INNOVATION, ENTREPRENEURSHIP, MANAGEMENT SERIES

SMART INNOVATION SET

Volume 3

Big Data, Open Data and Data Development

Jean-Louis Monino
Soraya Sedkaoui

iSTE
WILEY

Foreword

The world has become a digitalized place, and technological advancements have multiplied the ways of accessing, processing and disseminating data. Today, new technologies have reached a point of maturity. Data is available to everyone throughout the planet. In 2014, the number of Internet users in the world was 2.9 billion, which is 41% of the world population. The thirst for knowledge can be perceived in the drive to seize this wealth of data. There is a need to inquire, inform and develop data on a massive scale. The boom in networking technologies – including the advent of the Internet, social networks and cloud computing (digital factories) – has greatly increased the volume of data available. As individuals, we create, consume and use digital information: each second, more than 3.4 million emails are sent throughout the world. That is the equivalent of 107,000 billion emails per year, with over 14,600 per person per year, although more than 70% of them are junk mail. Millions of links are shared on social networks, such as Facebook, with over 2.46 million shares every minute. The average time spent on the Internet is over 4.8 hours per day on a computer and 2.1 hours on a cellphone. The new immaterial substance of “data” is produced in real-time. It arrives in a continuous stream flowing from a variety of generally heterogeneous sources. This shared pool of all kinds of data (audio, video, files, photos, etc.) is the site of new activities aimed at analyzing the enormous mass of information. It thus becomes necessary to adapt and develop new approaches, methods, forms of knowledge and ways of working, all of which involve new paradigms and stakes as a new ordering system of knowledge must be created and put into place. For most companies, it is

difficult to manage this massive amount of data. The greatest challenge is interpreting it. This is especially a challenge for those companies that have to use and implement this massive volume of data, since it requires a specific kind of infrastructure for the creation, storage, treatment, analysis and recovery of the same. The greatest challenge resides in “developing” the available data in terms of quality, diversity and access speed.

Alain IOZZINIO
E-PROSPECTS Manager
January 2016

Key Concepts

Before launching into the main text of this book, we have found it pertinent to recall the definitions of some key concepts. Needless to say, the following list is not exhaustive:

– *Big Data*: The term Big Data is used when the amount of data that an organization has to manage reaches a critical volume that requires new technological approaches in terms of storage, processing, and usage. Volume, speed, and variety are usually the three criteria used to qualify a database as “Big Data”.

– *Cloud computing*: This term designates a set of processes that use computational and/or storage capacities from remote servers connected through a network, usually the Internet. This model allows access to the network on demand. Resources are shared and computational power is configured according to requirements.

– *Competitive intelligence*: It is the set of coordinated information gathering, processing and dissemination activities useful for economic actors. According to the Marte Report, competitive intelligence can be defined as the set of coordinated information research, processing and dissemination actions aimed at exploiting it for the purpose of economic actors. This diverse set of actions is carried out legally with all data protection guarantees necessary to preserve the company’s assets, with the highest regard to quality, deadlines and cost. Useful information is needed at the company or partnership’s different decision-making levels in order to design and put into place strategies and techniques coherently

aimed at achieving company-defined objectives and improving its position in the competitive environment in which it operates. These kind of actions take place in an uninterrupted cycle that generates a shared vision of company objectives.

– *Data*: This term comprises facts, observations and raw information. Data itself has little meaning if it is not processed.

– *Data analysis*: This is a class of statistical methods that makes it possible to process a very large volume of data and identify the most interesting aspects of its structure. Some methods help to extract relations between different sets of data and thus draw statistical information that makes it possible describe the most important information contained in the data in the most succinct manner possible. Other techniques make it possible to group data in order to identify its common denominators clearly, and thereby understand them better.

– *Data governance*: It constitutes a framework of quality control for management and key information resource protection within a company. Its mission is to ensure that the data is managed in accordance with the company's values and convictions, to oversee its quality and to put mechanisms into place that monitor and maintain that quality. Data governance includes data management, oversight, quality evaluation, coherence, integrity and IT resource security within a company.

– *Data journalism*: The term designates a new form of journalism based on data analysis and (often) on its visual representation. The journalist uses databases as his or her sources and deduces knowledge, meaningful relationships or intuitions from them that would not be accessible through traditional research methods. Even

when the article itself stands as the main component of the work, illustrating ideas through graphs, diagrams, maps, etc., is becoming more important day by day.

