodies principles, means, and methods for the transformation of data in order to hide its information content, establish its authenticity, prevent its undetected modification, prevent its repudiation, and/or prevent its unauthorised use. The OECD (1997) *Council Recommendation concerning Guidelines for Cryptography Policy* set forth a number of Principles concerning cryptography policy.

- **Anonymisation** is a process by which an entity's identifying information are excluded or masked so that its identity cannot be reconstructed. Researchers have shown, however, that anonymised data *linked* with additional data, can be de-anonymised.

- **Pseudonymisation** is the process by which the most identifying attributes (*i.e.* identifiers) are replaced by an artificial identifier (i.e. pseudonym). It is required for many applications for which some kinds of persistent identifier is needed, in particular for applications based on two-way communication and transactions, for which complete anonymity would be a barrier.

- **Unlinkability** of two or more data sets means that the data processor cannot sufficiently distinguish whether data entries within these datasets are related or not. Unlinkability ensures that a user may make multiple uses of resources or services without others being able to link these uses together. Anonymisation and pseudonymisation can be a first step to enable unlinkability, they do not however guarantee unlinkability.

- Last, but not least, **noise addition, misinformation, and disinformation**: can also be effective barriers to information extraction. Misinformation is false or inaccurate information that is spread unintentionally, in contrast to disinformation, which is false or inaccurate information spread intentionally to mislead. Noise addition techniques refer to techniques used to "mask" sensitive data attributes, while keeping the totality of the data set "statistically close" to the original data sets.

86. **Public sector data**, in respect to the OECD (2008a) *Recommendation for Enhanced Access and More Effective Use of Public Sector Information* (PSI), is data that is generated, created, collected, processed, preserved, maintained, disseminated, or funded by or for the Government or public institutions. It is characterised by being: (i) dynamic and continually generated, (ii) directly produced by the public sector, (iii) associated with the functioning of the public sector (e.g., meteorological data, geo-spatial data, business statistics), and (iv) often readily useable in commercial applications with relatively little transformation, as well as being the basis of extensive elaboration. It can be broken down further in: (i) *public corporate data*: public sector data produced and provided by public corporations, and (ii) *government data*: public sector data produced and provided by (central and local) government administration. **Private sector** data in contrast is data that is generated, created, collected, processed, preserved, maintained, disseminated, and funded only by the private sector.

87. **Open data** is data primarily characterised by *non-discriminatory access* or "access on equal terms" as stated in the OECD (2005) *Recommendation on Principles and Guidelines for Access to Research Data from Public Funding*. In other words, open data is data for which access is not limited based on users' identity or intended use of the data. Open data can maximize the economic and social value of data. It is an enabler for open innovation as discussed in the OECD (2010) *Innovation Strategy* and the OECD (2013d) project on *Knowledge Networks and Markets*. Openness is not a binary attribute but a **continuum** ranging from **closed data** (access only by the data controller) to access by the public. Three key factors affect the degree of openness:

- Technological design (incl. e.g. availability on the web, machine readability, and linkability),

- Intellectual property rights (IPRs) (incl. legal regimes such as copyright as well as other IPRs applicable to databases and trade secrets), and

- Pricing with marginal cost pricing being recommended by the OECD (2005) *Recommendation on Principles and Guidelines for Access to Research Data from Public Funding and the* OECD (2008a) *Recommendation for Enhanced Access and More Effective Use of Public Sector Information*.

88. **Personal data** is defined by the OECD (2013c) *Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data* as "any information relating to an identified or identifiable individual (data subject)". Any data that are not related to an identified or identifiable individual is therefore **non-personal data**. However, data analytics have made it easier to relate seemingly non-personal data to an identified or identifiable individual, thus transforming non-personal data into personal data. This suggests that, whether data conveys any personal information depends on a number of factors including, but not limited to, the data itself. To classify data as "personal" *ex ante* thus assumes that personal information can always be extracted. Equivalently, the reverse is assumed for "non-personal" data. These assumptions, however, may not be valid in all cases. For example, data that does not convey any personal "individualizable" information is typically not considered "personal data". But if linked with other data sets, suddenly the data can reveal additional information relating to an identified or identifiable individual (e.g. IP addresses). In an environment in which personal information is not a question only of the underlying data, but in particular that of data processing and linkage, the point of control for privacy protection shifts away from the data.

89. **Volunteered** data is provided by individuals when they explicitly share information about themselves or others. Examples include creating a social network profile and entering credit card information for online purchases. **Observed data** is captured by recording activities. In contrast to volunteered data where the data subject is actively and purposefully sharing its data, the role of the data subject is passive. Examples of observed data include location data of cellular mobile phones, and web usage behaviour. **Inferred data** are the output of data analytics. Examples include credit scores calculated based on an individual's financial history. It is interesting to note that personal information can be "inferred" from several pieces of seemingly anonymous or non-personal data.

90. **User created data** refers to data that has been made available by an individual as opposed to **machine generated data**. In contrast to user created content (UCC) as defined in OECD (2007), user created data can be assumed even where data have not been made public. The distinction between user created and machine generated data can raise issues related to IPRs.

91. **Structured data** are data that are based on a predefined *data model* (i.e. an abstract representation of "real world" objects and phenomenon). Such models can be explicit as in the case of a SQL database where the data model is reflected in the structure of the database's tables and their inter-linkages. The data model can also be implicit as in the case of structured web content or as in the case of web logs, where the underlying (implicit) model can be made explicit at relatively low cost (**semi-structured data**). In contrast, **unstructured data** are data that have no predefined data model and where such model cannot be cost effectively extracted. Typical examples include text-heavy datasets such as text documents and e-mails, as well as multimedia content such as videos, images and audio streams. Unstructured data account by far for the largest share of the global data volume. According to some estimates not even 5% of the digital universe can be considered structured or semi-structured data. It should be noted at this point, however, that the difference between structured, semi-structured, and unstructured data is becoming less important in the long term. With rising computing capacities, data analytics are increasingly able to automatically extract some structures embedded in unstructured data.

92. **Linked data** typically refers to structured data that are published so that they can be interlinked. Data linkage is a means to contextualise data and thus enable insights that are greater than the sum of its isolated parts (**data silos**). It is closely related to the concept of open data, for which the full benefits can only be achieved, if the isolated open data sets can be interlinked. Open standards play an important role in an interlinked data ecosystem. An example includes the "Web of Data", in which data are linked through *meta-data* standards such as the *Dublin Core metadata terms*.

93. **Meta-data** is data about entities including (**primary) data**. They provide the necessary context without which the data cannot be accessed, linked, or fully understood. Meta-data can be (i) descriptive (based on attributes used to search and find an entity), (ii) structural (describing the structure and organisation of an entity such as databases), and administrative (providing information to help manage a resource).

