Premier ministre

SECRÉTARIAT GÉNÉRAL POUR LA MODERNISATION DE L'ACTION PUBLIQUE

Chief Data Officer

2015 Report to the Prime Minister on data governance

Leverage data to modernise public action

July 2016

The decree of 16 September 2014 creates, under the Prime Minister's authority, a Chief Data Officer (CDO), attached to the Secretariat-General for Government Modernisation (SGMAP).

The CDO coordinates the administrations' action with regards to the inventory, governance, production, circulation and data use.

He organizes, in accordance with personal data protection and law protected secrets, the best exploitation of data and its wider circulation, in particular for the purposes of assessing public policies, improving public action and stimulating research and innovation.

Pursuing those goals, the CDO shall propose all relevant measures to the Prime Minister, including, when appropriate, legislative or regulatory reforms.

Under that decree, The Chief Data Officer is responsible for encouraging the better use of data by the administration — in particular by disseminating the culture and use of data sciences - and for encouraging the largest possible liquidity of data. The CDO can be consulted by anyone on any question relating to the circulation of data. Local and regional authorities, legal persons under public and private law entrusted with a public service remit may refer to him, on any matter related to data use by their services.

The Chief Data Officer's site allows to submit a case for consultation on data circulation and to follow the first results of data-driven projects :

http://agd.data.gouv.fr .

Finally, the decree states that the Chief Data Officer shall deliver each year to the Prime Minister a public report on the inventory, governance, production, circulation and use of data by administrations.

This report also describes existing government data, their quality and the innovative exploitations that they allow. It presents recent developments around data use and contains proposals to improve the exploitation and circulation of data between administrations.

TABLE OF CONTENTS

INTRODUCTION

Predicting and preventing car thefts, optimizing waiting times for emergencies, better targeting of customs controls, identifying energy deficient buildings, identifying companies that will soon be recruiting and informing the relevant job seekers, addressing the shortage of teachers in school, optimizing traffic lights to decongest and clean up the city centers, revising the price calculation formula for medicinal drugs, negotiating electricity purchases by anticipating and controlling consumption peaks, negotiating better public procurements, predicting the micro-economic effects of a tax reform, forecasting medical investment needs through the analysis of scientific literature... . all these uses of predictive analysis are within reach of public authorities. They hold enormous potential for efficiency, expenditure control and policy accuracy.

Predictive analysis is only one example of a set of new practices : "data-driven strategies", which for example allow :

- To regulate an industrial sector by a releasing relevant data — as the government is experimenting with geolocation data from taxis to enable them to benefit from digital services;
- To organize information so that individual public servants can take better decisions;
- To improve the day-to-day work of front-office staff by providing them with more real-time information;
- To increase the autonomy and freedom of choice for public service users — for example in predicting the likeliness of a successful legal procedure.

These promises carried by data sciences are at the core of the digital transformation of big companies and large cities worldwide and are one of the levers for the modernisation of public action. By creating the function of Chief Data Officer on September 16th 2014, the Prime Minister wanted to make data sciences tools available to public decision makers, a few months before the United States and Britain did likewise.

This implies integrating new skill sets in government teams: data scientists, statisticians with innovative profiles, computer geeks, keen on new data-processing methods, and concerned with the concrete impact and translation of their mathematical results. It requires high-quality data — that France does indeed produce and handle following a long-standing tradition, thanks to its high standard statistics and its commitment to the quality of public service. It also requires an enhanced culture of "data-driven strategies", an ambition to carry this type of change and the patience to continuously test and verify whether small changes can produce major improvements.

The correct implementation of these methods, however, requires a prior effective data governance, i.e. global management of data produced or held by government to ensure the quality, freshness, interoperability and availability in the correct formats, facilitating swift use and dissemination amongst public servants —from central and local governments —so that they receive the information necessary to the performance of their tasks in compliance with important statutory secrets that protect fundamental freedoms and the nation's fundamental interests. It also requires organisation that allows the State control and sovereignty over its data, processes and systems, and provides citizens the transparency they are entitled to claim.

To be used in ambitious government reforms, these methods must be integrated into a "data policy". The growing role of data and algorithms in society may sometimes cause concern, but it is essentially an increased capacity for public action, which we should know how to make good use of. It requires

Fighting car theft using data sciences

The Chief Data Officer and the office of technology and information systems of internal security (ST (SI) [2]) are working together to develop a predictive model of vehicle theft. The aim is to achieve an optimal allocation of gendarme and police patrols in the Oise department. Using more than 600 geographic and socio-economic variables as well as other indicators such as the weather and occurrence of thefts in neighbouring areas in the preceding days, the CDO*'s[1] team has developed a model showing promising results. Allocation of daily patrols in the 10 % of most risky areas (probabilities estimated by the model) would have covered more than half of the thefts that occured during the last five months of 2014. Direct cooperation with the national security services present on the ground is now planned to build a tool that will be tailor-made and adapted to the sectors problematic.

[1] All terms followed by a * are described in the annex.

transparency and independent checks and balances as well as controlled and democratic use of data, transparency of outcomes and results and cooperation with civil society. In 1978, the legislature had organised the main security features to protect privacy, creating a regulatory framework for analysis still valid today, forty years later . In 2015, other provisions, on the ethics of algorithms, on the opening of decision and public action, will probably be necessary.[2] This will require a major educational campaign and will generate important debates, especially on the weight of the individual interest against the public interest in regards to optimisation algorithms.

Therefore, in establishing the duties of a Chief Data Officer, the Prime Minister gave him the task of preparing each year an annual report on data governance, to measure progress in quality, circulation and use of data, but also to measure the administrative, political and social adoptions of these methods, and democratic strategy building based on these methods.

This first report, based on a year of investigation, exchange and experimentation with numerous public officials and many administrations intends to provide a framework of analysis, detect promises and illusions of data-driven science, present initial results, report the initial difficulties encountered and suggest first orientations.

It also tries to fit the ongoing revolution in a historical narrative: that of the administration and of the French State. Without repeating the already long, well-documented history of official statistics, it is important to bear in mind that these new methods did not, unlike what some pundits claim, appear with the rise of big data*.[3] They fit in a long history of relations between statistics and political power, marked by the establishment of key operators, by the introduction of a specific system of governance and a specific economic model, secured by processes, regulations and laws, shaped by the State's global IT architecture.[4] The fact that this history is now turned upside down by the rise of new large quantities of digital data, new tools and new players, but also new philosophies of action itself, does not preclude the need to learn, understand and estimate its fair value.

Therefore, this first annual report from the Chief Data Officer will, retaining its ambition to deliver practical and operational developments, put modern day data use back in its historical perspective.

The first part will analyse the role of data in public action, and in particular :

- The progressive development of increasingly diverse and accurate data, with the digital shift of data stemming from computerized management systems;
- The increasing use of these data and as a consequence the legitimate ambitions which may be placed in these new practices.

The second part will identify the main bottlenecks to efficient data use after a year of experimentation and interrogate the current principles — albeit unwritten — underpinning the existing data governance. In the absence of an explicit one, the rules governing relations between administrations or the pre-existent IT architecture choices produce governance by default.

Finally, the third part of this report will present the main paths towards public service efficiency through new data governance and present first provisions that can be implemented, some of them in the very near future.

Prepared under the responsibility of The Chief Data Officer it has received significant contributions from the SGMAP, including Etalab, the DINSIC, exchanges with the ANSSI, the APIE, the CGEIET, the INSEE, the CNIL and the ministerial IT departments, as well as the huge contribution of Simon Chignard, within Etalab.

Data sciences and employment

To help job seekers find work, you can look for the offer that best suits their needs. You can also identify companies likely to recruit them and encourage them to apply. "La Bonne Boite" a State start-up developed by Pôle Emploi and the SGMAP follows this direction. Using economic data describing the company and its employment history, the model predicts for each sector and each department the probability that a company will recruit within the next six months. Job seekers can then focus their research on companies with the highest probability of recruitment in the sector.

[3] Desrosières A. (2000) : La Politique des grands nombres : histoire de la raison statistique, Editions La Découverte) (2nd edition)
[4] For example, we can underline the importance given by the National Council of Resistance to the existence of independent, high-quality official statistics, which has directly led to the 1951 law on official statistics, and therefore both to the INSEE's structure and the statistical confidentiality which structures present information exchanges between companies and the State.

The Chief Data Officer, one year of action

During its first year of operation, The Chief Data Officer worked with 3 goals : gear data sciences to improve public policies, facilitate the flow of data and leverage ecosystems, both within and outside public administrations.

1/ Gear data sciences towards public policies

In 2015, the CDO set up a team of four "data scientists" offering their services to administrations. This team has succeeded in under a year in producing encouraging results with several voluntary ministries. Attention is drawn in particular to the following results :

- A mission with the State Purchase Centre, to analyse the power consumption of the State and thus lead a firmer control on purchases[5];
- A mission with the office of technology and information systems of internal security (ST (SI)[2] that has helped develop a prediction model of car theft at the department level;
- A mission with the teams of Pôle Emploi to predict with 80 % accuracy if a company will recruit a given profile in the next quarter, which enabled Pôle Emploi supported by SGMAP to develop "La Bonne Boîte"[6].

2/ Facilitate the flow of data within the administrations

The Chief Data Officer has put a referral procedure in place, to enable citizens facing difficulties with poor data flow to report issues.

In the course of 2015, a dozen of referrals from State officials or public bodies, researchers, journalists, companies and individual projects on the use of data have been processed. In addition, the Ministry of Health and Social Affairs has asked for the CDO's formal opinion in preparation of the health law draft.[7]

Support was provided to the team in charge of assessing data sales between administrations led by Mr Antoine Fouilleron of the Court of Auditors.

3/ Leverage ecosystems

In 2015, the CDO prepared a data sciences public procurement contract. This contract is part of the SGMAP's (Secrétariat Général pour la modernisation de l'action publique) general strategy to support administrations in their transformation projects. Starting 2016 they will therefor benefit from additional resources for data sciences projects.

Henri Verdier
Chief Data Officer
December 2015

[5] Documented by the SAE and the CDO :
https://agd.data.gouv.fr/2015/05/17/analyser-les-consommations-energetiques-des-batiments-publics/
[6] http://labonneboite.pole-emploi.fr/
[7] https://agd.data.gouv.fr/2015/04/02/avis-portant-sur-la-publication-la-rectification-et-la-reutilisation-des-informations-portant-sur-les-professionnels-de-sante/

1

Data at the core of public action

1. THE STATE : A DATA PRODUCER

The State has been producing data for its own needs and those of society for a long time. The excellence of its large data producing operators is widely recognised. The economic and social importance of cadastral, geographical and meteorological data, the scientific and democratic importance of reliable and independent public statistics, just like that of legal data embodying the constitutional principle of open courtrooms, are difficult to quantify since their uses are so ubiquitous.

With lower production costs, the emergence of new production strategies — possibly including citizens themselves — and the multiplication of non-scientific data, the nature and scope of data produced is undergoing a dramatic change in scale and even in nature, urging for new structures.