– *Data mining*: Also referred to as knowledge discovery from data, is intended for the extraction of knowledge from large amounts of data using automatic or semi-automatic methods. Data mining uses algorithms drawn from disciplines as diverse as statistics, artificial intelligence and computer science in order to develop models from data; that is, in order to find interesting structures or recurrent themes according to criteria determined beforehand and to extract the largest possible amount of knowledge useful to companies. It groups together all technologies capable of analyzing database information in order to find useful information and possible significant and useful relationships within the data.

– *Data reuse*: This practice consists of taking a dataset in order to visualize it, merge it to other datasets, use it in an application, modify it, correct it, comment it, etc.

– *Data science*: It is a new discipline that combines elements of mathematics, statistics, computer science and data visualization. The objective is to extract information from data sources. In this sense, data science is devoted to database exploration and analysis. This discipline has recently received much attention due to the growing interest in Big Data.

– *Data visualization*: Also known as “data viz”, it deals with data visualization technology, methods and tools. It can take the form of graphs, pie-charts, diagrams, mappers, timelines or even original graphic representations. Presenting data through illustrations makes it easier to read and understand.

– *data.gouv.fr*: The French government's official website for public data, which was launched on December 5th 2011 by Mission Etalab. In December 2013, data.gouv.fr was transformed deeply through a change in both the site's structure and its philosophy. It has, without doubt, become a collaborative platform oriented towards the community, which has resulted in better reuse of public data.

– *Dataset*: Structured and documented collection of data on which reusers rely.

– *Etalab*: This is a project proposed in the November 2010 Riester Report and put into place in 2011 which is responsible for implementing the French government's open data policy, as well as for establishing an almanac of French public data: data.gouv.fr.

– *Hadoop*: Big Data software infrastructure that includes a storage system and a distributed processing tool.

– *Information*: It consists of interpreted data and has discernible meaning. It describes and answers questions like "who?", "what?", "when?" and "how many?".

– *Innovation*: It is recognized as a source of growth and competitiveness. The Oslo Manual distinguishes between four types of innovation:

- *Product innovation*: Introduction of a new product. This definition includes significant improvements to technical conditions, components or materials, embedded software, user friendliness or other functional characteristics.

- *Process innovation*: Establishing a new production or distribution method, or significantly improving an existing one. This notion involves significant changes in techniques, material and/or software.

- *Marketing innovations*: Establishing a new marketing method involving significant changes in a product's design, conditioning, placement, promotion or pricing.

- *Organizational innovation*: Establishing a new organizational method in practices, workplace organization or company public relations.

– *Interoperability*: This term designates the capacity of a product or system with well-known interfaces to function in sync with other existing or future products or systems, without access or execution restrictions.

– *Knowledge*: It is a type of know-how that makes it possible to transform information into instructions. Knowledge can either be obtained through transmission from those who possess it, or by extraction from experience.

– *Linked Open Data (LOD)*: This term designates a web-approach proposed by supporters of the "Semantic Web", which describes all data in a way such that computers can scan it, and which links to it by describing its relationships, or by making it easier for the data to be related. Open public data is arranged in a "Semantic Web" format, such that its items have a unique identifier and datasets are linked together by those identifiers.

– *Open innovation*: It is defined as increased use of information and knowledge sources external to the company, as well as the multiplication of marketing channels for intangible assets with the purpose of accelerating innovation.

– *Open knowledge foundation network*: A British non-profit association that advocates for open data. It has most famously developed CKAN (open source data portal

software), a powerful data management system that makes data accessible.

– *Open data*: This term refers to the principle according to which public data (gathered, maintained and used by government bodies) should be made available to be accessed and reused by citizens and companies.

– *Semantic Web*: This term designates a set of technologies seeking to make all web resources available, understandable and usable by software programs and agents by using a metadata system. Machines will be able to process, link and combine a certain amount of data automatically. The semantic web is a set of standards developed and promoted by W3C in order to allow the representation and manipulation of knowledge by web tools (browsers, search engines, or dedicated agents). Among the most important, we can cite:

- *RDF*: a conceptual model that makes it possible to describe any dataset in the form of a graph in order to create knowledge bases;

- *RDF Schema*: language that makes it possible to create vocabularies, a set of terms used to describe things;

- *OWL*: A language that makes it possible to create ontologies and more complex vocabularies that serve as support for logical processing (interfaces, automatic classification, etc.);

- *SPARQL*: A query language for obtaining information from RDF graphs.