94. **Microdata** refer to data that has been collected and stored at the level of individual respondents or business entities. It is often derived from statistical surveys or administrative sources, such as business registers, patent applications or tax collections. Examples include data of the 2010 Community Innovation Survey (CIS) and the Micro-Data Lab of the OECD Directorate for Science, Technology and Industry (STI). Microdata underpin **sectoral data** and **macrodata,** which summarize details about individual respondents or business entities to aggregates at the sectoral or regional level respectively. The aggregation does not happen without information loss, which could otherwise provide important insights to researchers, e.g. as regards the heterogeneity of individual and business behaviour. Access to microdata provides researchers with much more freedom to investigate complex interactions and perform detailed analysis. In particular, microdata can be used for exploring relationships between two or more different data sets when data linkage is feasible. However, it also raises issues on confidentiality and privacy given that microdata are collected at the level of individual respondents or business entities, often guided by legislation on national statistics. National Statistic Offices (NSOs) have developed a number of strategies to give access to microdata, while protecting confidentiality and privacy.

95. **Real-time or near-time data** is data made available for processing immediately or almost immediately after collection. It is a dynamic stream of data, and is therefore sometimes viewed in contrast to **static data**. Real-time data is often used in situations where delays in data-driven decision making cannot be tolerated. Typical use cases include: (i) tracking and (ii) control systems (e.g. navigation systems). Analysts, for example, are using readily available data to make real-time "nowcasts" ranging from purchases of automobiles to flu epidemics to employment/unemployment trends in order to improve the quality of policy and business decisions.

96. **Big data** initially referred to data sets for which the (i) volume became an issue in terms of data management and processing. Further definitions highlighted that volume is not the only important characteristic of big data and pointed to (ii) velocity at which data is generated, accessed, processed and analysed (referring to real-time data), as well as (iii) variety (referring to linking unstructured data). These three properties – volume, velocity and variety – are considered by many to be the three main characteristics of big data and are commonly referred to as the three Vs. Others have also suggested a fourth V, Value, which is related to the increasing social and economic value to be obtained from the use of big data.

REFERENCES

Ackoff, R.L. (1989), "From Data to Wisdom", Journal of Applied Systems Analysis, Volume 16.

Acquisti, A., L. John and G. Loewenstein (2011), "What is Privacy Worth?", mimeo, http://pages.stern.nyu.edu/~bakos/wise/papers/wise2009-6a1_paper.pdf.

Alavi, M. and D.E. Leidner (2001) "Knowledge Management and Knowledge Management Systems: Conceptual Foundations and Research Issues", MIS Quarterly, Vol. 25, No. 1, pp. 107-136, March.

Alesso, H. P. and C. F. Smith (2008), *Thinking on the Web: Berners-Lee, Gödel and Turing*, 3 December, Wiley-Interscience.

Andrade, N.N.G, A. Martin and S. Monteleone (2013), "'All the Better to See You with, My Dear': Facial Recognition and Privacy in Online Social Networks", IEEE Security & Privacy, Vol. 11, No. 3, pp. 21-28, May/June.

D'Amuri, F. and J. Marcucci (2010), "Google it! Forecasting the US unemployment rate with a Google job search index", SSRN, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1594132.

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

Askitas N. and KN. Zimmermann (2010), "Google econometrics and unemployment forecasting", Technical report, SSRN 899, 2010, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1465341.

Australian Bureau of Statistics (1911), *Census of the Commonwealth of Australia, 1911*, 02-03 April, Volume I, Cat. no. 2112.0, available at: www.ausstats.abs.gov.au/ausstats/free.nsf/0/A81792B5E5298F70CA25783900108ACB/$File/1911%20Census%20-%20Volume%20I%20Statisticians%20Report.pdf.

Autonomy (2012), "How to Leverage Big Data to Monetize Customer Experiences", White Paper, www.marketingpower.com/ResourceLibrary/Documents/Whitepapers/Autonomy%20Whitepaper%20Final%202.28.2012.pdf.

Belkin, N. J. and S. E. Robertson (1976), "Information science and the phenomenon of information", Journal of the American Society for Information Science, 27, 197–204, DOI: 10.1002/asi.4630270402.

Berners-Lee, T. J. Hendler, and O. Lassila (2001), "The Semantic Web", Scientific American, 17 May, available at www.scientificamerican.com/article.cfm?id=the-semantic-web.

Bertolucci, J. (2013), "Big Data's New Buzzword: Datafication", InformationWeek, 25 February, available at: www.informationweek.com/big-data/news/big-data-analytics/big-datas-new-buzzword-datafication/240149288.

Blair, D. C. (2002), "Knowledge management: Hype, hope or help?" Journal of the American Society for Information Science and Technology, 53(12), 1019–1028, DOI: 10.1002/asi.10113.

Bracy, J. (2013), "Changing the Conversation: Why Thinking 'Data is the New Oil' May Not Be Such a Good Thing", Privacy Perspective, Privacy Association, 19 July, available at: www.privacyassociation.org/privacy_perspectives/post/changing_the_conversation_why_thinking_data_is_the_new_oil_may_not_be_such.

Brown, J. S. and P. Duguid (2000), "The Social Life of Information", First Monday, 5(4), 3 April, available at: http://firstmonday.org/ojs/index.php/fm/article/view/738/647.

Buckland, M. (1991a), *Information and information systems*, Greenwood Press, New York.

Buckland, M. (1991b), "Information as thing", Journal of the American Society of Information Science, 42(5), 351–360, available at: http://people.ischool.berkeley.edu/~buckland/thing.html.

Carrière-Swallow, Y. and F. Labbé (2010), "Nowcasting with Google Trends in an Emerging Market", *Central Bank of Chile Working Papers*, No. 588, July.

Chisholm, M. (2011), "What is Data Ownership?", BeyeNETWORK, 28 November, www.b-eye-network.com/view/15697.

Choi, H. and H. Varian (2009), "Predicting the Present with Google Trends", Discussion paper, Google, 10 April, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1659302.

Cleveland, H. (1982), "Information As a Resource", THE FUTURIST, December, available at http://hbswk.hbs.edu/pdf/20000905cleveland.pdf.

ComputerWeekly (2013), "How to manage unstructured data for business benefit", Analysis, last time accessed: 29 August, available at: www.computerweekly.com/feature/How-to-manage-unstructured-data-for-business-benefit.

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

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, CA, http://research.google.com/archive/mapreduce.html.

Debons, A., E. Horne and S. Cronenweth (1988), *Information science: An integrated view*. New York: G.K. Hall.

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

Department of Health and Human Services [United States] (2013), "Data Ownership", last time accessed: 18 November, http://ori.dhhs.gov/education/products/n_illinois_u/datamanagement/dotopic.html.