A very brief history of State-produced data

The construction of a modern State is accompanied by the gradual creation of a set of reference data necessary for its organisation and key to a functioning society. From the seventeenth century, royal power sought to codify and standardise parish records which at the time constituted the best source of knowledge on the population. In addition to these records, the authorities have developed a more detailed knowledge of the territory for defence and tax calculating purposes. The mapping launched under Louis XIV lead to the first comprehensive map of the French Kingdom. The Napoleonic land registers made it possible to allocate taxes to citizens, whilst the first statistical office organised a general population census in 1801. From the Old Regime on, maps and surveys conducted by the official hydrographic service were available to "vessels, both of trade and of war".[8]

Most of this data was built to meet the needs of the public authority and to support its development. Its original purpose often determines their present-day ministry of supervision : the land register is currently managed by the Directorate-General of public finance, and the Hydrographic and Oceanographic Service (SHOM) is under the Ministry of Defence.

At the end of the Second World War, official statistics and planning go hand in hand : the former describes the population and the national economy, the latter guides and orients development with a series of great projects.

The computerisation of the administration, carried out on a large scale in the 1970s, gives a new pulse to registers. In 1974, the administration only had 200 computers.[9] The models were very costly, centralised, and used exclusively for the most complex tasks : social security and health system management. The records had become "information systems", improving processing capabilities, and allowing hitherto unprecedented overlapping. The State was the first, along with the large banks, to seize the opportunity of IT development.

The data produced by the State, and more generally by all different kinds of public services come in varying forms and are used for a wide range of purposes. This variety of nature and use partly explains the complexity of data management and the difficulty of drawing up a single governance for data and information as diverse (statistical information, meteorological, geographical, administrative, fiscal, management system information, large databases co-produced with companies, researchers and citizens, etc.). The aim here is not to fully classify data produced by the State but only to raise awareness among the various decision-makers. They need to know the scale and complexity of the subject.

What is a digital data ?

Digital data is the basic description, of digital nature, represented in coded form, of a reality (measure, transaction, event, etc.) that has to be :

- gathered, registered,
- treated, handled, processed
- saved, archived
- exchanged, distributed, communicated.

Depending on the intended use, data might be closed (reserved to some people or organisations), shared (subject to specific contractual constraints — specific licences — or terms of use) or open (open to all users and all legal uses). [10]

A piece of information is a set of data aggregated for man-use. To be usable, the data needs to be accompanied by metadata (literally data about data) which make it possible to describe it as accurately as possible (origin, production method, destinations, legal rules, etc.).

[8] Decree of the King Council, 1773
[9] Nora Minc report on computerization — 1978
[10] Source: Data spectrum, Open Data Institute (https://theodi.org/data-spectrum)

It is common practice to organise the data in three broad categories, namely data allowing :

- To identify or designate persons or things : individuals in the human resources information system (SIRH), equipment (vehicle identification), infrastructure (roads), identification of persons (civil status, the national identification register of natural persons), organisations (the register of legal units), rules (the law is structured and organised to identify particular parts), events, motivations and budgets (organic law relating to finance laws, the "LOLF"). This identification plays a vital role in any society. And the State plays a central role in the definition of minimum rules or standards applicable for society and for itself, to identify objects or persons (e.g. civil status);
- To describe, characterise things to use them, interact with them, study them, etc. For example, data to describe a firm : its activity, its geographical locations, its size, its turnover, its relations with the administration, legal obligations, etc.
- Or to decide to drive, to steer... to make decisions.[11]

The major producers of public information

In order to fulfill its duties, the State has established operators dedicated to the production of data. INSEE, IGN Météo France or the INED, INSERM, the ONEMA are known for their expertise and the quality of their production.

The National Institute of Statistics and Economic Studies (INSEE) produces, analyses and disseminates statistical information on the economy, society and the French territories. This information is linked to the macroeconomic, demographic and social sectors. The INSEE is in charge of the civil status registers of companies (SIRENE) and the annual population census.

The National Geographic Institute (IGN) has the responsibility, since the post-war period, of mapping France and its territories. It has kept adapting its production to the challenge of digitisation, including through the production of large-scale repositories or the provision of a Geoportal. The National Institute of Health and Medical Research (INSERM) is the only public French research body entirely dedicated to human health and its researchers produce very large volumes of data, including on epidemiology.

The reputation of these major producers is established at an international level and they participate in European and international standardisation efforts. The INSEE represents France at the European Statistical System and the Director of the IGN does the same within the United Nations Committee of Experts on the geographic information management (UN GGIM).

New sources and new data production methods

Like businesses and citizens, the State is acknowledging the advent of large amounts of new types of data.

We too often tend to limit the ongoing revolution to explosion in the amount of data produced.[12] However, the diversity of sources and data production methods is probably even more remarkable.

"Google knows more about France, than the INSEE" claimed two French IT researchers in 2013. The formula, largely taken up, hit the spot. It even got quoted in the preamble to the meeting of the National Statistical Information Council dedicated to the use of big data in statistics. However what this formula doesn't show is the complementarity between sources. The INSEE compiles statistics using robust scientific methods, in an objective and transparent fashion. Public statistics are consistent internally, and over time. They allow comparisons between regions and in most cases also between countries.

The main digital players, in turn, collect data in a non-scientific way, using sensors, traces of use, or direct contribution of Internet users. They have neither the robustness nor the completeness of scientific data. However, they gradually, by their sheer volume, create a form an imprint of reality, which may in turn be interpreted and used to produce active knowledge.

[11] You can also file data uses along three axes. The first is the identification and description of users. The second covers all the data necessary for the performance of the state's missions (defense, justice, education, health, work, etc.). The last axis relates to the resources that public services use for their missions, assets both physical (furniture, equipment, premises) and non-material, as well as persons (staff, partners, suppliers). Finally, a distinction must be made between permanent data and flowing data.
[12] 2, 5 trillion bytes are produced daily in the world and 90 % of the existing data stock has been generated over the last two years according to IBM.

To use this data, statisticians must learn to deal with a new set of issues. They must reverse their usual approach from adapting the data collection at the desired level of precision to assessing the degree of reliance that can be placed on data whose main purpose was not statistical information. Waze, providing a GPS identifying congestion, has created maps of many emerging countries cities simply by analysing the routes shaped by mobile phone users. Apple collects a lot of biometric data, like many companies offering fitness accessories. Telephone operators, e-commerce platforms and digital players are amassing knowledge on society and the economy.

The data produced by the administrations and data produced externally are not conflicting, they may even complement each other. The INSEE is thus developing a follow-up project of the consumer price index based on anonymous checkout receipts. Many researchers have shown that carefully analysed social data could shed light on key issues very difficult —or expensive— to examine on the sole basis of survey data.

Last but not least of the ongoing changes : producing essential data no longer is the sole prerogative of the State. Voluntary contributors from OpenStreetMap are mapping the country at high speed. Co-production of essential data with the multitude is not a futuristic scenario : The national address repository is the product of collaboration between the National Geographical Institute, La Poste, The Chief Data Officer and the Association OpenStreetMap France.[13]

There are new sources as well as new ways of producing data. The State is not absent from this revolution : The management systems it uses also produces data in the form of traces, which may now be used. Its officials may use terminals to collect extensive information, and the State is progressively demonstrating its ability to enter the world of common assets creation.

Four examples of new sources of data used for public policies

For the detection of unknown side-effects of medicines

In the US, the Food and Drug Administration (FDA) has now reached an agreement with Google to identify the unknown side-effects of medicines. Anonymised inquiries from the search engine can notably identify side effects which arise after the medical treatment and which are today sometimes underestimated by the current pharmacovigilance apparatus.

To monitor the spread of epidemics (malaria, dengue fever, Ebola virus)

Data from Twitter and Google are used in many countries (including Brazil and Singapore) to monitor the spread of communicable diseases such as dengue. Healthmap analysed thousands of live data sources and identified the Ebola outbreak almost a week before the alert was formally triggered by the concerned countries.

To improve the availability of transport in a city

Under the program Data4Development, Orange, the telecom operator, has made available to the scientific community, anonymised data and in particular the location of mobile phone users in Ivory Coast and Senegal, to observe origin-destination flows within a city. They are then converted into journeys at the existing transport network level. Thanks to this, an IBM team has helped improve the transport system of Abidjan, so as to increase the number of lines and user satisfaction, both in terms of possible journeys and waiting time.

To identify economic and food crises

The United Nations Secretariat Global Pulse program analyses data from Twitter to track the evolution of opinion in each country. By doing so, they were able to detect food crisis due to the surge in price of agricultural raw materials in real time. This analysis does not replace the official measure of inflation, but complements it, offering a real-time view.

[13] See details at: http://adresse.data.gouv.fr/

2. THE STATE AS A DATA USER

If the State is a data producer, it is primarily because it uses data itself. Public policies are transformed by he digital revolution. "Data-driven strategies" represent new opportunities for those who master data to action and use them as regulatory tools. Production or use of open data, as a new tool, broadens the range of policy options available to the State.

Organising the State, organising society, changing traditional practices

Data is necessary for the daily functioning of the State and the public services

To carry out its tasks efficiently, the State must mobilize fresh high-quality data. This data operates at each stage of public action : diagnosis, programming, implementation and evaluation. Its uses are numerous, ranging from the management of staff in schools to the preparation of tax reforms as well as planning works or many investment decisions.

The State, like most large organisations, has massively automated and optimised a lot of processes, and is now handling a growing amount of data. Its IT systems manage the day-to-day data necessary for the functioning of the authorities, such as the data on the identity of individuals and on vehicles both used on a daily basis by law enforcements.

The re-use of administrative records by the administrations whose primary purpose is to make information available to the public has been going on for a long time. Public statisticians increasingly use data files held by tax or social administrations to produce statistics on employment, business activity and income. Administrative records avoid resorting to investigations costly for both respondents and investigators. Those records also meet the increasing demand for data at detailed geographical levels or detailed nomenclature, in particular when supporting public policies.

Thus, the finely located information available to the INSEE in many areas, in particular regarding equipment or income (they have for several years been released by "tiles" of 200 m per side each), allows it to provide very precise quantitative insights in support to the relevant administrations. For example, one might cite the essential contribution of the localized tax revenue data to urban policy reforms and their follow-ups. The mobilisation by the INSEE of an innovative methodology based on data grid-statistics allowed the identification of new priority districts on the sole basis of the inhabitants' income criterion. This approach has been translated into law and two decrees.[14] These areas are now intended to be tracked over time, through a set of statistical indicators.

The new challenges facing public authorities — e.g. security, urban development or energy transition — are increasingly complex and involve a large number of different actors : governmental departments, authorities, companies and third-sector stakeholders. Sharing data will follow this movement of inter-departmental work.

Data for housing policies

42 billion euros of public money is spent annually on housing. The amount dedicated to producing data, including statistics, and to steer these public policies is around 30 million euros.[15] This share reflects the difficulty of obtaining data in a field where, local authorities are playing an increasingly important role.

The thematic debate on housing open data, conducted jointly by the National Board of Housing and Etalab showed that :

- The relevant data is not always available, due to a lack of coordination between numerous actors involved and given the complexity of the subjects dealt with;
- Existing data is not always available at the most detailed level, making it hard to take the diversity of local situations into account..