– *Semi-structured information*: It is worth noting that the boundary between structured information and unstructured information is rather fuzzy, and that it is not always easy to classify a given document into one

category or the other. In such a case, one is no doubt dealing with semi-structured information.

– *Smart Data*: The flood of data encountered by ordinary users and economic actors will bring about changes in behavior, as well as the development of new services and value creation. This data must be processed and developed in order to become “Smart Data”. Smart Data is the result of analysis and interpretation of raw data, which makes it possible to effectively draw value from it. It is, therefore, important to know how to work with the existing data in order to create value.

– *Structured information*: It can be found, for example, in databases or in programming languages. It can thus be recognized by the fact that it is arranged in a way such that it can be processed automatically and efficiently by a computer, but not necessarily by a human. According to Alain Garnier, the author of the book *Unstructured Information in Companies*, “information is structured when it is presentable, systematic, and calculable”. Some examples include forms, bills, pay slips, text documents, etc.

– *Text mining*: This is a technique that makes it possible to automate processing of large volumes of text content to extract the main tendencies and statistically assess the different subjects they deal with.

– *Tim Berners-Lee*: He is the co-inventor of the Semantic Web. He is very active and engaged in data.gov.uk. In particular, he has defined a five star ranking system to measure the Semantic Web openness level for putting a dataset online.

– *Unstructured information*: Unlike structured information, unstructured information constitutes the set of information for which it is impossible to find a

predefined structure. It is always intended for humans, and is therefore composed mainly of text and multimedia documents, like letters, books, reports, video and image collections, patents, satellite images, service offers, resumes, calls for tenders, etc. The list is long.

– *Web 1.0*: This term refers to the part of the Internet that makes it possible to access sites composed of web pages connected by hyperlinks. This Web was created at the beginning of the 1990s. It creates a relationship between an edited site that publishes content or services and Internet users who visit it and who surf from site to site.

– *Web 2.0*: This term designates the set of techniques, functions and uses of the World Wide Web that have followed the original format of the Web. It concerns, in particular, interfaces that allow users with little technical training to appropriate new Web functions. Internet users can contribute to information exchanges and interact (share, exchange, etc.) in a simple manner.

– *Web 3.0*: (also known as the Semantic Web). This is a network that allows machines to understand semantics, which is to say the meaning of information published online. It expands the network of Web pages understandable by humans by adding metadata that is understandable by a machine and that creates links between content and different pages, which in turns allows automatic agents to access the Web in a more intelligent manner and to carry out some tasks in the place of users.

Introduction

Today data comes from everywhere: GPS tracking, smartphones, social networks where we can share files, videos and photos, as well as online client transactions made possible through the intermediary of credit cards. Of the 65 million people in France, 83% are Internet users and 42% (or 28 million) are on Facebook. More than 72 million telephones are active, and the French people spend a daily average of 4 hours online. Mobile phone users spend over 58 minutes online, and 86% of the population is on a social network. The French people spend over 1.5 hours per day on social networks.

Developing this massive amount of data and making access to it possible is known as “Big Data”. This intangible data comes in a constant stream and its processing is especially challenging in terms of knowledge extraction. This is why new automatic information extraction methods are put into place such as, for example, “data mining” and “text mining”. These are the sorts of processes behind radical transformations in the economy, marketing and even politics. The amount of available data will increase greatly with the appearance of new connected objects on the market that are going to be used more and more.

Some objects we use in our daily lives are already connected: cars, television sets and even some household appliances. These objects have (or will one day have) a chip designed to collect and transfer data to their users through a computer, a tablet or a smartphone. More importantly, these objects will also be able to communicate with one another! We will be able to control equipment in our homes and in our car by simply logging onto our smartphone or

some other device. This phenomenon is known as the "Internet of Things".

Box I.1. *Zero marginal cost society*

The American economist, Jeremy Rifkin, predicted the development of a new society of wealth and abundance brought about by technology, especially by the Internet of Things and 3D printing. New technologies would modify socio-economic relationships to the point of significantly reducing profits for capitalist enterprises. In the world of the Internet, the advent of the *zero marginal cost society* has already taken place. As information has become dematerialized and as it has become possible to reproduce and distribute it with near-zero marginal costs, radical changes have come about in these industries' business models.

