Standardization and big data

Charles Huot,
Temis

Abstract:
During 2014 and 2015 under the auspices of AFNOR, a work group — 32 persons from various fields (health, energy, the armed forces, documentary services…) and experts in technology representing public authorities and research laboratories — reflected on the question of standardization and big data. As chair of the Big Data Alliance, AFNOR, the French standards organization, proposed that the author head this group along with Jean-François Legendre as rapporteur to AFNOR's Committee of Strategic Orientation. Twelve months later, a white book (*Données massives ou "Big data": Quels défis pour la normalisation?*) was published and presented to the public on 15 June 2015. The principal points in this publication are updated…

The economic, industrial, technical and societal prospects of the collection and processing of big data give rise to major opportunities and, too, risks that we must understand, anticipate and control. One characteristic of big data is that they concern all sectors and fields where breakthrough technology is present: smart cities, e-health, smart factories, smart networks, connected devices, etc. For years now, data management technology has been based on transactional procedures for storing and querying structured databases; but this no longer suffices owing to the development of the social media on the Web and mobile computing (smart phones, tablets, sensors and connected devices). New techniques work on big data and analyze them in (nearly) real time.[1]

Developing and processing big data requires leverage. Standards are tools voluntarily set for supporting firms and addressing the issues related to big data. They are levers for opening markets and building up the confidence of all parties. Their development and adoption are a competitive issue. In the ever more complex and multiform national, European and international environments, stakeholders must become aware of the leverage provided by standards. They must work out an appropriate strategy in this key field so as to place big data at the service of the parties concerned.

Herein, "big data" refers to the processing of masses (teraoctets) of information made up of (very often) heterogeneous data (from multimedia, sensors, the social media, telephones, etc.) by extremely fast means (for real-time decision-making) and, if need be, unconventional means (*e.g.*, NoSQL databases). What is at stake in the standardization of the collection, processing and use of huge sets of data, often heterogeneous and unstructured? These data come from the Internet, the social media (corporate or general), communication networks, sensors in smart networks, mobile devices, security arrangements, factories, etc. This discussion of the stakes of standardization will not be restricted to technical questions, despite their importance. Attention will be focused on standards as a means for proposing to firms and to interface users shared practices and forms of organization that are adapted to the deep changes made necessary by the business models based on big data.

[1] This article has been translated from French by Noal Mellott (Omaha Beach, France). The translation into English has, with the editor's approval, completed a few bibliographical references.

Issues, prospects and opportunities

Uses of big data have advanced thanks to several technological breakthroughs for integrating, processing, interpreting and representing heterogeneous data from various sources. These data can thus take center place in business models and add more value to activities as services are better targeted thanks to the optimization of business processes and advances in knowledge about the economic environment. Various fields of business can thus converge, and this probably marks the upsurge of new business models that redefine the relations between producers, distributors and consumers or between goods and services. These relations have thus become much more complex, and it is thus very hard for an organization to make decisions. What is new about big data is the need to use gigantic volumes of data from the many and diverse sources now accessible.

Drawbacks

Controlling the data

Firms are aware of the importance of the data that they hold. When strategic reasons forbid the sharing of data, using the data is inherently limited, since it increases the risks of information being leaked. Since the arrangements for using big data depend on heavily integrated systems and call for specialized third-party services, organizations are sensitive to events and incidents involving security. In cases when data may be shared, questions arise about licenses, intellectual property rights and the traceability of the uses to which the data are put. The right to use them leads to questions about their non-repudiation. The traceability of data and their uses is an even more important issue in the case of “open data”. This holds, in particular, for using data from cultural works and research (with the problem of identifying the authors). Practices, rules and standards already exist but in delimited fields that big data is upending.

Legal security and ethics

The handling of personal data raises questions about how to anonymize them and whether encryption should be reversible. For some parties (local authorities, financial organizations, etc.), their image might be at stake given the effects and deviations of a “Big Brother” sort related to the massive collection of data. This risk also exists for big corporations that want to control their reputation on the social media and manage customer relations. Societal concerns might also restrict the procedures for using big data or presenting the results of using them.

For a firm, the issue in using big data is competition: the possibility for making profits and staking out a position in new business fields. For public authorities, it means acquiring the capacity for optimizing operations and proposing new services to citizens. For citizens, it is the possibility to become conscious actors in the big data ecosystem and gain new services that improve the quality of life. To benefit from this new business model, it is necessary to decompartmentalize organizations and play on the diagonal nature of business processes. This implies transforming the existing organization. A firm or organization has to be capable of capturing information and, too, of realizing the worth of its own legacy of information. This means understanding and controlling the value of data (what can and cannot be shared — and under what conditions) with a full awareness of the issues related to intellectual property rights and regulations (about personal data, in particular). For public authorities, data are essential for town and country planning.

Firms as well as local authorities must take cognizance of several technical aspects:

- First of all, the quality of the information drawn from big data is directly linked to the quality of the data sets being used. The quality of the input data is important since they come from various (often diverse) sources, and this has repercussions on data analytics and semantics.
- Secondly, interoperability is useful, especially for the capture of data, the specifications applied to metadata, the filtering and extraction of information, and the reporting of results.
- Finally, it is necessary to see to the security of the data in order to vouchsafe their integrity and confidentiality. The full flow of data as they are being processed must be made secure; and the means must exist for non-repudiation (at the source) and approbation (of the result).

All this calls for specific skills and qualifications (data analysts, statisticians, specialized attorneys…) and, too, for the requisite infrastructure and technology.