Duda, R., P. E. Hart, and D. G. Stork (2000), *Pattern Classification*, 9 November, Second Edition, Wiley-Interscience.

Dumbill, E. (2012a), "What is big data? An introduction to the big data landscape", O'Reilly Radar, 11 January, http://radar.oreilly.com/2012/01/what-is-big-data.html.

Dumbill, E. (2012b), "Big data market survey: Hadoop solutions", O'Reilly Radar, 19 January, http://radar.oreilly.com/2012/01/big-data-ecosystem.html.

Dumbill, E. (2011), "Data is a currency: The trade in data is only in its infancy", O'Reilly Strata, 23 February, available at: http://strata.oreilly.com/2011/02/data-is-a-currency.html.

Eliot, T. S. (1934), *The Rock*, London: Faber & Faber.

Engelsman, W. (2009), "Information Assets and their Value", in S. Gopalan (Ed.), *Knowledge Assets: Concepts and Measurements*, The Icfai University Press, Hyderabad, India.

Eurostat (2013), "How to Apply for Microdata?", September, http://epp.eurostat.ec.europa.eu/portal/page/portal/microdata/documents/How_to_apply_for_microdata_access.pdf.

Fazekas, M. and T. Burns (2012), "Exploring the Complex Interaction Between Governance and Knowledge in Education", *OECD Education Working Papers*, No. 67, OECD Publishing. doi: 10.1787/5k9flcx2l340-en

Garner, B. A. (1999), *Black's Law Dictionary*, 7 th edition, West Group, St. Paul, MN.

Gartner (2011), "Gartner Says Solving 'Big Data' Challenge Involves More Than Just Managing Volumes of Data", Press release, www.gartner.com/it/page.jsp?id=1731916.

Gentile, B. (2011), "The New Factors of Production and the Rise of Data-Driven Applications", *Forbes*, 31 October, www.forbes.com/sites/ciocentral/2011/10/31/the-new-factors-of-production-and-the-rise-of-data-driven-applications/.

Giovannini, E. (2012), "Micro foundations of macro data: increasing data quality and exploiting the 'true wealth' of National statistical institutes", Sixth ECB Statistics Conference on "Central bank statistics as a servant of two separate mandates: price stability and mitigation of systemic risk", 17-18 April, https://www.ecb.europa.eu/events/pdf/conferences/stats6th/Session_3_2Mr_Giovannini.pdf.

Glanz, J. (2013), "Is Big Data an Economic Big Dud?" The New York Times, 17 August, available at: www.nytimes.com/2013/08/18/sunday-review/is-big-data-an-economic-big-dud.html.

Gleick, J. (2011), *The Information: A History, A Theory, A Flood*, Fourth Estate, London.

Grossman, L. (2010), "Twitter Can Predict the Stock Market", Wired, 19 October, available at: www.wired.com/wiredscience/2010/10/twitter-crystal-ball/.

Hastie, T., R. Tibshirani, J. Friedman (2011), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Second Edition, 23 December, Springer.

Hawkins, D. T. (2001), "Information science abstracts: Tracking the literature of information science. Part 1: Definition and map", Journal of the American Society for Information Science and Technology, 52(1): 44-54, 7 December, DOI: 10.1002/1532-2890, available at: http://ms.lzu.edu.cn/wwhhss/Documents/Donald%20T%20Hawkins%20Information%20Science%20map.pdf.

Hess, C., and E. Ostrom, (2007). *Understanding Knowledge as a Commons: From Theory to Practice*. MIT Press.

Hey, J. (2004), "The Data, Information, Knowledge, Wisdom Chain: The Metaphorical link", working paper, December, available at: www.dataschemata.com/uploads/7/4/8/7/7487334/dikwchain.pdf.

von Hippel, E. (1988). *The Sources of Innovation*. New York: Oxford University Press.

Hjørland, B. (2002), "Principia informatica: Foundational theory of information and principles of information services", in H. Bruce, R. Fidel, P. Ingwersen, and P. Vakkari (Eds.), *Emerging frameworks and methods*, Proceedings of the Fourth International Conference on Conceptions of Library and Information Science (CoLIS4), pp. 109–121, CO: Libraries Unlimited, Greenwood Village.

Hjørland, B. (1998), "Theory and metatheory of information science: A new interpretation", Journal of Documentation, 54(5), 606-621, DOI: 10.1108/EUM0000000007183.

Hoberman, S. (2009), *Data Modeling Made Simple*, 2nd Edition, Technics Publications, LLC.

Hollerith, H. (1894), "The Electrical Tabulating Machine", *Journal of the Royal Statistical Society*, 57(4) December, pp. 678-689, Wiley, available at: www.jstor.org/stable/2979610.

Hoshovsky, A.G., and R.J. Massey (1968), "Information science: Its ends, means, and opportunities", Proceedings of the American Society for Information Science Annual Meeting, 5, 47-55.

IDC (2012), "The Digital Universe In 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East", IDC, The Digital Universe Project, December, available at: www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf.

Information and Privacy Commissioner Ontario [IPC] (2000), "Should the OECD Guidelines Apply to Personal Data Online?", A Report to the 22nd International Conference of Data Protection Commissioners, Venice, Italy, September, available at: www.ipc.on.ca/images/resources/up-oecd.pdf.

International Organization for Standardization [ISO] (2009a), *ISO/IEC 15408-1/2/3:2005 - Information technology — Security techniques — Evaluation criteria for IT security*, http://isotc.iso.org/livelink/livelink/fetch/2000/2489/Ittf_Home/PubliclyAvailableStandards.htm.

International Organization for Standardization [ISO] (2009b), *ISO/IEC 27000:2009 - Information technology — Security techniques — Information security management systems - Overview and vocabulary*, http://www.iso.org/iso/catalogue_detail?csnumber=41933.

James, G., D. Witten, T. Hastie, and R. Tibshirani (2013), *An Introduction to Statistical Learning: with Applications in R*, 12 August, Springer.

Janert, P. K. (2010), *Data Analysis with Open Source Tools*, November, O'Reilly Media.

Kroes, N. (2012), "Digital Agenda and Open Data: From Crisis of Trust to Open Governing", European Commission - SPEECH/12/149, 05 March, Bratislava, available at: http://europa.eu/rapid/press-release_SPEECH-12-149_en.htm.

Krugman, P. (2012a), "Rise of the Robots", The New York Times – Blog, 8 December, available at: http://krugman.blogs.nytimes.com/2012/12/08/rise-of-the-robots/.

Krugman, P. (2012b), "Capital-biased Technological Progress: An Example (Wonkish)", The New York Times – Blog, 26 December, available at: http://krugman.blogs.nytimes.com/2012/12/26/capital-biased-technological-progress-an-example-wonkish.

Lakoff, G., and M. Johnson (1980), *Metaphors We Live By*, The University of Chicago Press, Chicago and London.