This phenomenon has attracted the interest of operational decision-makers (marketing managers, finance chiefs, etc.) seeking to benefit from the immense potential involved in analyzing data hosted by companies in real-time. In order to meet the Big Data challenge, measures must be taken, including incorporating tools that make more restrictive data processing possible and actors capable of analyzing that data. This will only be possible if people become more aware of the benefits of "data development". When databases are organized, reorganized and processed by statistical methods or econometric modeling, they become knowledge.

For a company, it is essential to have access to more and more data about the environment in which it operates. This will make it possible to scrutinize not classes of behavior, but individual cases. This explains why this revolution has

brought with it the emergence of so-called “start-up” companies whose objective is to process the data known as Big Data automatically. We certainly find ourselves in front of one of the elements of what some people are calling the “new industrial revolution”. The Internet and the digital and connected objects have opened up new horizons in a wide array of fields.

Box I.2. *A data access example*

Data access makes it possible to enrich quantitative and qualitative analyses. Client contacts can be analyzed through data collected by a call center. This kind of product can also be offered in a limited quantity, as does e-prospects. It is necessary to develop that data by exploring the content of emails and voice calls, and to match that information with browsing activities on the company website. Beyond that, it is also possible to study messages exchanged on social networks (Facebook, Twitter, LinkedIn, etc.) in order to identify new trends or to identify the products that are being most talked about.

E-prospects[1] has developed an innovative service based on a community platform that makes it possible to buy and sell contacts and meetings through a range of products and countries.

Research and development with the University of Montpellier's TRIS laboratory[2].

The program is supported by the Languedoc-Roussillon region and by the "Transferts L-R" program[3].

Example I.1. *The startup E-PROSPECTS*

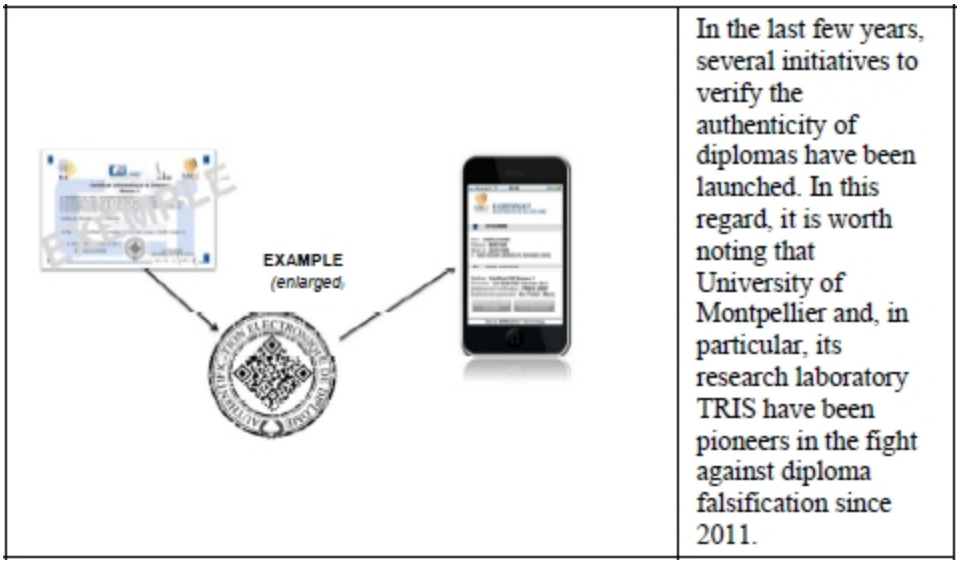
In order to get the full potential out of data, it must be available to all interested parties with no additional obstacles and at reasonably accessible costs. If data is open to users [MAT 14], other specialized data processing companies can be created. This activity will meet the needs of users without them having to develop models and equations themselves.

Open Data, beyond its economic and innovative potential, involves a philosophical or ethical choice[4]. Data describes collective human behavior, and therefore, belongs to those whose behaviors it measures. The cultivation of these phenomena depends on the availability of data that can be communicated easily.