It is urgent to better organise the collection, recovery and sharing of housing data and secure their quality and re-use potential.[16]

[14] See in particular the Town Planning Act of 21 February 2014 and the 2013 activity report of the INSEE, pages 20 to 22.
[15] Inspectorate General of the INSEE 1.7.25 — General Council of Environment and Sustainable Development 009075-02 "report on the organisation of the statistical service in the field of housing" — http://www.ladocumentationfrancaise.fr/var/storage/rapports-publics/144000532/0000.pdf
[16] National Board of Housing, Etalab (2015): open data in the field of housing, summary of discussions

Reference data is essential to the functioning of society

Similarly, the State has had a long-standing production of reference data (data used by a wide range of entities which use them frequently, e.g. geographical official codes, the SIRENE register, the land register, etc.). Many economic and social activities are based on the quality and availability of those data, classifications, or essential data repositories.

With the digital revolution, and the strengthening of data usage by new services, new reference data is identified. Precise geolocation of buildings has, for example, become essential for new services. At the same time, with the reduction in data production costs, the capacity of the State to define standards is partly called into question. A standard is considered one if it is recognised as such by the users, and not when it's unilaterally defined.

In this world of default standards (and not just norms), preserving France's capacity to act is, for example, ensuring that the business identifier remains the SIREN number, provided by the INSEE, and not an identifier assigned by a third party, e.g. a financial information company. Providing the "official" data is not enough : It must be of quality, comprehensive, up-to-date and made available 24/7 via APIs*[18] with high-quality service. Many key repositories became obsolete in just a few years. Twenty years ago, human knowledge was classified according to the categories defined by the Library of Congress (Dewey), and by the BNF (Rameau). These two typologies have clashed briefly when Yahoo! tried to index the web according to a similar logic. Google's approach, using algorithms based on the hyperlinks defined by users has shifted this reasoning: repositories became obsolete, not as means of classifying and prioritising, but as a means of finding a document.

Data is a strategic asset whose value is in re-use more than it is in its initial use. The GPS's example is enlightening. The satellite positioning system developed by the US in the late 1970s became operational in 1995. Its use was initially restricted to that by the US Army, and was then gradually extended to civil use, following a decision by President Clinton. Today, the GPS has become the key platform for the functioning of many industries, from aviation to agriculture and transport. Europe, China and Russia strive to deploy their own network of satellites so their economies won't solely rely on this infrastructure controlled by a single super power.

New action strategies

"Software is eating the world" as digital revolution players like to recall, quoting Marc Andreesen's famous words.[19] Businesses, key actors, strategies and tools which have made the digital revolution possible are transforming many fields of human activity. To fulfill its tasks and master its costs the public authority has to take up these tools, methods and strategies.

The example of the cadastral reference system

A large body of literature has emphasised the economic impact of a key repository very common in France : the cadastre.

Without a cadastre, the public authority has very little visibility when charging tax and securing property rights, as shown by the failure of Greece to levy property tax, a problem sometimes invoked as key factor in the Greek public debt crisis.

Without a cadastre, moreover, it is almost impossible for landowners to use their property to borrow and therefore to invest. Numerous analyses show that the absence of a cadastral system hampers the emergence of a middle class in developing countries.[17]

Uses also significantly changed : Once solely used by the State for tax purposes, modern cadastre is now used by local governments (urban planning, infrastructure management) but also by private enterprises (security of real estate transactions, mortgage credit support, etc.).

The land register is no longer just a tool for a specific use (levy the land tax) it has become a strategic element, a standard on which many actors coincide in order to exchange. In that respect, the cadastre is a reference enabling the coordination between stakeholders, which allows them to join forces and achieve the best possible balance.

Reference data has similarities with currency which is produced and guaranteed by the State, to allow for an exchange between actors and the smooth functioning of the economy.

[19] Marc Andreessen (2011): Why software is eating the world, The Wall Street Journal

Data sciences : New fields of action

What are data sciences ?

The term was coined by Jeff Hammerbacher (Facebook) and DJ Patil (LinkedIn now White House's Chief Data Scientist) in 2008. It refers to people who analyse data, not in order to produce reports or statistics, but to improve the product or service in the organisation for which they work. Data Scientists thus have both the ability to analyse the data, the capacity to write computer code and the ability to come up with new usages.

For example, in 2006, Jonathan Goldman, newly recruited at LinkedIn, observed that data could predict a user's network. He then imagined the "People you may know" module and ran some tests. This model encountered great success and has played an important role in the development of the social network. Similarly at Eventbrite, Frenchman Paul Duan developed highly original and efficient approaches to fraud detection using algorithmic models.

Likewise, large companies have developed new digital uses for statistical methods and algorithms statistical learning. Facebook uses user data to predict potential friends, LinkedIn to predict professional contacts, Netflix to predict relevant movies and Amazon to predict the products you may purchase.

Data sciences therefore use various methods of statistics and "machine learning *", linear regressions, logistic regressions, decision trees, random decision forests and segmentation algorithms and various data visualisation methods to devise new applications.[20 21]

Data sciences assist authorities in their decision making. This requires relevant data, but also that the solutions offered are easy to enable and have a measurable impact :

- Easy to enable when working on concrete operational software : they do not simply illustrate, observe or even understand, but use data to support decision-making. Do I have to buy this product and at what price ? How to plan resources to address needs ? Where to begin ?
- Measurable by working on decisions whose impact is quantifiable and measurable to ensure that that measures in turn nurture ("train") the algorithm.

With data sciences, data therefore does not only describe reality or help decisions : it falls squarely within the policy process.

Developments occurring at the border between data and action are without doubt the essential dimension of data sciences' revolution. To fully understand this we have to acknowledge just how much those new tools widen the range of actions available to the decision maker.

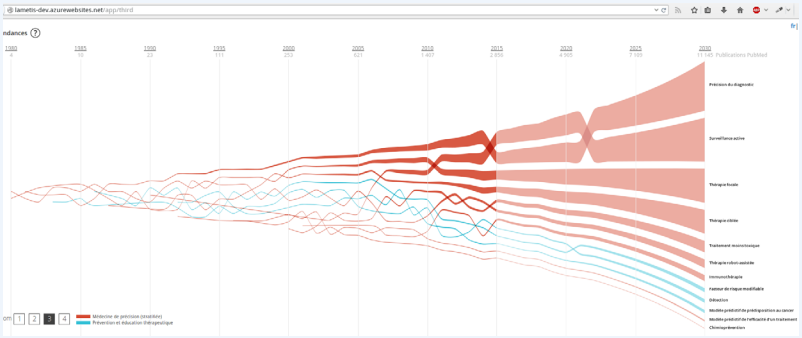
A simple example. In a traditional scheme, the statistician seeks to identify verified causation to intervene in a linear process. If, for example, he can prove that speed is causing road accidents, speed limits will automatically produce a decrease. In this action, it is essential to understand that "correlation does not mean causation", because we would otherwise work on factors which are not genuine determinants. With the capacity to process massive data in real time, and therefore to measure, on a daily basis, the effects of a decision it becomes less important to distinguish correlations from causality. It is indeed possible to test an action's effectiveness on a daily basis and modify it as soon as it appears to no longer to produce the desired effects. This approach is not without epistemological and sometimes ethical questions. Some of those have been discussed in a famously provocative Wired article in 2008.[22] The fact remains that, as far as action is concerned, this approach is working.

This focus on "data-to-action" explains a number of strategies in North American cities. The city of New York has identified buildings with high fire risk to assist the fire-fighters in their preventive action (going within a few weeks, from 10 % to 78 % positive controls); "data-to-action" also helped dealing with Hurricane Sandy's aftermath using methods inspired by probabilistic approaches.[23]

[20] Press G. (2013): A very short history of data sciences, Forbes.com
[21] Davenport T., Patil DJ (2012): Data Scientist, the sexiest job of the 21st century, Harvard Business Review
[22] Anderson, C. (2008): The end of theory, the data deluge makes the scientific method obsolete, Wired magazine
[23] Flowers M. (2013): NYC by the numbers, annual report to the Mayor of New York

An example of health big data which does not use personal data

Data sciences development is also the development in anlysis of new types of content such as images or texts. For example, to predict the next technologies in curing prostate cancer and imagining the hospital of tomorrow, La Métis, supported by The Chief Data Officer analysed all the publications of the US Library of Medicine (Medpub), freely available and re-usable publications on the topic going back to 1980. The purpose of this analysis was to identify patterns of diffusion of innovations and to predict the volumes of publications on each of the identified innovations for the next 15 years. From those it's possible to infer trends and practices, and therefore recommendations on hospital investments.



Reading : The analysis shows that the accuracy of diagnosis and active surveillance are the two major trends in the scientific literature on curing prostate cancer.

The data visualisation developed by La Métis with the CDO allows public authorities to have a clearer picture of solutions for detection, diagnosis and treatment of prostate cancer in 2030 and practitioners and experts to have a more complete picture of the scientific literature on the subject.[24]

[24] https://agd.data.gouv.fr/2015/11/25/cancer-de-la-prostate-a-quoi-ressemblera-le-parcours-de-soin-en-2030/

Sharing and opening data to create social and economic value

Open data has long been analysed in France as a consequence of the public's right to information, which it undoubtedly is in part. But it is important to emphasise that it may also be a very efficient policy shaping tool. Sharing data on www.data.gouv.fr, the State and local authorities are fostering the re-use by society and can drive innovation, bring the country to an open innovation approach, or even correct market imperfections and improve public data.

The provision of free and open data and its re-use creates economic and social value, through five mechanisms :[25] transaction costs reduction, innovation, information asymmetries reduction, collaboration and feedback loops. Those mechanisms, described below, are not exclusive ; they can combine for the same data.

The efficiency : reduce transaction costs

Opening data allows for a better use of available resources by public and private actors. The theory of transaction costs is that every economic transaction generates costs (cost of searching for information). By providing free public data we reduce the transaction costs, both in preparations and in transactions themselves. The availability of data is a source of efficiency and effectiveness, both for the authorities and for private parties. Several experiments in France and abroad confirm this mechanism of value creation. In Australia, the transaction costs resulting from the sale and distribution of Australian geographical data have been assessed before they were made available for free in 2002, between 17 % and 33 % of revenues. The annual gain of this opening has been estimated at 1,7 million dollars per year for the reduction of transaction costs.[26] In Denmark, the government has launched a program called "basic data". The aim is to establish a free information infrastructure around three core databases including the company register and essential spatial data. The benefits of this project are estimated at 35 million euros every year for the public sector (efficiency) and 70 million euros for the private sector (creation of new services).[27]

Innovation and transformation

The second mechanism of value creation is tied to the use of open data by the public and private sectors to create new products and services. The overall capacity of innovation plays a decisive economic role. It allows not only growth but also raises significant structural changes.

Open data is a factor for innovation for those who re-use. In the Netherlands, opening meteorological data enabled the creation of a highly dynamic re-user ecosystem : Private actors' income increased by 400 %, the number of data uses rose 300 %. These activities have generated a return of 35M€ for the Dutch public finances, in taxes and additional charges.