Laney, D. (2001), "3D Data Management: Controlling Data Volume, Velocity, and Variety", META Group, 6 February, http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.

Le Coadic, Y. F. (2004), *La science de l'information*, Collection : Que sais-je? (No. 2873), Paris: PUF.

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

Loukides, M. (2010), "What is data science? The future belongs to the companies and people that turn data into products", O'Reilly Radar, 2 June, http://radar.oreilly.com/2010/06/what-is-data-science.html.

Lucky, R. W (1989), *Silicon Dreams: Information, Man, and Machine*, 1st edition, July, New York: St. Martin's Press.

Luhmann, N. (1996), *Soziale Systeme*. Frankfurt am Main: Suhrkamp.

Luhn H. (1958), "A Business Intelligence System", IBM Journal of Research and Development, 2,4, Page 314, available at: http://domino.watson.ibm.com/tchjr/journalindex.nsf/c469af92ea9eceac85256bd50048567c/fc097c29158e395f85256bfa00683d4c!OpenDocument.

Machlup, F. (1983), "Semantic Quirks in Studies of Information", in F. Machlup and U. Mansfield (eds.), *The Study of Information: Interdisciplinary Message,* Wiley, New York.

Mayer-Schonberger, V. and K. Cukier (2013), *Big Data: A Revolution That Will Transform How We Live, Work, and Think*, 5 March, Eamon Dolan/Houghton Mifflin Harcourt.

Merelli, E. and M. Rasetti (2013), "Non locality, topology, formal languages: new global tools to handle large data sets", International Conference on Computational Science, ICCS 2013, Procedia Computer Science 18 (2013) 90 – 99, doi:10.1016/j.procs.2013.05.172.

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

MGI: McKinsey Global Institute (2011), "Big data: The next frontier for innovation, competition and productivity", McKinsey & Company, June, www.mckinsey.com/~/media/McKinsey/dotcom/Insights%20and%20pubs/MGI/Research/Technology%20and%20Innovation/Big%20Data/MGI_big_data_full_report.ashx.

MIT Technology Review (2010), "Twitter Mood Predicts The Stock Market", MIT Technology Review, 18 October, available at: www.technologyreview.com/view/421251/twitter-mood-predicts-the-stock-market/.

Mivule, K. (2013), "Utilizing Noise Addition for Data Privacy, an Overview", Proceedings of the International Conference on Information and Knowledge Engineering (IKE 2012), Pages 65-71, arXiv:1309.3958v1.

Moody, D. and P. Walsh (1999), "Measuring the Value of Information: An Asset Valuation Approach", in The Seventh European Conference on Information Systems (ECIS'99), Copenhagen Business School, available at: wwwinfo.deis.unical.it/zumpano/2004-2005/PSI/lezione2/ValueOfInformation.pdf.

Morris, C.W. (1938), *Foundations of the theory of signs*, The University of Chicago Press, Chicago.

Narayanan, A., V. Shmatikov (2007), "How To Break Anonymity of the Netflix Prize Dataset", 22 November, http://arxiv.org/abs/cs/0610105v2.

NISO (2004), *Understanding Metadata*. NISO Press.

Nonaka, I., and H. Takeuchi (1995), *The Knowledge Creating Company*, Oxford University Press.

Nunberg, G. (1996), *Future of the book*, UCA Press, Berkeley, CA.

Oborn, E., Barrett, M., and Racko, G. (2010), "Knowledge translation in healthcare: A review of the literature". Working Paper Series 5/2010, Cambridge Judge Business Scholl, available at: www.jbs.cam.ac.uk/research/working_papers/2010/wp1005.pdf.

OECD (2013a), "New Sources of Growth: Knowledge-Based Capital – Key Analyses and Policy Conclusions", 25 July, available at: www.oecd.org/sti/inno/knowledge-based-capital-synthesis.pdf.

OECD (2013b), "Exploring Data-Driven Innovation as a New Source Of Growth: Mapping the Policy Issues Raised by 'Big Data'", *OECD Digital Economy Papers*, No. 222, OECD Publishing. doi: 10.1787/5k47zw3fcp43-en.

OECD (2013c), Recommendation of the Council concerning Guidelines governing the Protection of Privacy and Transborder Flows of Personal Data, C(80)58/FINAL, as amended on 11 July 2013 by C(2013)79, available at www.oecd.org/sti/ieconomy/2013-oecd-privacy-guidelines.pdf.

OECD (2013d), "Knowledge Networks and Markets", OECD Science, Technology and Industry Policy Papers, No. 7, OECD Publishing. doi: 10.1787/5k44wzw9q5zv-en.

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

OECD (2013f), "Outline for a Report on Privacy in a Data-Driven Economy: The Role of Data in Promoting Growth and Well-Being: Module 5", DSTI/ICCP/REG(2013)5.

OECD (2013g), "The OECD Model Survey on ICT Usage by Businesses: Proposal for the 2nd Revision", [DSTI/ICCP/IIS(2013)2], 12 November.

OECD (2012a), *Knowledge Networks and Markets in the Life Sciences*, OECD Publishing. doi: 10.1787/9789264168596-en

OECD (2012b), "Exploring the Economics of Personal Data: A Survey of Methodologies for Measuring Monetary Value", *OECD Digital Economy Papers*, No. 220, OECD Publishing. http://dx.doi.org/10.1787/5k486qtxldmq-en.

OECD (2012c), "Big Data and Statistics: Understanding the Proliferation of Data and Implications for Official Statistics and Statistical Agencies", DSTI/ICCP(2012)11.

OECD (2011), "Quality Framework and Guidelines for OECD Statistical Activities", STD/QFS(2011)1, 17 January, http://search.oecd.org/officialdocuments/displaydocumentpdf/?cote=std/qfs%282011%291.

OECD (2010), *The OECD Innovation Strategy: Getting a Head Start on Tomorrow*, OECD Publishing. doi: 10.1787/9789264083479-en

OECD (2009), "Smart Sensor Networks: Technologies and Applications for Green Growth", *OECD Digital Economy Papers*, No. 167, OECD Publishing. doi: 10.1787/5kml6x0m5vkh-en

OECD (2008a), *OECD Recommendation for Enhanced Access and More Effective Use of Public Sector Information*, 16 June, C(2008)36, available at: www.oecd.org/internet/ieconomy/40826024.pdf.

OECD (2008b), "Radio-Frequency Identification (RFID): A Focus on Information Security and Privacy", *OECD Digital Economy Papers*, No. 138, OECD Publishing. doi: 10.1787/230618820755.

OECD (2004), *Declaration on Access to Research Data from Public Funding*, C(2004)31/REV1, 30 January, available at: http://acts.oecd.org/Instruments/ShowInstrumentView.aspx?InstrumentID=157.