The Internet Age has detonated a boom in information research. Companies are flooded by the wealth of data that

results from simple Internet browsing. In other words, they are forced to purchase pertinent information to develop high added value strategies that allows them to succeed in the face of incessant changes in their business environment. Industrial strategies now rely strongly on the capacity of companies to access strategic information to better navigate their environment. This information can, thus, become the source of knew knowledge (knowledge pyramid).

The process of gathering, processing and interpreting information is not limited to defining ideas, but also consists of materializing them in order to ensure improved knowledge production that leads to innovation. Competitive intelligence allows each company to optimize its service offerings in qualitative and quantitative terms, as well as to optimize its production technology.



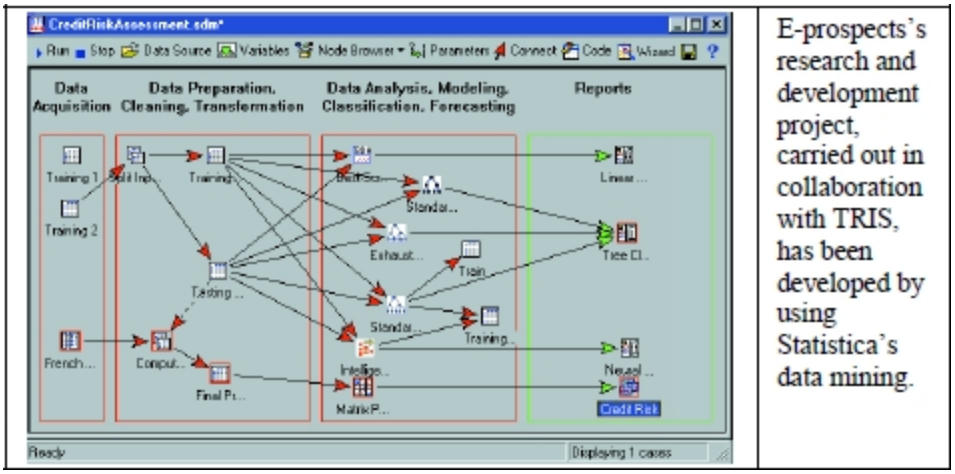
Example I.2. *Information processing C2i certificate security and massive processing*[5] *by QRCode*[6]

Beyond the advent of ICT and of increased data production, dissemination and processing speeds, another element has recently become critically important: time. The importance of time carries with it a notion of information circulation

speed. This prompts companies to rethink their strategies beyond the challenges involved in processing large volumes of data. The value of a given piece of data increases in time and depends on the variety of uses it is given.

In this sense, companies must possess the capacity to absorb the entirety of data available, which allows them to assimilate and reproduce knowledge. This capacity requires specific skills that make it possible to use that knowledge. Training “data scientists” is, therefore, indispensable in order to be able to identify useful approaches for new opportunities, or for internal data exploitation, and in order to quantify their benefits in terms of innovation and competitiveness. However, Big Data is just a single element in the new set of technical tools known as “data science”.

Data scientists have the task of extracting knowledge from company data. They hold a strategic function within the firm, and to that end, must be in command of the necessary tools. They must also be able to learn on the go and increase their understanding of data mining regularly, as the volume of data requires increasing skills and techniques.



Example I.3. *Data mining and Statistica software*

When confronted with this multiplicity of data, companies are driven to apply sophisticated processing techniques. In fact, technical competence in data processing is today a genuine strategic and useful stake for companies' competitive differentiation [BUG 11]. Processing this mass of data plays a key role for tomorrow's society because it can be applied in fields as varied as science, marketing, customer services, sustainable development, transportation, health and even education.

Big Data groups together both processing, collection, storage and even visualization of these large volumes of data. This data, thus, becomes the fuel of the digital economy. It is the indispensable raw material of one of the new century's most important activities: data intelligence. This book shows that the main challenges for Big Data revolve around data integration and development within companies. It explores data development processes within a context of strong competition.

More specifically, this book's research brings together several different fields (Big Data, Open Data, data processing, innovation, competitive intelligence, etc.). Its interdisciplinary nature allows it to contribute considerable value to research on the development of data pools in general.

I.1. The power of data

Companies are very conscious of the importance of knowledge and even more so of the way it is "managed", enriched and capitalized. Beyond all (financial, technical and other) factors, the knowledge that a company has access to is an important survival tool, whether it is market knowledge, or legal, technological and regulatory information.