Several studies show that lowering or removing fees causes a mechanical increase in data re-use.[28] For example, opening the IGN's large-scale reference data to the public service delegates resulted in a multiplication by 20 of downloads and social benefits estimated at 114M€/year, for a loss of revenue from frees of only 6M€/year.[29]

In France, data.gouv.fr hosts and organises a community of over 600 organisations, half of them public bodies and more than 10,000 users, who have published a total of 90,000 resources, 21,000 datasets, and more than 1,300 reuses. The Dataconnexions contest organised by Etalab since 2012, identified 200 high-potential start-ups. This system makes it possible to give an initial impetus to these projects, contributing to their growth and consolidation. Snips, who won the contest's 3rd edition became within a few months a company with 35 employees.

25 Jetzek T., Avital,M. (2013): The generative mechanisms of open government data, ECIs 2013 proceedings
26 Mr de Vries (2012): Re-use of public sector information, report for Danish Ministry for Housing, Urban and Rural Affairs
27 Ministry of Finance of Denmark (2012): Good basic data for everyone — a driver for growth and efficiency
28 Vickery, G. (2010): Review of recent studies on PSI re-use and related market developments
29 Trojette A. (2013): Open data: The exceptions to the principle of free access are all legitimate? report to the Prime Minister

The reduction of information asymmetries

The third mechanism generating value is linked to the reduction of information asymmetries by means of transparency. Information asymmetry occurs when a player has better or more complete information than other actors involved in a transaction or communication. Information asymmetry leads to sub-optimal situations. Open data can reduce those asymmetries at various levels. At the macroeconomic level, transparency is a tool for fighting corruption as acknowledged by the World Bank. At the microeconomic level, the availability of detailed data on public procurements allows all stakeholders to have the same level of information. Respondents can know the last successful tenderer for a public contract and the market conditions, thereby enabling them to better tailor their response. The number of replies and their quality is improving, which is also a condition of the effectiveness of public procurements. In France, the purchase department of the State and The Chief Data Officer have also conducted an analysis of energy use in buildings, which has identified consumption profiles. This analysis, as well as the underlying data have been made available to potential energy suppliers .[30]

Feedback loops to influence behaviour

Sharing data also encourages feedback loops. Sharing real-time information about a system allows its operators to change their behaviour, assess the effects of those changes and adjust dynamically. Speed indicators panels installed at the entrances of cities and on some road sections operate in accordance with this principle. Their presence reduces average speed by 10 %. Similarly, transmitting traffic predictions to drivers one hour before congestion is expected ultimately helps reducing its intensity. Feedback loops have wide-ranging applications for public action. The US Department of Labour publishes on a quarterly basis since 2010 the list of the worst 500 companies with respect to the application of health and safety at work legislation.[31] Not being on this list represents a major challenge for employers and provides a strong incentive to better protect their employees. In France, the Ministry of economy, industry and the digital sector did likewise by publishing in November 2015 a list of 5 big companies who received the highest fines for repeated late payments.[32]

Collaboration to produce, enrich and improve data

Collaborative data production is not strictly speaking a novelty, participatory science benefits from a long tradition in the fields of botany, observation of biodiversity or even astronomy.[33] Digital gives a new impulse to these practices and widens their scope.

The availability of open data creates the conditions for collaboration between many actors, both public and private. This collaboration around data is a new action strategy.

Indeed, cooperation generates economies of scale. Thus the Platform data.gouv.fr enables everyone to enrich, improve and share datasets. Since the end of 2013, numerous enrichment examples have been documented. Files on accidents resulting in death or injury have been subject to numerous improvements by users : cleaning, correction of duplication, addition of INSEE and geographical codes (postal codes). Similarly, users of the site could signal errors and propose adjustments to producers (alert geocoding errors, missing or incomplete address, missing data), setting in motion a drive towards continuous improvement of data quality.

Collaboration can improve the quality of existing data. It may also be a lever for data production. The national address repository (BAN) results from pooling of data from the IGN, La Poste and data produced by OpenStreetMap contributors.[34]

This positive momentum must be organised : The user must be able to access documentation ; he may even be involved in its drafting. Collaborative organisation may help in cleaning files. In fact, there is little that individual users can do on their own when it comes to correcting sets of incorrect data. We would greatly benefit if for each dataset, we'd create a shared space for dialogue and exchange of codes to ensure that those who want to go collaborative can.

[30] Chief Data Officer (2015): Analyse the energy consumption of public buildings, available on agd.data.gouv.fr
[31] « Severe Violator Enforcement Program », US Department of Labor : https://www.osha.gov/dep/
[32] http://www.economie.gouv.fr/dgccrf/sanctions-delais-paiement
[33] See in particular the Vigie Nature program led by the Muséum National d'Histoire Naturelle (http://vigienature.mnhn.fr)
[34] http://adresse.data.gouv.fr

Regulation through data, a new form of public action

Public authorities are facing new challenges : Online or mobile services have a direct and immediate impact on well-established economic sectors, whether it's mobility (Uber, Blablacar) or short stay accommodation (Airbnb, Bedicasa).

Most of these companies would not exist without their data, given the central place they occupy in their business models. They provide the indispensable ingredient to trade : trust, by analysis of the transactions and reciprocal rating of participants. Confidence that the potential customer must have before purchasing the service offered by a third party is not certified by labels or diplomas but by analysis of past transactions and reciprocal rating data translated into grades.

The data is also continuously used to improve the service : Uber can predict the areas where demand will be the greatest at a given time and hence encourage drivers to go after those spots or modify the prices on the basis of supply and demand. Airbnb constantly analyses users' search history, and therefore knows what it must recommend to meet their preferences. It can also help the guests to decide the best rental charge (at least the one that maximises the income of the platform).

These new activities require a new form of regulation. Regulators, both national and local, are increasingly aware of the importance of data in these business models. New forms of regulation are appearing : "regulation through data" or "Regulation 2.0".[35]

The first method is to exchange data against the authorisation to engage in a territory. The city of New York has relaxed conditions of exercise of Uber in exchange for data on journeys, drivers and the demand for mobility in each point of the city and at all times. With this data, the city may move from a system of ex ante control (ex-ante authorisation to exercise by means of licences) to a subsequent moderation (continued validity of that authorisation by the data analysis).

The municipal authority of San Francisco seeks to combat the "gentrification" of some neighbourhoods blamed on Airbnb. It has recently set up an office dedicated to very short duration renting. Its goal is to encourage owners to comply with local laws which provide that hosts cannot rent their accommodation for more than 90 days per year without being present on site. To monitor this, it should be possible to access the company's data, which Airbnb has so far refused to share. Here, data is a negotiation tool in the balance of power between these platforms and the cities in which they operate.

The second form of regulation through data is for the regulator, to play an active role in the emergence of digital platforms. In France, Article 1 of Law No 2014-1104 from October 1st 2014 on taxis and transport vehicles provides for the establishment of a geo-located taxi register. he idea is to allow taxis to perform pick-ups thanks to digital tools. Thanks to this register — and to apps using it, it will be possible to book a taxi immediately on one's smartphone, independently of its dispatch centre ("all customers can see all taxis').[36] The le.taxi platform, developed by The Secretariat-General for Government Modernisation and the Ministry of Interior, is currently being deployed. It marks an important step forward in regulating through data. By organising the data flow (geo-positioning of vehicles) and asking for strong rules (free and neutral platforms), the State adapts its role to the digital revolution.

[35] Grossman N. (2015) : White Paper : Regulation, the Internet Way. A Data-First Model for Establishing Trust, Safety, and Security | Regulatory Reform for the 21st Century, Mimeo

[36] See http://le.taxi/

2

The lack of data governance as a hurdle to fullfill the potential of data

For a long time, the State has organised its operation around scientific and administrative data. However, the organizational and technological strategies and the rules governing data pertaining today result from organizational choices, as well as technological and legal frameworks prior to the ongoing digital revolution. For understandable historical reasons, it was just like all big organisations, more interested in the construction of undeniable knowledge than in the dissemination of data for greater purpose. Focused on reliability, safety and cost efficiency, it ignored the needs for interoperability, accessibility and usability, therefore tolerating a culture of silos, differing formats, degraded quality, excessive subcontracting and an overall loss of sovereignty and autonomy over its own data.

1. THE UNAWARENESS OF AVAILABLE DATA

As of today, no one knows precisely the size of data held by public administrations. This unawareness is the first obstacle to the full exploitation of data by the public authorities : What is not known is not mastered.

This can lead to a loss of opportunity as it deprives authorities of valuable and de facto available information. This is all the more important at a time when it is increasingly easy to cross-check data and multiply its potential. No one can identify all the information detailed at the municipal level and this restricts the analytical capabilities of territories and the possibility of observing new cross correlations.

Furthermore, ignorance of data possessed by the administration — or the difficulty to access it- sometimes leads to duplication of production work. Thus a repository of geo-located addresses was constructed in parallel by the INSEE, the DGFIP and La Poste.

This should not obscure the efforts undertaken for several years in attempt to draw up such repository.

As early as 1978, the article 17 of the CADA law introduced an obligation to keep a register of public information. The idea is to list all documents produced by an authority in a single entry. Today, the registers are predominantly published for documentation : very relevant to statistical study and research, they are much less so to identify the various existing databases within an administration. The mapping approach focusing on information systems ("digital land use planning") identifies the main softwares and the corresponding databases by function. On the one hand, an approach based on the document, on the other hand, an approach by the IT system : both are necessary but neither are sufficient.

Why is it difficult to get a comprehensive overview of the data controlled by the State ? This first year of exchange and cooperation with administrations helped to identify the main difficulties hindering such an exhaustive survey :

- Awareness varies according to the stakeholders. The major data producers have organised themselves accordingly. But today number of administrations sometimes indirectly produce data without considering them as such and without questioning their potential usefulness for third parties;
- A difficulty in distinguishing the data from the system producing it : The data is sometimes so closely integrated into the information system that it becomes very difficult, if not impossible to extract. This issue, linked to the design of information systems will be developed in the next chapter. It is common in the ministries, to designate a database by the name of the software producing it (SIV, PATRIM);
- Considerable differences in the way data is stored and disclosed, from tabular files to databases associated with work software. This diversity sometimes makes the identification of relevant data by the producers themselves more difficult;
- The same source may contribute to several different databases, managed by different actors. It is not always easy to establish the origin of data;
- Multiple databases on the same subject ; there are as many bases as there are different approaches and possible purposes. The health data mapping established by Etalab identifies no less than half a dozen databases on cancer, without there truly being any redundancy. Some are related to the follow-up treatment of long-term patients (cohorts), others to the management of hospital care. Each has its own purpose.

The importance of data management

Public policies are often based on "data of authority", the data used by policy makers who are aware of its existence, its uses, scope, and possibly the risks of misuse.

However, as a result of the digital revolution, the majority of existing data is now produced in large computerized management systems, and is not known and identified as such.

A famous story in the open data community tells of this large municipality looking for data concerning cultural practices for its new Open Data Portal. It took almost a year to realise that the municipal library management software held a treasure : The list of borrowed books on a daily basis formed a sociology of cultural practices, helping to understand seasonal trends, identify new correlations between types of books, recommend books to borrow, etc.