OECD (1997), *Recommendation of the Council concerning Guidelines for Cryptography Policy*, 27 March 1997 - C(97)62/FINAL, http://webnet.oecd.org/oecdacts/Instruments/ShowInstrumentView.aspx?InstrumentID=115&InstrumentPID=111.

OECD (1996), *The Knowledge-Based Economy*, OECD, Paris, available at: www.oecd.org/dataoecd/51/8/1913021.pdf.

Ohm, P. (2010), "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization", *UCLA Law Review*, Vol. 57, p. 1701-1777.

Oppenheim, C., J. Stenson and R. Wilson (2004), "Studies on Information as an Asset III: Views of Information Professionals", Journal of Information Science, 30(2): 181-190, available at: http://jis.sagepub.com/content/30/2/181.full.pdf.

De Panizza, A. and G. de Prato (2009), "Streamlining Microdata for the Analysis of ICT, Innovation and Performance", Joint Research Centre, 24120 EN, December, http://ftp.jrc.es/EURdoc/JRC54908_TN.pdf.

Peirce, C. S. (1958), *Writings of Charles S. Peirce*: *A chronological edition*, A.W. Burke (Ed.) Vol. VII–VIII. Indiana University Press, Bloomington.

Peirce, C. S. (1931), *Collected papers of Charles Sanders Peirce*, C. Hartshorne and P. Weiss (Eds.), Vol. I–VI, Harvard University Press, Cambridge, MA.

Pfitzmann, A. and M. Hansen (2010), "A terminology for talking about privacy by data minimization:: Anonymity, Unlinkability, Undetectability, Unobservability, Pseudonymity, and Identity Management", v0.43, 10 August, http://dud.inf.tu-dresden.de/Anon_Terminology.shtml.

Poli, R. (2001), ALWIS: Ontology for knowledge engineers, Doctoral dissertation, University of Utrecht, the Netherlands, avaialble at: www.academia.edu/3120636/Alwis_Ontology_for_Knowledge_Engineers.

Rao, L. (2011), "Index And Khosla Lead $11M Round In Kaggle, A Platform For Data Modeling Competitions", TechCrunch.com, 2 November, available at: http://techcrunch.com/2011/11/02/index-and-khosla-lead-11m-round-in-kaggle-a-platform-for-data-modeling-competitions/.

Repo, A. J. (1986), "The dual approach to the value of information – an appraisal of use and exchange values", Information Processing & Management, 22(5): 373-383, available at: www.sciencedirect.com/science/article/pii/0306457386900725.

Repo, A. J. (1989), "The Value of Information – Approaches in Economics, Accounting, and Management Science", Journal of the American Society for Information Science, 40(2):68-85, available at: http://onlinelibrary.wiley.com/doi/10.1002/%28SICI%291097-4571%28198903%2940:2%3C68::AID-ASI2%3E3.0.CO;2-J/abstract.

Rotella, P. (2012), "Is Data The New Oil?", Forbes, 02 April, available at: www.forbes.com/sites/perryrotella/2012/04/02/is-data-the-new-oil/.

Russel, S. and P. Norvig (2009), *Artificial Intelligence: A Modern Approach*, Third Edition, 11 December, Prentice Hall.

Russom, P. (2007), "BI Search and Text Analytics: New Additions to the BI Technology Stack", TDWI Best Practices Report, TDWI, Second Quarter 2007.

Senn, J.A. (1990), *Information Systems in Management*, 4th Edition, Wadsworth Publishing Belmont, CA.

Schmitz, H. (2003), *Was ist Neue Phänomenologie?*, Koch Verlag, Rostock.

Schwartz, J. (2000), "Intel Exec Calls for E-Commerce Tax", The Washington Post, 6 June.

Scofield, M. (1998), "Issues of Data Ownership", Information Management, 1 November, http://www.information-management.com/issues/19981101/296-1.html.

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

Shilakes, C. and J. Tylman (1998), "Enterprise Information Portals: Move Over Yahoo!; the Enterprise Information Portal Is on Its Way", Merrill Lynch, 16 November.

Silver, N. (2012), *The Signal and the Noise: Why So Many Predictions Fail – but Some Don't*, The Penguin Press, New York.

Skyrme, D, (1994), "Ten Ways to Add Value to Your Business", Managing Information, 1(3):20-25, available at: www.skyrme.com/pubs/tenways.htm.

Sopek, M. (2011), "Semantic Web is NOT for machines to understand !!!", Mirek's Blog - Passionate Reading, 07 December, available at: http://sopekmir.blogspot.fr/2011/12/semantic-web-is-not-for-machines-to.html.

Sowa, J. F. (1992), "Semantic Networks", *Encyclopedia of Artificial Intelligence*, Stuart C. Shapiro (Ed.), Second Edition, Wiley, available at: www.jfsowa.com/pubs/semnet.htm.

Spiekermann, S., J. Grossklags and B. Berendt (2001), "E-privacy in 2nd Generation E-Commerce: Privacy Preferences Versus Actual Behavior", *Proceedings of the ACM Conference on Electronic Commerce*, pp. 38-47.

Stiglitz, J., P. Orszag and J. Orszag (2000), "Role of Government in a Digital Age", Computer and Communications Industry Association, October, www.ccianet.org/CCIA/files/ccLibraryFiles/Filename/000000000086/govtcomp_report.pdf.

Stonier, T. (1993), *The wealth of information*, Thames/Methuen, London.

Stonier, T. (1997), *Information and meaning – An evolutionary perspective*, Springer, Berlin.

Suhoy, T. (2009), "Query indices and a 2008 downturn: Israeli data", Technical Report, Bank of Israel, 2009, www.bankisrael.gov.il/deptdata/mehkar/papers/dp0906e.pdf.

Surowiecki, J. (2011), "A Billion Prices Now", *The New Yorker*, 30 May.

Swartz, A. (2013), *Aaron Swartz's A Programmable Web: An Unfinished Work: Synthesis Lectures on the Semantic Web: Theory and Technology*, February, doi:10.2200/S00481ED1V01Y201302WBE005, Morgan & Claypool Publishers.

Thorp, J. (2012), "Big Data Is Not the New Oil", HBR Blog Network, 30 November, available at: http://blogs.hbr.org/cs/2012/11/data_humans_and_the_new_oil.html#disqus_thread.

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

Tuomi, I. (2009), "Theories of Open Innovation", last time accessed: 07 October 2013, available at: www.meaningprocessing.com/personalPages/tuomi/articles/TheoriesOfOpenInnovation.pdf.

Ubaldi, B. (2013), "Open Government Data: Towards Empirical Analysis of Open Government Data Initiatives", *OECD Working Papers on Public Governance*, No. 22, OECD Publishing. doi: 10.1787/5k46bj4f03s7-en

United Kingdom Government (2012), Open Data White Paper, June, available at: http://data.gov.uk/sites/default/files/Open_data_White_Paper.pdf.