Conclusion

For big data and artificial intelligence to develop (in particular, as business models), several key problems related to data must be addressed:

- The interoperability of information systems and data is essential in order to be able to use the masses of available data in an overarching program.
- Security has to figure among the technical specifications of the systems for collecting, storing and processing data. Security also implies legal questions about ownership (of the data and of the results of processing them) and licenses (for working on the data).
- To abide by the restrictions on personal data (under the supervisory authority of the CNIL in France), legal requirements must be taken into account before implementation. A key technique for making it easier to use data is to anonymize them, as pointed out in the insert.
- A specific organization must be foreseen to set the conditions for managing (pooling, etc.) and transmitting data among private, public and mixed (public/private) stakeholders.

Table

Voluntary standard on the methodology for assessing the risk of the re-identification of anonymized data: An example from the field of health

Using big data in the health field entails manipulating and processing highly sensitive information. For this reason, CEDRIC, a laboratory at the Conservatoire National des Arts et Métiers (CNAM) has designed models for quantifying the risk of a set of data being re-identified, in particular when data sets are crossed and processed statistically. It used the same hypotheses about data sampling as Latanya Sweeney (2002). The CNAM asked AFNOR (the French Standardization Association) to establish a referential document for this work.

This fast, original method being proposed estimates the probability of a re-identification within a reasonable time. It might eventually be used to quantify the damages ensuing from data theft, for example. Given the financialization and judiciarization of dangerous data, such a tool could help judges make more objective decisions.

While electronic technology is rich with possibilities in the health field, it also bears risks, mainly related to the governance of data and the processing of highly sensitive data. The EU's recent General Data Protection Regulation (GDPR) applies when the data concern private persons. Furthermore, regulations control the access to health data in France, as in the rest of Europe, and state the conditions under which the data may be used. The general rule is that health data may not be used unless exceptions are foreseen.

To ward off these risks, one frequently used method in big data and in health is to make the data anonymous by stripping them of whatever could be used to identify persons. Various techniques do this, some of them incorporated in standards. The risk of a "data breach" is not negligible; and even "anonymized" data could be reworked to re-identify the data subjects:

— 72.7% of re-identification attacks have taken place since 2009 (all sectors together).

— The main data affected are ADS data sets (Excel spreadsheets plus graphs).

Having a method for assessing the risk of re-identification inherent in the anonymization process is, therefore, important. Current techniques for evaluating how robust anonymized databases are use "brute force" and are, therefore, costly in time and computing power. As a consequence, they are used sparingly.

To the best of our knowledge, there is not yet any standard for assessments of the risk of anonymized data being re-identified. At stake in this standardization is the promotion of the state-of-the-art method designed by CNAM.

Bibliography

BERA M. (2017) "Big data et anonymisation", *Le CNAM. Le Blog*, 5 October. Available at: http://blog.cnam.fr/technologie/les-big-data/big-data-et-anonymisation-947632.kjsp?RH=1516962150988.

EXECUTIVE OFFICE OF THE PRESIDENT (2015), *Big Data: Seizing Opportunities, Preserving Values* (Washington, DC: The White House), interim progress report of February, 11p. Available via https://obamawhitehouse.archives.gov/sites/default/files/docs/20150204_Big_Data_Seizing_Opportunities_Preserving_Values_Memo.pdf.

PRESIDENT'S COUNCIL OF ADVISORS ON SCIENCE & TECHNOLOGY (2014) *Big Data and Privacy: A Technological Perspective* (Washington, DC: The White House), report to the President, May, 76p. Available via https://bigdatawg.nist.gov/pdf/pcast_big_data_and_privacy_-_may_2014.pdf.

GARNIER A. (2013) *Big Data et réseaux sociaux: mythes & réalités. La déclinaison pour les Réseaux Sociaux d'Entreprise* (Paris: Groupement Français de l'Industrie de l'Information). Available from http://www.gfii.fr/fr/document/big-data-et-reseaux-sociaux-mythes-realites-ladeclinaison-pour-les-reseaux-sociaux-d-entreprise-un-livre-blanc-d-alain-garnier-ceo-de-jamespot.

HAMEL M.P. & MARGUERIT D. (2013) "Analyse des Big Data. Quels usages? Quels défis?", Commissariat Général à la Stratégie et à la Prospective, *La note d'analyse*, 8, 12 November, 12p. Available via https://www.strategie.gouv.fr/publications/analyse-big-data-usages-defis.

HENKE N., BUGHIN J., CHUI M., MANYIKA, J., SALEH, T., WISEMAN, B. & SETHUPATHY, G. (2016) "The age of analytics: Competing in a data-driven world", (McKinsey Analytics) December. Available at: https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/the-age-of-analytics-competing-in-a-data-driven-world.

HUOT C. (2015) Editor: *Livre Blanc. Données massives: Big data, impact et attentes pour la normalisation* (Paris: AFNOR), June, 66p. Available via https://docplayer.fr/11353189-Livre-blanc-donnees-massives-big-data-impact-et-attentes-pour-la-normalisation.html.

ISO/IEC JTC 1 (2014) *Big data*, preliminary report (Geneva: ISO). Available via https://www.iso.org/iso/big_data_report-jtc1.pdf.

MANYIKA J., CHUI M., BROWN, B., BUGHIN, J., DOBBS, R., ROXBURGH, C. & BYERS, A.H. (2011) "Big data: The next frontier for innovation, competition, and productivity" (McKinsey Digital) May. Available at https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation.

SWEENEY L. (2002) "Achieving k-anonymity privacy protection using generalization and suppression", *International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5), pp. 571-588.