The opening of this data raises new questions in return. Such data even anonymous could potentially indicate phenomenon of ghettoization and disclose personal information. This raw data was not used in scientific work and policy communications. It was necessary to make the most out of it and analyse its exact scope before finally releasing it.

Such data from major management systems have become a key subject of data governance. They ask some open questions. It is crucial to learn how to "own" and integrate them into the usual process of the administration.

2. THE STATE'S IT SYSTEM IS NOT DATA-ORIENTED

The architecture choices predate the data revolution

Over the past year, The Chief Data Officer has worked with many authorities, and has initiated concrete projects with a lot of them. This bottom-up approach has shown how access to relevant data is difficult to administrations themselves.

In most cases, difficulties arise from IT system design choices based on the constraints and priorities prior to the development of new data uses. In particular, the IT software and their databases are often designed and optimised to meet one or more objectives in a set period of time, to the detriment of other objectives related to the re-use of data. These difficulties are, inter alia :

- A heavy technological burden. The data is often being produced or treated in very old systems for which development skills are very rare, and therefore expensive, and in which the reversibility of data has not been sought after;
- A structure lacking foresight of sharing needs. For example combining data that can be shared and reused with data covered by legal secrets (e.g. when personal data are mixed with the whole system), or missing essential metadata (like the rights associated with the data) which prevents the extraction;
- A structure organized around the needs of management applications: omitting unstructured data extraction for unexpected uses;
- The lack of accessible reference material on 24 hour/7 day basis. Even when administrations agree to transfer their data, they are often perceived as information (or files) that need to be transmitted at regular intervals, therefore not ensuring an optimal data freshness. Only a few administrations have taken remedial strategies such as platforms and APIs*, and organised themselves to make their reference data available to other applications (for consultation and real-time synchronization).

This situation is neither a public nor a French specificity. It reflects the history of information technology, undergoing a continuous digital revolution.

Not so long ago, computing was designed on paper, to minimise innovation oriented investment, and used as a static resource at the service of an organisation. In particular, its value was rarely estimated (and therefore rarely steered) according to its transformative potential for an organisation or for the value chain. This situation has generated a portfolio of original applications fragmented, with an excessive reliance on service providers, increasing opacity of expenditure, and hence the major failure of some "great computing projects".

This is not only harmful to the proper use of data in the conduct of public policies. It led to an overall loss of the capacity for action of the State, raising complex issues beyond computing problems. As underlined by Michel Volle,[37] information systems reflect as much as they shape the processes of organisations. Today, the need to exchange information between public officials, to cooperate in the production or improvement of such data, is deeply foreign to the design of information systems as much as it is the administrations themselves.

The appointment of the DISIC (inter-minsterial director for information and communication systems)) in 2011, has launched the process of response to this situation. The creation of a Chief Data Officer in 2014 and the integration of the DISIC in the inter-ministerial directorate for digital, information and communication systems (DINSIC), which, since September 2015 also includes the Chief Data Officer, will make it possible to combine in a single strategy the back-office and the front-office, infrastructure and user experience, thus opening up new opportunities. In particular, it will create, through inter-ministerial action, a State IT system in line with the best current practices : extreme availability of critical applications, agile innovation capacity, flexibility, open system architecture, continuous improvement of expenditure, management of service-related costs...

The strategy of the State as platform drafted by the DISIC and the ministerial IT departments in 2014 and 2015, lays down the fundamentals of this digital revolution in public computing, favouring the interoperability of systems based on platforms and APIs.[38] However, there is still a long way before it can become a basic principle of the State IT.

The State does not retain sufficient control over its information system

Various constraints facing the actors in the information system of the State (deadlines set for completion, difficulties related to human resources, etc.) have sometimes led to the loss of independence of the State in favour of third party actors in mastering its information system :

- Subcontracting is sometimes made without creating or maintaining an internal team with knowledge of the products created, able to steer the service provider and call into question its proposals, to implement a change of subcontractor without risk to the continuity of the service, and if necessary capacity to internalize certain tasks (e.g. carrying out extraction and data manipulation, without being charged by the-task with prohibitive pricing and deadlines).[39]
- The property of the State on its data may receive an expressed or implied weaver : when, it becomes technically impossible to recover stored data in an information system, when the localization of reference data is no longer available or accessible when the tools to address them no longer exist, when their migration in open and up-to-date formats is no longer possible, or even when the State agrees to a partial loss of intellectual property (rights abandoned to suppliers, sometimes include "intellectual property" of some dimensions essential to data use : metadata, database structuring, data, and logical lock-in to freeze owners and thus the State's ability to use its own data).

This also has an impact on the security of the information system of the State.

Not being able to easily recover data is an overall regrettable and sometimes dangerous loss of efficiency and responsiveness. Thus, in the treatment of a cyber-attack in a ministry the safety data logs were only available through the subcontractor, subject to payment and with a significant delay, which slowed down the analysis. Information system security can call for a massive data processing capacity.[40]

The outsourcing of data hosting or processing may also reveal security vulnerabilities. It is essential to support the widest possible opening for data that as to be shared (open data), however, for data that is not meant to be open on behalf of legal secrets or security features of the CADA, it is mandatory to continuously question where it is hosted and treated. This attention will become increasingly important to the development of cloud-like services, but also with the imminent development of machine learning "as a service", where it will become possible within a single click to remotely deal with massive amounts of data for a few thousand dollars.[41]

The issue of the independence of the State and its capacity for action must once again become the central issue in the design of information systems, and in particular, construction of the legal and contractual framework.

Independence does not mean isolation. Independence is the capacity to define one's objectives and the means to achieve these objectives. Each public purchasing decision, each delegation of competence, each decision of subcontracting should be approached in the light of this central question : Does the public authority retain full capacity to act ? Recovering the data produced by the delegated activity, accessing the software source code, building bridges of interoperability between systems, maintaining the necessary human capacity to steer the fate of a project... all these ambitions should be at the heart of the purchasing decisions and specifications of the IT projects. The ability to use data towards public action depends on them, and many other abilities do as well.

[37] Volle M. (2006): IT : life with automats , Economica
[38] http://references.modernisation.gouv.fr/strategie-du-si-de-letat
[39] Corrective actions have been initiated; it includes project evaluation criteria on the basis of which the DINSIC delivers opinions while conducting audits.
[40] Source: Interview with ANSSI
[41] Morin-Desailly C. (2013): The European Union, colony of the digital world? Report drawn up on behalf of the Committee on European Affairs of the Senate

3. THE ADMINISTRATIVE CULTURE DOES NOT ENCOURAGE SHARING AND COOPERATION BETWEEN ADMINISTRATIONS

The difficulty of sharing data within the administrations

No one will be surprised to find in this report, that our administrative culture is far from being based on cooperation and data sharing.

There are several reasons for this : The fear of violating statutory secrets, the general feeling that "file exchange" is forbidden, fear of the hierarchy's reactions, fear of being overwhelmed by demands while the means necessary for routine missions tend to decline... all these reasons are legitimate and deserve serious answers.[42]

But they must not force us to lose sight of the central objective. Inter-ministerial cooperation with the decentralised services, and with local authorities, is still not sufficient. Information systems have been designed in logic of ministerial silos financed under separate budgets and this ultimately penalizes the effective functioning of the State.

As this is the nature of data itself : Even if it is produced by an authority for its own goals, it can serve many other purposes. The success of open data, this radical way of sharing is a testimony to this : the database of road accidents, time stamped and geo-localised is not only relevant to the police but also to urban planners ; groundwater pollution should be compared with river pollution ; house prices are interesting to the Ministry of Housing and to the tax administration. Many more examples exist.

The main issue is not so much the individual conduct of the agents — many of them are convinced of the benefits of cooperation — but in the administrative culture itself, as a set of norms, standards and practices.

Administrations are often reluctant to share data because they consider, often wrongly, that data production is so closely linked to their role and their activity that they would serve little purpose to third parties. In the absence of an actual ownership right, it is common for data producers to claim a moral right to it. The fear of misuse or misunderstanding of the data — or fear of highlighting their insufficient quality — is often invoked.

This lack of data sharing has very important implications for the State, both with loss of efficiency and effectiveness. Data essential to the steering and evaluation of public policies are thus not made available to all stakeholders. Administrations sometimes replicate already existing databases of which they are not aware. More generally, the State deprives itself of the expected benefits from better data circulation and usage.

In emergency situations, as we have recently experienced, the ability to exchange data rapidly and to interconnect can become extremely critical, particularly in the field of civil protection, public information and real-time information collection. It seems essential to quickly organize an inter-ministerial work on how to prepare a capacity for rapid exchange of data in times of need.

[42] If it is framed by the law "Law on IT and Liberties" for understandable reasons of protection of privacy, it is not prohibited. In addition, the Guidelines concern only the files containing personal information. As such, the principle of sharing and passing of non-personal information poses no problem. The cross of personal data is not impossible but subject to prior authorization by the CNIL.

Referrals to the Chief Data Officer (CDO)

The decree No 2014-1050 from September 16th 2014 provides that anyone can ask any question they have on data held by the state to the Chief Data Officer. Local and regional authorities, legal persons governed by public law and legal persons under private law entrusted with a public service may consult on any issue linked to the use of data by administrations.

A dozen referrals were received during this first year of exercise:

- half of them come from individuals or companies developing projects using government data;
- four referrals are from public administrations (mainly from local authorities);
- The remaining applications come from a non-profit, a journalist, and the Court of Auditors.

These referrals illustrate instances of poor data flow resulting in an inefficiency of authorities :

- Communities are denied access to data on individual agricultural parcels which would enable them to better coordinate the fight against environmental pollution;
- The stakeholders responsible for the prevention of accommodation unfit for habitation do not have data on building fires which would be useful for them in order to better assess the risks associated with each house;
- The public institution responsible for managing the assets seized and confiscated (AGRASC) do not have access to judicial, administrative and financial data essential to the effectiveness of its public service mission. This is slowing down the handling of cases and reduces the institution's potential to tackle the State's general budget.[43]

Furthermore, The Chief Data Officer delivered an opinion on the publication, rectification and re-use of information on health professionals. This opinion responds to a referral of the Ministry of Health and Social Affairs as part of the preparation of the draft health law.[44]

[43] Extract from the annual report for 2014 from the management and recovery of the assets seized and confiscated agency
[44] The referral processes as well as the opinions are published on the website of the CDO: agd.data.gouv.fr. In 2015, around a dozen referrals have been received and dealt with.

4. MANAGING BUDGETS HINDERS SHARING AND COOPERATION BETWEEN ADMINISTRATIONS

The benefit from sharing data is often collective, uncluding other public authorities or other actors in the economy, while the necessary efforts to share data rely on the producer (or distributor) alone, who won't usually see the short and medium term interest it has in sharing.

How to reconcile budgetary management and data sharing ?

Making data available to third parties represents a cost and sometimes a lack of profit, for administrative entities. However, these entities often act in a very restrictive and very vertical system. Such is the spirit of the LOLF which frameworks programs on an area of actions, objectives and means. However, those objectives almost never include making data available to third parties. For the head of the administrative budget of the entity the free -or at a very low cost- data sharing means an increase in costs, without any additional income, and almost no impact on objectives as part of the tasks assigned to him.