United Nation [UN] Global Pulse (2012), "Big Data for Development: Opportunities & Challenges", Global Pulse White Paper, May, available at: www.unglobalpulse.org/sites/default/files/BigDataforDevelopment-UNGlobalPulseJune2012.pdf.

Unwin, A., M. Theus, and H. Hofmann (2006), *Graphics of Large Datasets: Visualizing a Million*, Springer.

Warden, P. (2011), "Why you can't really anonymize your data", O'Reilly Strata, 17 May, http://strata.oreilly.com/2011/05/anonymize-data-limits.html.

Wellisch, H.H. (1996), *Abstracting, indexing, classification, thesaurus construction: A glossary.* American Society of Indexers, Port Aransas, TX.

Zax, D. (2011), "Is Personal Data the New Currency?", MIT Technology Review, 30 November, available at: www.technologyreview.com/view/426235/is-personal-data-the-new-currency/.

Zeleny, M. (1987), "Management Support Systems: Towards Integrated Knowledge Management", Human Systems Management, 7(1987)1, 59-70, DOI 10.3233/HSM-1987-7108, avai

Zins, C. (2007a), "Conceptual Approaches for Defining Data, Information, and Knowledge", Journal of the American Society for Information Science and Technology, 58(4):479-493, 1 February, Wiley InterScience, DOI: 10.1002/asi.20508, available at: www.success.co.il/is/zins_definitions_dik.pdf.

Zinow, R. (2012), "Big Data, Mobile and Cloud combined – the new paradigm shift?", SAP, Presentation at the ECD Conference, 23 May, http://ecd-conference.de/wp-content/blogs.dir/46/files/2011/03/Zinow_SAP_1545_2.pdf.

Zins, C. (2007b), "Conceptions of Information Science", Journal of the American Society for Information Science and Technology, 58(3):335-350, 1 February, Wiley InterScience, DOI: 10.1002/asi.20507, available at: www.success.co.il/is/zins_conceptsof_is.pdf.

NOTES

[1] Business model offers free service to customers, and a premium level of the service is available for a fee (see for example Dropbox).

[2] Knowledge-based capital (KBC) comprises a range of assets. These assets create future benefits for organisations but, unlike machines, equipment, vehicles and structures, they are not physical. This non-tangible form of capital is, increasingly, the largest form of business investment and a key contributor to growth in advanced economies (OECD, 2013a). One widely accepted classification groups KBC into three types: (i) *computerised information* (software and databases); (ii) innovative property (patents, copyrights, designs, trademarks); and (iii) economic competencies (including brand equity, firm-specific human capital, networks of people and institutions, and organisational know-how that increases enterprise efficiency) (see Corrado et al., 2009).

[3] In 2010, the OECD launched an organisation wide (horizontal) project on *New Sources of Growth (NSG): Knowledge-Based Capital*. The outcomes of the first phase (*NSG I*) provided evidence of the impact on growth, and the associated policy implications, of KBC (OECD, 2013a). The second phase of the project *(NSG II)*, which started in 2013 under the auspices of the Committee for Information, Computer and Communications Policy (ICCP), aims to study in more depth those issues highlighted under NSG I as those that need to be further analysed. This included among other the analysis of the increasing social and economic value of data and the policy implications (see OECD, 2013b). For more information about the OECD horizontal project see http://oe.cd/kbc.

[4] This and other key properties of data will be elaborated in more detail later in the report.

[5] See for example OECD (1998), which defines knowledge-based economies as "economies which are directly based on the production, distribution and use of knowledge and information".

[6] See for example Hess and Ostrom (2007) according to whom "knowledge […] refers to all intelligible ideas, information, and data in whatever form in which it is expressed or obtained". See also Daniel Bell, cited in Cleveland (1982), who defines information as "data processing in the broadest sense" and knowledge as "an organized set of statements of facts or ideas […] communicated to other".

[7] Estimates suggest that the share of unstructured data in businesses could be as high as 80% to 85% and it remains largely unexploited or underexploited (Shilakes and Tylman, 1998; Russom, 2007; ComputerWeekly, 2013). See section below on the data taxonomy for more details about "structured" vs. "unstructured" data.

[8] As Speier, et al. (2007) explain: "Information overload occurs when the amount of input to a system exceeds its processing capacity. Decision makers have fairly limited cognitive processing capacity. Consequently, when information overload occurs, it is likely that a reduction in decision quality will occur."

[9] See also Raphael Capurro (cited in Zins, 2007a), who has argued that knowledge is "no-thing" (contrary to "information-as-thing" as suggested by Buckland (1991a), and Poli (2001 cited in Zins, 2007a), who has highlighted that knowledge is "not entirely reducible to information and data".

[10] Information science is an interdisciplinary field related to library and documentation science, computer science, cybernetics, and philosophy (epistemology). In contrast to these disciplines, however, information science is primarily concerned with the analysis, collection, classification, manipulation, storage, retrieval, movement, and dissemination of information. Information science also deals with data and knowledge, which are considered together with information as interrelated building blocks of information science (see Hawkins, 2001; Zins, 2007a for a literature review on the evolution of information science).

[11] This is illustrated through metaphors like "data is the seed, information is the crop, and knowledge is the harvest". But again these metaphors should be take with caution.

[12] The hierarchy is also known as the knowledge pyramid or in the knowledge management domain as the DIKW (data, information, knowledge, wisdom) hierarchy, but with an additional fourth concept "wisdom". For more details about the DIKW hierarchy see Zeleny (1987) and Ackoff (1989).

[13] See philosophy (epistemology), cybernetics, and information science.

[14] The spillover effects are even stronger due to the fact that knowledge tend to be abundant; although "what is scarce is the capacity to use them in meaningful ways" (OECD, 1996).

[15] See Zins (2007a) citing Elsa Barber, Thomas A. Childers, and Michael Buckland (1991a; 1991b).

[16] See Zins (2007a) citing Shifra Baruchson-Arbib.

[17] See Zins (2007a) citing Michal Lorenz, who defines knowledge as "tacitly or consciously grasped and interiorized content of information related and meaningfully integrated into a unifying frame of experience among other information contents interiorized in the same way, the complex of which reflects subjective understanding of environment".

[18] See Zins (2007a) citing Carol Tenopir. The term "internalisation" (as well as "externalisation") should be used with caution as they are based on implicit assumptions that have been challenged by some scholars. This includes for example the assumed existence of an outside vs. inside world. Some scholars such as Schmitz (2003), however, have challenged these concepts as modern mental constructs.

[19] See Zins (2007a) citing Lena Vania Pinheiro, Anna da Soledade Vieira, and Aldo Barreto respectively. See also Machlup (1983, cited in Hess and Ostrom, 2007), who defines knowledge as the "assimilation of the information and understanding of how to use it".

[20] See Zins (2007a) citing Aldo Barreto and Jo Link-Pezet.

[21] See Zins (2007a) citing Jo Link-Pezet.

[22] See Zins (2007a) citing Aldo Barreto.

[23] See Zins (2007a) citing Caroline Haythornthwaite.

[24] For Carol Tenopir, for example, knowledge is "internalized or understood information that can be used to make decisions" (Zins, 2007a).

[25] See Zins (2007a) citing Thomas A. Childers and Yishan Wu respectively.

[26] See Zins (2007a) citing Thomas A. Childers, Haidar Moukdad, and Anna da Soledade Vieira.

[27] See Zins (2007a) citing H. M. Gladney.

[28] See Zins (2007a) Maria Pinto, who refers at Belkin and Roberston (1976) and Blair (2002).

[29] See also Michael Buckland's explanation (cited in Zins, 2007a): "By extension the word 'knowledge' is used more loosely for (1) what social groups know collectively; and (2) what is in principle knowable because it has been recorded somehow and could be recovered even though, at any given time, no individual knows (or remembers) it".

[30] The increased capacity to externalise knowledge could be a source for *capital-biased technological change*, which tends to "shift the distribution of income away from workers to the owners of capital" (Krugman, 2012a; 2012b).

[31] Some definitions see information, for example, as a "human-exclusive" phenomenon, while others ascribe information also to non-human biological and/or physical systems (see Gleick, 2011). Hjørland (2002, cited in Zins, 2007a), for example, defines information in terms of biological signals and mechanisms.

[32] See Zins (2007a) citing Hanne Albrechtsen, Michael Buckland, Raya Fidel, and others.

[33] See Zins (2007a) citing Hanne Albrechtsen and Lena Vania Pinheiro. Information is therefore perceived as the "little atoms of content" (Nunberg, 1996).

[34] See Zins (2007a) citing Hanne Albrechtsen, Elsa Barber, Raphael Capurro, and Maria Pinto.

[35] See Zins (2007a) citing Maria Pinto

[36] See Zins (2007a) citing Aldo Barreto.

[37] See Berners-Lee et al. (2001).

[38] See Zins (2007a) citing Quentin L. Burrell, Charles H. Davis, and Caroline Haythornthwaite. See also Morris (1938).

[39] According to Raphael Capurro (cited in Zins, 2007a): "A 'message' is a 'meaning offer' while 'information' refers to the selection within a system and 'understanding' to the possibility that the receiver integrates the selection within his/her preknowledge – constantly open to revision i.e. to new communication – in accordance with the intention(s) of the sender. The receiver mutates each time into a sender."

[40] See Zins (2007a) citing Haidar Moukdad and Ronald Rousseau.

[41] See Berners-Lee et al. (2001).

[42] See Zins (2013a) citing Michal Lorenz, and Roberto Poli.

[43] As Lucky (1998) explains: "Perhaps information itself is best described in terms of *organization*". See also Zins (2007a) citing Michel Menou, Scott Seaman, and Anna da Soledade Vieira.

[44] Senn (1990 cited in Engelsman, 2009), for example, defines information as "data presented in a form that is meaningful to the recipient". It would be however wrong to conclude that data is "unprocessed information" as in Hey (2004), because information *per se* has meaning, while data does not.

[45] *Google Now*, for example, is a voice enabled intelligent personal assistant, which can autonomously identify and process information such as flight information, hotel reservations and restaurant reservations from personal email accounts.

[46] See Zins (2013a) citing Quentin L. Burrell, Charles H. Davis, Birger Hjørland, Donald Kraft, Charles Oppenheim, Roberto Poli, and Irene Wormell.

[47] One could argue at this point that the structure of the data thus reflects its embedded meaning, and as a consequence that the interpretation of data can be based on a structural analysis of the data. This is in line with, for example, Merelli and Rasetti (2013), who explore the idea of a "global geometric vision of data space" that could allow to explore their "structure and hidden information patterns". As a result Merelli and Rasetti (2013) "propose an approach that exploits topological methods for classifying global information into equivalence classes and regular languages for describing the corresponding automaton as element an of hidden complex system."

[48] Hadoop is an open source programming framework for distributed data management, that was inspired by a paper by Google employees Dean and Ghemawat (2004). It was funded initially by Yahoo!, deployed and further developed by Internet firms such as Amazon, Facebook, and LinkedIn, then offered by traditional providers of databases and enterprise servers such as IBM, Oracle, Microsoft, and SAP as part of their product lines, and is now used across the economy for data-intensive operations

[49] Luhn (1958)'s definition is based on Webster's dictionary definition of (i) intelligence: "the ability to apprehend the interrelationships of presented facts in such a way as to guide action towards a desired goal", and (ii) business: "a collection of activities carried on for whatever purpose, be it science, technology, commerce, industry, law, government, defense, et cetera."

[50] This is also reflected in the French translation "donnée".

[51] See Zins (2013a) citing Elsa Barber, Nicolae Dragulanescu, H.M. Gladney, Yves François Le Coadic, and Haidar Moukdad. See also Debons, et al. (1988).

[52] See Zins (2013a) citing H.M. Gladney, Caroline Haythornthwaite, Gordana.Dodig-Crnkovi, and Richard Smiraglia talking about "evidence". See also Stonier (1993, 1997).

[53] See Zins (2013a) citing .Hamid Ekbia, Gordana.Dodig-Crnkovi and William Hersh.

[54] See Zins (2013a) citing Caroline Haythornthwaite and Yishan Wu.

[55] See Zins (2013a) citing Donald Kraft and Gordana.Dodig-Crnkovi.

[56] See Zins, (2013a) citing Caroline Haythornthwaite.

[57] See Zins (2013a) citing Aldo Barreto and Caroline Haythornthwaite.

[58] See Zins (2013a) citing Aldo Barreto, Anthony Debons, Nicolae Dragulanescu, and Haidar Moukdad.

[59] See Zins (2013a) citing Yishan Wu.

[60] See Zins (2013a) citing Elsa Barber, and Richard Smiraglia.

[61] See Zins (2013a) citing Elsa Barber.

[62] See Mayer-Schonberger and Cukier (2013)

[63] See Zins (2013a) citing Elsa Barber.

64 The fact that information is often seen as "data with meaning", "interpreted data", or "structured data", explains why information and data are often used as synonyms. In many cases, there is no need to make the difference, namely when the information is perfectly reflected in the data, so that with the data one gets the information. However, there are many cases where this is not true, for example, when a user is not able to extract any meaning out of the data, either because he or she lacks the contextual knowledge and/or the analytic capacity (skills and technologies) or just simply because the data is encrypted and he/she does not have the key to decrypt the data. Furthermore, the context dependency implies that the same data set can lead to different information being extracted if for example used in different contexts.