In addition, if the costs of openness can be anticipated in the case of open data (exporting files not generating additional costs if the data is accessed or used by a large number of actors) it is more difficult to anticipate the possible success of a controlled opening and a number of authorities fear that they will become victims of their own success. One possible answer to address this tension is to include the task of data dissemination within the tasks of the authority, to empower decision-makers in charge of budget management. This approach may also need to supplement budget allocations to the concerned administrative entities to ensure data sharing.

The government sells itself data

The sale of data by government departments or other administrative authorities and operators to other state agencies is a significant financial failure both in terms of economy, and efficiency.

Firstly it is unproductive from an accounting point of view. Indeed, it is not a zero-sum game at State level, but a net loss related to the costs of those transactions. There is, beyond the sale of some big repositories to a few large customers, a multitude of internal transactions of low value (less than EUR 500) which are carried out annually between administrations, authorities and operators. Processing (invoicing, accounting and administrative follow-up) and regulating such transactions incurs costs disproportionate to the amounts committed.

The central State itself is familiar with this situation where the major producers themselves are forced to find sources of income to complement public funding granted to them. The business model of those producers should be reviewed to ensure that the common benefit is maximized. The decision to make the large-scale reference system (RGE) free for public service missions was a first step in this direction.

Selling oneself data also represents an opportunity cost for the State. Some administration choose not to acquire data which would be relevant to their tasks ; others are building their own databases to become independent from third parties, risking duplication and wasting of scarce resources.

Finally, this hampers the aim of providing the State with a single information system, consistent and of quality. In particular it generates an amount of duplication, copies of datasets, and de facto errors.

The findings of the Fouilleron Mission

At the request of the Director of the Office of the Prime Minister, Mr Antoine Fouilleron, auditor at the Court of Auditors, carried out an in-depth study on these sales of data between administrations, handed over to the government at the end of November 2015 ; it leads to the following recommendations :[45]

Proposal No 1 : Establish the principle of free exchanges of data between administrations under their public service mission in the law and only provide for paid exchanges if there is complex data treatment and survey co-financing.

Proposal No 2 : Reaffirm, within a circular from the Prime Minister, the principle of completely free exchange of data between authorities, including for complex data, and extend this principle to relations between the State and its operators to carry out public service tasks.

Proposal No 3 : Ensure the neutralization of the budgetary flows established in regards to paid data exchange of exchange by data base transfers in the Finance Bill for 2017.

Proposal No 4 : Accompany the implementation of the principle of free exchanges of data between administrations through the deployment of infrastructure and services which support the standardisation and normalisation of these exchanges. Draw up a standard license for exchange of data between administrations.

Proposal No 5 : Deepen the analysis on non-budgetary bottlenecks to smooth the data flow between administrations, build a databases directory of administrations and objectify the legal constraints that might restrict dissemination of data covered by legal secrets.

[45] Fouilleron A. (2015): Paid data exchanges between administrations, report to the Prime Minister

5. DISINCENTIVES STEMMING FROM THE APPLICATION OF "LEGAL SECRETS"

"The CNIL will never agree.", "There is statistical secrecy." "This is under fiscal secrecy.", "We might breach confidentiality agreements." ,"Transparency opposes the medical confidentiality." "I need hierarchical instructions." "The data is not that good and I will be responsible for the errors." "I might be allowed to transmit data but I am not obliged to and I do not know what my superiors would think." "This application is to be examined by the administrative court"... anyone wishing to publish or simply to use public data often encountered some of these replies and many more.

It is disappointing and damaging to see the CNIL thus exploited even before any referral by the administrations, even though the CNIL allows almost all data treatments, provided that their conditions of implementation, including security, are satisfactory.

The 'legal' secrets" are numerous. Each one of them has its own history, its reference framework, its scope and its limits. Protectors of fundamental freedoms, fundamental interests of the nation or necessary for the exercise of State functions, they are legitimate. Some of them are even likely to be strengthened at a time when computing power is released into society as a result of the dissemination of personal computing coupled with the emergence of big data*, making it possible to deduce new information from seemingly mundane data. Most of the concerns are legitimate. The fact remains that their legal basis are frequently questionable, as they tend to confuse different issues and heterogeneous aspects. This climate of ambiguous anguish becomes an obstacle to good data governance in the efficiency of public policy.

Several studies and reports have pointed to shortcomings and even contradictions, of the legal and regulatory framework. Those findings are important, but after some years of experience, they appear secondary in recognition of the approximate application of legal secrets.

Uncertainty in the application of legal secrets

A secret is not the destruction of information.

On the contrary, a secret is information which is known to some, and which must not be transmitted to others.

In a democracy, this barrier was erected to either protect people or the nation's fundamental interests.

The most important, looking at a legal secret, is therefore who it opposes, and under what conditions. Thus, medical confidentiality isn't opposable to patients, but to their close relatives, it does not concern the doctor-patient relation. Professional secrecy does not protect the professional's secrecy but the secrecy of whoever discloses information to him. The military confidentiality clearance does not give access to all classified documents — access is provided on a need to know basis, in order to limit the risk related to information, that if compromised would be likely to harm the nation's interests. The "statistical secrecy" is not a prohibition to produce statistical results on individuals. It was based on a specific question : The obligation made to companies to submit information, laid down by the 1951 Law in return for which the State swore not to use this information to monitor the companies' undertakings and on the other hand not to reveal such information in a manner which would jeopardise the confidentiality of business secrets.

Moreover, it should be noted that, over time, the legal protection of trade secrets settled in certain habits, which have gradually extended the circle of initial limits, or did not follow the changes in the data, practices and uses.

Over the past year, The Chief Data Officer was denied many data on the ground that the CNIL would not accept their transmission. Apart from the fact that it was the CDO himself that was required to obtain authorisation for treatment of the CNIL, effective verification showed that nine times out of ten, the CNIL had not been consulted and did not oppose access to this information.

And the list of these approximations could be extended without difficulty to tax secrecy, business confidentiality, criminal liability of public officials in the case of disclosure of a trade secret, statistical confidentiality, etc.

Today, the shady perception of statutory secrets has more impact on the functioning of the state than the secrets themselves. It is essential for the proper functioning of the State, that the rule of law, and only the law, is replaced at the heart of data exchange practices and to teach administrations that cooperation should be the rule and that the limits imposed by statutory secrets must be applied in full respect of the spirit of the law.

Some necessary adjustments

After a year of experiments, the CDO considers that the vast majority of difficulties faced in data-sharing come mainly from excessive precautions that extend unduly the scope of statutory secrets. However, it is not impossible that certain legal difficulties come from legal frictions that could be corrected.

Opinions differ, e.g. regarding the relationship between the law on access to official documents and the law on information technology and liberties.

These two texts, in fact, present different reasoning, and handle overlapping concepts that are not to be confounded.

The law on information technology and liberties, adopted on 6 January 1978, regulates the automatic processing of personal data. It responded to the risks found in nascent computerization.

The law on access to administrative documents, adopted on 17 July 1978, lays down the right of access to administrative documents, progressively enriched with a right of re-use. The transparency of public action is of course limited by other interests protected by law, such as the intellectual property of a third party, the fundamental interests of the nation and privacy. To ensure the protection of privacy, it provides that access to information whose disclosure would undermine the protection of private life, medical confidentiality and commercial and industrial confidentiality is barred, save to the interested parties. Over time, the case-law of the CADA has defined the concept of private life and in particular that some personal data of public figures, for example, were not concerned by privacy.

The CADA Law then adds another security on re-use stating that "public information containing personal data may be made available for reuse when the person concerned has consented or if the administrative holder is able to render it anonymous or, if lacking this anonymization*, a legislative or regulatory provision authorizes it."

The competent authorities, and in particular the CNIL and CADA, have developed a set of rules enabling them to coordinate those two texts in accordance with their respective missions.

The question of their application to open data has been discussed on numerous occasions, notably in the Council for public publishing and administrative information guidance and administrative information (COEPIA) who published a memento on the protection of personal information in the context of openness and public data sharing, or between the CNIL, the CADA, the Office of Legal and Administrative Information (DILA) and the mission Etalab to ensure all the safeguards on the re-use of large databases of case law. [46] [47]

But the fact remains that the coexistence of these two approaches, with regard to a third issue, namely that of the movement of data within the administration, generates much hesitation within the administration and would deserve further development and real political inter-ministerial leadership.

[46] Council for public publishing guidance and administrative information (2013): Guide to the protection of personal information in the context of open data
[47] See, for example: https://www.data.gouv.fr/fr/datasets/cass/

6. SPREADING DATA SCIENCES

Good data governance is a prerequisite for the development of data sciences but, conversely, the development of data sciences is without doubt essential to efforts that require good data governance.

Indeed, within a context of strong administrative and budgetary constraints, it is difficult to engage any organization in an effort, without showing them in practice, how these efforts can quickly make it possible to simplify or to optimise their daily operations and the success of their missions.

In other words, the construction of a better data governance will be driven by usage, and development of uses is an integral part of the strategy to build a better data governance.

Therefore, The Chief Data Officer began its work with a small team of four data scientists, offering its services to administrations and which in under a year has succeeded in producing a number of encouraging results with the support of voluntary administrations. Attention is drawn in particular to the following results :

- Work with the State Purchase Centre, to analyse the power consumption of the State and thus lead to a firmer control on purchases;[48]
- Work with the technologies and information systems of the security department has helped develop a prediction model of car theft at the department level;
- Work with the teams of Pôle emploi to predict with 80 % accuracy when a company will recruit a given profile in the next quarter, which enabled Pôle Emploi supported by SGMAP to develop "La Bonne Boîte".[49]

These initial experiences, which will be continued during the year 2016, demonstrate the importance of the paradigm shift represented by the arrival of data sciences and "big data". This change mostly comes from data users and use of data for action". Mathematics used by traditional statistics and those used in data sciences are basically the same, even if new tools and methods are regularly updated. The Institut Mines Telecoms (IMT) and the Groupe des écoles nationales d'économie et de statistique (GENSE) have begun to teach those data sciences and big data* approaches. It is much more in the public action field that a fundamental change is happening and is sometimes proving difficult to accept.

The digital revolution is a new logic of action based on data. This logic of targeted action differs from scientific logic, geared towards the production of certain, verifiable and reproducible information. Thus, data scientists are more likely to accept working with imperfect data than traditional scientists, given that their action-oriented logic integrates this uncertainty. For example, The New York fire department bases its fire prevention visits on an algorithm, whose detected correlations might not be causalities. But each week, they make sure it continues to be effective. Data is increasingly used in the civil servants action at the counter and not to establish a strategy imposed by the hierarchy. Finally, data that used to serve as a tool for knowledge is increasingly used for real-time decision making or by the public service users themselves.[50]

These attitudes and action strategies are still scarce in the administrations (as they are also in private companies). Their widespread use is both the objective of the CDO, and the most important motivation for an administration to improve its data governance.

[48] This work is documented by the STC and on the site of the CDO : https://gd.data.gouv.fr/2015/05/17/analyser-les-consommations-energetiques-des-batiments-publics/
[49] http://labonneboite.pole-emploi.fr/
[50] For example, the winner of the hackathon organized by the CNAF and Etalab for the 70th anniversary of social security was a team that proposed to work with statistics of visits at the offices of the family allowance funds to offer users to come together according to their expectations, so that their visits are both an opportunity for mutual assistance between entitled persons and an opportunity for the CAF to mobilize staff specifically trained on the issues in question.

3

First guidelines for a good data governance

The first part of this report underlined the quality and accessibility requirements of State-produced data and the potential of data sciences, not only to evaluate and guide public policies but also to deeply reform them.

The second part highlighted the main obstacles on this journey, technical, organizational, cultural and juridical ones.

This last part will touch on the necessity of a long-haul labour, both interdepartmental and focused in order to reach a significant transformation. Just like the DISIC, with the different ministries, instigated a unified State IT strategy, such an effort will be necessary in building a State data system.

This effort will be led by the Chief Data Officer and punctuated by the annual report on data governance provided by its founding decree.

Nevertheless, it is already possible to engage in the following months in projects aimed at first substantial improvements.

1. START WITH CONCRETE DEVELOPMENTS

In the digital world, focusing on solving measurable concrete problems very often is the best strategy. Many objective reasons exist to justify this change, and Mike Bracken, the historic figure of the British Government Digital Service, has done a great job at presenting them on his blog.51 In a famous article "The strategy is delivery, again" he explains that too many digital projects start from an abstract political vision we try to translate in processes onto IT systems, then look to find users and settle over time. On the contrary, lean digital processes start from user needs, come up with new services, and build their infrastructure based on those needs, reshape their service in accordance with political decisions and organize the permanent feedback between political ambitions and user opinions.

This belief guided many projects led inside the SGMAP:

- Concrete and operational data sciences projects by the chief data officer's team;
- Development of APIs and geographic (api.carto) and company (api.entreprises) data access interfaces now used by dozens of administrations and built into many services;
- Support to simulation models (Open Fisca);
- State start-ups like simplified public procurements, Mes Aides, La Bonne Boîte or Le Taxi who, trying to successfully solve concrete problems, also try to create open APIs or resources build to be used by other systems;
- The creation of France Connect, a resource open to all administrations, enabling citizens to connect to any public service, but, more importantly, to themselves provide for data exchange between different systems to simplify public service;
- Support to collaborative projects (like the National address repository).

Those developments are small compared to the size of the State's information systems. They anchor the data strategy on actual practices and progressively create a network of usages and interconnection possibilities inside the IT system, more interoperability, more agility and more action capacity, in permanent link with users.

The DINSIC will continue and amplify this concrete developments strategy to streamline the State's IT system, notably through government as a platform. It will deepen and strengthen its support to similar initiatives stemming from other departments.

2. REVEAL THE AVAILABLE DATA

Mapping all of data available in the State's IT system is a daunting task, often told to be impossible.

The 1978 CADA law (article 17) does project that administrations must frequently release a repository of the "administrative documents" in their possession. But, although the case law from the Commission on Access to Administrative Documents States that files, databases and sometimes software are considered "documents" within the law's reach, most of the repositories only deal with published documents and leave out data and information. It is true that 1978's context greatly differs from 2016's. In four decades, we went from documents being neatly organized and hosted on a limited number of computers with a clear supervision from public servants, to a general system computerization, massive distribution of cells and an explosion in management data, big data and real-time data feeds.

The DISIC initiated work on the State's IT architecture, and was able to make a "land use plan" which identifies the main families of data and provides initial guidance to data researchers.

In the context of work on the opening of health data, ETALAB proposed a first inventory of the main databases available in the health system.[52] This project has made it possible to measure the difficulty of an approach based on the survey of administrations, as well as the difficulty to describe in a single benchmark the database content, the sources of data, their accuracy, granularity, and frequency of updating, their owners and the various secret protecting such data.[53]

These experiences have demonstrated the difficulty of any approach that is too linear, centralised and exhaustive.

This difficulty is compounded by the fact that, as already pointed out, many important information come from the management system and are rarely seen as source of information by the authorities responsible.

Therefore, The Chief Data Officer will launch, in 2016, a data mapping project that will be collaborative and open to all administrations willing to participate and take benefit of it.

[52] See https://www.data.gouv.fr/fr/datasets/cartographie-des-bases-de-donnees-publiques-en-sante/
[53] It has taken nearly three interviews with each administration concerned to achieve the desired outcome.

3. HELPING THE STATE'S IT SYSTEMS EVOLVE

The report highlighted that the ability to use one's own data depended on IT mastery. It also stressed the birth, in all big organisations, of next generation computing, not only thought as a static tool serving the pre-existing organisation, but rather a way to address new challenges :

- Capacity to deliver quickly and in an iterative manner the new digital skills;
- Minimising operating costs of previous applications;
- Meeting the level of service required by the digital era, for example in terms of simplicity, security and availability.

The State's platform strategy, prepared by the DISIC and the ministerial departments, will be at the core of this development. It should guarantee specific requirements for data use.

It should, in particular, ensure data extractability by design. The choice of architecture and metadata should aim to maintain the capacity to separate the data according to the limits on their use.[54] It will also be important to develop tools or efficient extraction query methods (without being obliged to use an ad hoc provision necessarily limiting market) and to modify the process governing the creation and updating of databases. The ambition of exportability of data should become an ongoing priority for the design of the new systems.

This strategy will also focus on the choice of architecture and governance to advance towards the use of data in real time (often real time data are only assessed once per year and are therefore not usable on-the-run).

New audit rules of State IT projects will be offered to the IT departments and then systematically added to the process of monitoring projects by the DINSIC, such as control of :

- Freedom of data exploitation and associated models: ensuring that participants in the projects, and in particular the software providers and publishers do not pursue intellectual property rights that may limit the ability of extraction, use and/or re-use of data;
- The capacity of an IT system to distinguish data subject to statutory secrets and privacy, and those which are not subject to those constraints;
- The technical conditions of extractability: the capacity of an IT system to allow extraction of data in a format that is reusable, by means of a "full" dump, and, when appropriate, using an open interface like an API or owner-free web services;
- The use of benchmarks, and existing ontology to promote the re-use of existing systems rather than creating parallel bases;
- Conditions for access to data for purposes of data sciences but also for other administrations (movement of the data within the State);
- The ability of the IT system to supply, if possible in a direct and machine-readable way, open data portals like www.data.gouv.fr.

Finally, the DINSIC must also ensure access to resources for all ministries allowing them to test the practical potential of data sciences (for example: give access to capacity calculation: high-performance computers/servers and the possibility to install and work with open resources like the language R or Python with pandas or scikit-learn libraries, all of which are massively used by data scientists.

[54] For example, the aggregation in the same system of data covered by secrecy and data not covered has again recently, prevented data mining, which could have allowed substantial savings.

4. BREAKING DOWN ADMINISTRATIVE SILOS

To realize the value of data, it is essential to remove artificial barriers between departments and to promote inter-ministerial collaboration. In this context, The Chief Data Officer will build over the coming months on projects in progress. Some of them aim at providing a general framework for such cooperation. The first one addresses the sale of data between administrations. The mission on data sale within the administration entrusted by the Prime Minister to Mr. Antoine Fouilleron helped quantify those sales as well as the transaction costs and the associated loss of opportunities.[55] The Chief Data Officer within the SGMAP, will accompany data producers in transforming their business models.

This first task will provide a framework more favourable to the free movement of data and its better use by the public authority. However, breaking the barriers between administrations cannot be decreed: it must be practised. Funding related to the Future investment program (PIA) will be a privileged tool to promote inter-ministerial collaborations. 21 projects were selected under the call for proposals on the industrialisation of open data.[57] The ministries were strongly mobilised on this occasion, notably through a dedicated event "Project Camp"

The Chief Data Officer organises a network of correspondents and experts throughout the administrations. This network is aimed at all those engaged in (or wishing to participate in) data sciences within ministries and, potentially, at local and regional levels. It aims at promoting the exchange of good practices as well as to provide these officials with a necessary good-will and care needed for them to perform their tasks. The year 2016 will allow for the structuring and consolidation of the network. Various initiatives to establish data administrators in ministries or amongst operators are being studied. The Chief Data Officer supports those initiatives if they observe two principles :

- Assume this function in the spirit of the chief data officers many private companies and many authorities have,by nominating a person specifically in charge of data-based decisions, or data-based public policies. The aim is not to create an additional layer of control, but to lead to a transformation of the action, which must therefore be given a quality of data, a capacity to extract and handle data, legal certainty, technical and data sciences skill;
- Organise ab initio the networking of these initiatives around the CDO so that these various initiatives give rise to a genuine intelligence and collective action.

What financing for major data repositories?

There are three models for financing major data repositories:

The producer shall charge the person who must register: for example a firm or an association can be charged for registration and/or publication ;

The producer shall pay the re-user, according to a model of royalties, mostly related to the degree of use ;

— the producer is financed directly by the public service, like the "basic data" approach adopted in the Netherlands and Denmark.

No European country strictly applies one or other of those three models; they rather use differentiated approaches by types of data and/or producer.

The POPSIS, study commissioned by the European Commission looked into the ratio of coverage of costs by revenues from the sale of data from a global point of view (considering the entire budget of the producer and not those directly linked to the production of a database in particular).[56] In half of the cases studied by POPSIS, income covers only 1 % of total budgets of producers, and most of the time the rate of recovery of costs does not exceed 5-15 %

[55] Fouilleron A. (2015): Paid data exchanges between administrations, report to the Prime Minister
[56] http://ec.europa.eu/information_society/newsroom/cf/dae/document.cfm?doc_id=1158
[57] With the creation of the 'I welcome public" database which identifies the public-access buildings (PAB), industrialization of the process of anonymization of health data, the le.taxi, project, the creation of a platform of the geographical nautical information

5. A NEW DOCTRINE FOR THE APPLICATION OF LEGAL SECRETS

Clarifying the application of the doctrine of legal secrets

"There is no rule without secret." Discretion, confidentiality, protection of citizens, businesses and national security are the fundamental duties of the State. Bearing in mind those principles, the State also has a duty to apply the statutory secrets with accuracy and care, in line with the spirit and the letter of the law. As indicated in the second part of this report a secret is not a mystery : It organizes a division between those who have to access information and those who do not need to. This partition is to be made to the exact extent desired by the legislator or the administrative authority which introduced a statutory secret.

The Chief Data Officer therefore requires administrative authorities to give the utmost attention to the doctrine of implementation of statutory secrets of their respective administrations. If necessary it can provide support to the authorities wishing to clarify, consolidate or revise this doctrine.

The Chief Data Officer suggests that for each individual database established and protected by a secret, the relevant administration documents for itself and for authorized third parties what specific data are protected by the secrecy and on what basis.

For example, it might be appropriate, within a framework yet to be defined, to ask the Council of State to specify this doctrine and especially the condition of its concrete implementation by administrations.

However, for the specific purpose for which statutory secrets have been defined, The Chief Data Officer considers that all precautions must be taken and particularly supports the idea that the data held by the administration and covered by secrecy within the scope of the Criminal Code must be stored and processed on the national territory.