65 See Zins (2013a) citing Nicolae Dragulanescu and William Hersh.

66 In computer science, digital data is measured through a "unit of information", which is also used to measure the capacities of data storage system or communication channel. The unit is the "bit", a contraction of binary digit (see Lyman and Varian's (2003) study on "How Much Information?").

67 More than 30 million interconnected sensors are now deployed worldwide, in areas such as security, health care, the environment, transport systems or energy control systems, and their numbers are growing by around 30% a year according to MGI (2011). This trend is confirmed by available sales figures. According to the Semiconductor Industry Association for instance, sensors and actuators are the fastest-growing semiconductor segment with growth in revenue of almost 16% (USD 8 billion) in 2011.

68 The other principles of the OECD (1997) *Council Recommendation concerning Guidelines for Cryptography Policy* include: (iv) Standards for Cryptographic Methods; (v) Protection of Privacy and Personal Data; (v) Lawful Access; (vi) Liability; and (vii) International Co-operation.

69 "Pseudonym" comes from Greek "pseudonumon" meaning "falsely named" (pseudo: false; onuma: name). Thus, it means a name other than the "real name" (Pfitzmann and Hansen, 2010).

70 NLP involves at one extreme, "counting word frequencies to compare different writing styles. At the other extreme, [it] involves 'understanding' complete human utterances, at least to the extent of being able to give useful responses to them" (Bird et al., 2009).

71 See Zins (2013a) Birger Hjørland, Raphael Capurro, and Gordana.Dodig-Crnkovi.

72 See http://www.oecd.org/sti/ict/broadband

73 See http://www.oecd.org/sti/broadband/oecdbroadbandsubscribercriteria2010.htm.

74 The terminology used to draw the boundary between (public sector) content and data may seem imperfect, because digital content (including e.g. cultural archives, artistic works and so on), are ultimately stored on disk drives and/or exchanged over networks in the form of digital data. However, as argued in the previous section on "Data", although digital, this data cannot be directly indexed, searched, or otherwise analysed by software as public sector data can be, and because it has it is unstructured and has not been "datafied". But it is not the form (structured or not) that matters here, because in the long run the boundaries between structured and unstructured are blurring as data analytic capacities increases and the cost of extracting information from unstructured data sources such as document and even multimedia content is falling towards that from structured data sources such as databases. What matters is the fact that PSC typically applies to cultural content held by cultural establishments.

75 See http://stats.oecd.org/glossary/detail.asp?ID=2199.

[76] In countries such as the United States and the United Kingdom, the concept of government is used more broadly as a synonym for the public sector. As a consequence "government data" refers to public sector data.

[77] See http://stats.oecd.org/glossary/detail.asp?ID=2130.

[78] Many governments have established open data websites/portal where they published government data which are not open as defined in this paper (see Ubaldi, 2013).

[79] "Open source as a development model, promotes a) universal access via free license to a product's design or blueprint, and b) universal redistribution of that design or blueprint, including subsequent improvements to it by anyone." (see http://en.wikipedia.org/wiki/Open_source).

[80] The meeting was organised by Tim O'Reilly of O'Reilly Media and Carl Malamud of Public.Resource.Org. See https://public.resource.org/8_principles.html (last visited on 07 November 2013).

[81] See www.nationalarchives.gov.uk/doc/open-government-licence/version/2/.

[82] See http://storage.googleapis.com/books/ngrams/books/datasetsv2.html.

[83] See www.openstreetmap.org/copyright.

[84] Business model offers free service to customers, and a premium level of the service is available for a fee (see for example Dropbox).

[85] Open innovation is a concept that describes the "use of purposive inflows and outflows of knowledge to accelerate internal innovation, and expand the markets for external use of innovation". This includes proprietary-based business models that make active use of licensing, collaborations joint ventures, etc… Here "open" is understood to denote the arms' length flow of innovation knowledge across the boundaries of individual organisations (OECD, 2012).

[86] In November 2011, Kaggle raised USD 11 million from investors, including 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. Hal Varian, Google's Chief Economist, described Kaggle as "a way to organize the brainpower of the world's most talented data scientists and make it accessible to organizations of every size" (Rao, 2011).

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

[88] RFID may be considered as one of a group of automatic identification and data capturing technologies which also includes bar codes, biometrics, magnetic stripes, optical character recognition, smart cards, voice recognition and similar technologies.

[89] For data linkage to be effective, however, a positive coordination of sample design across surveys should be adopted, while negative sampling coordination across surveys (i.e. any company selected for one survey would be excluded from other surveys) and over time (i.e. the same company would not be selected again for a certain number of years) has been extensively used in Europe and elsewhere to reduce the statistical burden on respondents (OECD, 2013g).

[90] "Researchers must submit a research proposal and must be an employee of a research entity […] or a senior (Ph.D.) students under guidance of a supervisor employed by the research entity" (Eurostat, 2013).

[91] According to Hoberman (2009), "a data model is a wayfinding tool for both business and IT professionals, which uses a set of symbols and text to precisely explain a subset of real information to improve communication within the organization and thereby lead to a more flexible and stable application environment."

[92] For example, at the OECD-APEC (2012) workshop, Anticipating the Needs of the 21st Century Silver Ageing Economy, held 12-14 September 2012 in Tokyo, Japan, participants concluded that the multi-factorial nature of Alzheimer's disease (AD) will require sophisticated computational capabilities to analyse big streams of behavioural, genetic, environmental, epigenetic and clinical data to find patterns. In neurodegenerative research, many organisations are building big data repositories and contributing to the development of databases and global data-sharing networks. In the United States alone, the Alzheimer's Disease Neuroimaging Initiative and the Parkinson's Disease (PD) Progression Markers Initiative gather brain images and biological fluids from people with or at risk for AD and PD, respectively. The US National Alzheimer's Coordinating Center has amassed longitudinal records from more than 25 000 people, and recently started assessments for fronto-temporal dementia as well. Records from those who inherited an AD-linked gene are part of the Dominantly Inherited Alzheimer Network.

[93] The *Dublin Core metadata terms* were endorsed in *IETF (Internet Engineering Task Force) RFC 5013* and *ISO (International Organization for Standardization) Standard 15836-2009*.

[94] See also Dumbill (2012a), for which "big data" is "data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or doesn't fit the strictures of your database architectures. To gain value from this data, you must choose an alternative way to process it".

[95] This definition originated from the META Group (now part of Gartner) in 2001 (see Laney, 2001).

[96] Business model offers free service to customers, and a premium level of the service is available for a fee (see for example Dropbox).

[97] The key concept here is information. Data and knowledge are seen as manifestations of information.