CNIL's "packs of conformity"

Among the various statutory secrets, privacy probably holds a special place. it is a fundamental right. The CNIL, an indpendant authority regulates the automatic processing of personal data and therefore intervenes on a number of decisions relative to private life. It it is also one of the issues that matters the most tocitizens, because it is mixed with various different questions, illegal eavesdropping scandal PRISM and other revelations, or the growing awareness of the possibility to predict behaviours thanks to seemingly mundane data, cross-checked with other information. It is therefore probably also the least well known legal secrecy, where you find the greatest contradictions and unfounded assertions.

Aware of this difficulty, the CNIL has launched in 2014 a new tool, the "packs de conformité". *"Packs of conformity constitute an operational response to the needs of professionals regarding the application of the law "Informatique et Liberté" (law on IT and Liberties). Working in close coordination between the CNIL and the players in the sector, legal tools of simplification and facilitation, single authorisations (simplified standards, exemptions etc.) and good practice specially adapted to an occupational sector are being developed. These packs also allow anticipating on the expected changes to the draft EU Regulation on data protection. For the controllers it allows a substantial simplification of formalities in favour of a more dynamic relationship with the regulator. In this sense, the ability to report on compliance with the law is becoming an essential issue."*[58]

This is more generally in the search for new regulatory tools, such as labelling which make it possible to integrate data protection vis-à-vis their customers, whilst being legally secure vis-à-vis the regulator. In three years, the CNIL has therefore issued 60 labels and the label "data protection governance" is, according to the number of candidates (including, for the first time, local authorities) likely to undergo a great success.

[58] http://www.cnil.fr/en/linstitution/actualite/article/article/les-packs-de-conformite-un-succes-grandissant/

The development, with voluntary actors of a compliance-pack tailored for the public authority is an extremely promising path. The CNIL and the DINSIC discussed this perspective, for which several ministerial IT departments have expressed their interest. The CNIL could therefore launch them as early as 2016, in close consultation with the CDO.

Facilitating anonymization

A number of databases held by the public administration contains personal information and therefore cannot be directly published. On the other hand, different methods exist to dispel this personal nature. What's at stake here is "de-personalization" or, more precisely, withdrawing the possibility of re-identification. Once those methods are applied, information can be published.

Today, the implementation of these techniques requires investment in a very specific field. This represents a significant cost for administrations wishing to develop a data-sharing strategy as described in this report. They must then alone establish an anonymization process.

For good data governance, this step should be facilitated by a CDO-led visibility of an expertise-hub on these issues. This hub should include a better grasp of the issues related to data anonymization, not only from a legal point of view, but also from a technical viewpoint (ability to automate the large-scale anonymization). It would ensure the technological monitoring of anonymization solutions for different types of data (data tables, geolocalised data, network data, etc.). It must, as far as possible, provide the tools to measure the degree of personal data in the dataset and help eliminate it. It accompanies the administrations in that process and seeks to provide an open source anonymization toolkit.

6. DISSEMINATING THE CULTURE OF DATA

The data governance outlined in this report serves the transformation of public action. Reciprocally, it will succeed only if administrations are rewarded with their efforts. It is the impact of the data strategies, in terms of effectiveness, cost control, quality of working life that will encourage administrations to engage in the necessary reforms.

The Chief Digital Officer showed, throughout 2015, the concrete possibility of supporting administrations in their own projects and of obtaining verifiable results.

The CGEIET report showed the existence of pockets of competences within the administration and the recurrence of requests for support (both technical and legal, and sometimes of computing power or work environments adapted to big data*).

The year 2016 should primarily help to build new data sciences projects bringing concrete and verifiable results. The Chief Data Officer will work to support all these initiatives, aiding the work of administrations that wish to do so, and to network these skills within the various administrations, to create a community of practices, stimulate learning from peers, pooling resources and contribute to the rise of the quality of all these to the benefice of public action.

The data sciences public procurement contract

In the course of 2015, the CDO prepared a public procurement for data sciences support.

This market falls within the general strategy of the SGMAP to support administrations. A dozen suppliers will be available through subsequent contracts.

Through the SGMAP, administrations will thus have the possibility to access additional resources for data sciences projects.

Various aspects can be treated in three important components of the data sciences :

- *Search and exploration of data ;*
- *Data analysis : detection of predictive models, weak signals, classifications, etc.*
- *Data recovery.*

Contractors will be expected to use public data, open or not. A secure environment for both administrations and citizens has been established for this purpose (including data protection).The National Agency for the security of information systems and communication assesses firms on these aspects in the context of the examination and intervenes in the drafting and awarding of subsequent contracts.

The framework agreement highlights the need for possible skill transfer to the administrations. On the basis of the subsequent contracts, the administration will be fully autonomous by the end of the service. The count shall be carried out under the auspices of the administrator general government data by five different ministries.

The contract will be in place and usable in 2016.

At the request of the Secretary-General of the concerned ministry, the administration may require the SGMAP to examine the appropriateness of the support depending on the nature of the project and the number of applications received.

CONCLUSION

The governance of data is particularly important in the context of a set of important changes :

- changes in the nature of the data produced or held by the administration ;
- changes in the tools and methods for processing such data ;
- changes in the action strategies allowed by these new data and these new methodologies.

This first report shows that the State, like most large private or public institutions is not yet ready to grasp the full potential of such data. Change will be collective, progressive, and will follow the logic of the data-based action.

This report provides initial commitments and first recommendations that can be implemented as early as 2016 :

1. Further development of fast, specific and de-compartmentalizing projects, illustrated by FranceConnect, the National address repository, APIentreprises and State start-ups and summarised by the strategic framework of the State-as-platform. This logic is moving towards a unification of the State's IT strategy, through platforms and APIs, aiming at secure and concrete interoperability, in compliance with informational self-determination;
2. Launch, in 2016, a collaborative mapping of the data available in the State, open to all administrations who wish to participate and benefit from it;
3. Include the ability to extract and use the data as a criteria in the examination of IT projects by the State;
4. Help and develop inter-ministerial collaborations, both through support of the CDO, by prioritizing funding for innovative projects, through the development of thematic APIs (taxes, health, etc.) to connect ministries IT systems;
5. Create, starting from the CDO, collective competence, both technical and legal, in data anonymization;
6. Clarify the doctrine of legal secrets;
7. Request the CNIL to launch the conformity packs with voluntary administrations;
8. Disseminate the new uses of the data by direct cooperation with the CDO or by using the public sector contract the SGMAP prepared for administrations, and sharing results between administrations and with the public.

Those first initiatives will raise awereness around the issues of data management.

They may only be understood in a logic deeply committed to action. The data revolution sets the State, just likes economic and social players, at the forefront of innovation. However, innovation is driven by usage. Citizens and administrations, actual users and their habits select out of the promises offered by technology which ones will become relevant, sustainable and actual practices.

Many other issues need to be addressed in the coming years, such as the democratic accountability of this new power, the ethics of data, new strategies of public action, the economic role of public data, national sovereignty. Such questions will be posed worldwide. The digital revolution, and at its forefront, the data revolution, is a new industrial revolution. It redefines the economic and social balance and requires intense work to reach the creative synthesis needed to bring about new equilibriums.

In this revolution, France will have to find its own way. It will do so with even more clarity, and all the more force, knowing that it has its destiny under control, that it has chosen to master its tools and practices and speaking from experience.

GLOSSARY

A/B testing : A/B is a technique for testing two different versions (A and B) of a message or an interface to determine which is the most effective from the viewpoint of the recipient of a message or user.

CDO : France's Chief Data Officer, the function was created by Decree of the Prime Minister on 16 September 2014. The CDO coordinates action by the administrations on the inventory, governance, production, movement and use of data by administrations.

Anonymisation : anonymization of data is a technique aimed at changing its structure in order to make very difficult or impossible to re-identify natural or legal persons or entities concerned (source : Wikipedia).

API : Applications Programming Interface allow software to provide services to other software in a simple way. For example : the geocoding API located at data.gouv.fr can transform a postal address in geographical coordinates (latitude, longitude).

Big data : Big data refers to both data with certain characteristics — rich and voluminous — but also, by extension, the use made thereof.

Data : Digital data is the basic description of a digital nature, represented in coded form, of a reality (measure, transaction, event, etc.).

Reference data : Reference data are data frequently used by multiple public and private actors, and whose quality and availability are critical for these uses, e.g. geographical data in state repositories.

Pivot data : a key-data links several sets of data, such as for example the SIRET number of a business.

Data governance : set of principles and practices that are designed to ensure a better exploitation of the potential of data.

Register : In administration, a register is a book in which administrative information is entered.

Machine learning : From artificial intelligence, machine learning is a set of techniques where algorithms can "learn". I.e. they can improve themselves by including new data.

BIBLIOGRAPHY

Anderson C. (2008) : The end of theory, the data deluge makes the scientific method obsolete, Wired Magazine

Andreessen M. (2011) : Why software is eating the world, The Wall Street

Bracken M. (2013) : On Strategy : The strategy is delivery. Again., Available at mickebracken.com

Chief Data Officer (2015) : Analyse the energy consumption of public buildings, available on agd.data.gouv.fr

General Council of Economy, Industry, Energy and Telecommunications (2015) : Best practices for the big data and analytics in administration: A further step, report to the Minister of the economy, industry and the digital sector, Secretary of State responsible for State reform and simplification and the Secretary of State with responsibility for the digital sector

National Board of Housing, Etalab (2015) : open data in the field of housing, summary of discussions

Council public publishing guidance and administrative information (2013) : Guide to the protection of personal information in the context of open data

Davenport T., Patil DJ (2012) : Data Scientist, the sexiest job of the 21st century, Harvard Business Review

De Soto H. (2005) : The Mystery of Capital: "The Mystery of Capital: Why Capitalism Triumphs in the West and Fails Everywhere Else", Flammarion

De Vries M. (2012) : Re-use of public sector information, report for Danish Ministry for Housing, Urban and Rural Affairs

Desrosières A. (2000) : La Politique des grands nombres: histoire de la raison statistique, Editions La Découverte (2nd Edition)

Flowers M. (2013) : NYC by the numbers, annual report to the Mayor of New York

Fouilleron A. (2015) : The data exchange for consideration between administrations, report to the Prime Minister

Grossman N. (2015) : White Paper: Regulation, the Internet Way. A Data-First Model for Establishing Trust, Safety, and Security | Regulatory Reform for the 21st Century, Mimeo

Jetzek T., Avital, M. (2013) : The Generative Mechanisms Of Open Government Data, ECIS 2013 Proceedings

Ministry of Finance of Denmark (2012) : Good basic data for everyone — a driver for growth and efficiency

Morin-Desailly C. (2013) : The European Union, colony of the digital world? Report drawn up on behalf of the Committee on European Affairs of the Senate

Press G. (2013) : A very short history of data sciences, Forbes.com

Trojette A. (2013) : Open data: The exceptions to the principle of free access are all legitimate? Report to the Prime Minister

Vickery, G. (2010) : Review of recent studies on PSI re-uses and related market developments

Volle M. (2006) : IT: life with automats, Economica

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE

Premier ministre

SECRÉTARIAT GÉNÉRAL POUR LA MODERNISATION DE L'ACTION PUBLIQUE