CHIEF DATA OFFICER

DATA AS AN ESSENTIAL INFRASTRUCTURE

REPORT
TO THE PRIME MINISTER
ON THE MATTER OF DATA
IN THE ADMINISTRATION
2016-2017

La documentation Française

CHIEF DATA OFFICER

DATA AS AN ESSENTIAL INFRASTRUCTURE

La **documentation** Française

Foreword

"Throughout history, the greatest disruptions have occurred when what was once a scarce resource became an abundance. The transition from hunter gathering to an agrarian lifestyle; the invention of the printing press; the rise of new manufacturing processes and the industrial revolution; and the World Wide Web giving rise to our networked world. In each case what was a scarce resource became an abundance: more food, more literacy, more mass produced products, more data. These 'revolutions' of abundance have invariably seen the emergence of new structures in society, new forms of governance, new sources of wealth, new opportunities and, we should note, new inequalities. This has happened in the past and it is happening now in the data revolution"

These were the opening lines of the report[1] commissioned two years ago by Emmanuel Macron then Minister of Economy and Finance, and The Chancellor of the Exchequer, to Sir Nigel Shadbolt, co-president of the Open Data Institute and the Chief Data Officer.

In order to prepare the French administration to this revolution allowing it to use its data to create new services and better manage its public policies, a Chief Data Officer (CDO) position was created in September 2014. This mission requires both the capacity to locate and use the data, as well as the mastering of data science's tools to apply these methods with the ambition of improving the public service.

Inspired by the emergence of Chief Data Officers in large American cities and private companies, this position was then implemented for the first time at a governmental level.

The modest team (4 persons) was to face a huge challenge: bringing the State into the age of data. In other words: creating the conditions in which the State gains control over its data, puts it to good use, shares it – whilst respecting legal secrecy – in order to attribute to each its maximal potential value and more importantly, learn to use it to conceive and pilot public policies.

In 2014 and 2015, the CDO supported several ministerial projects, proving the value of data science in helping administrations in solving pinpointed issues.

1 "Data Driven Growth", a report by the Anglo-French working froup on the data economy (2016) co-chaired by the French Chief Data Officer, Henri Verdier and Sir Nigel Shadbolt, co-founder of the Open Data Institute. It can be downloaded from economie.gouv.fr: https://bit.ly/2MHpVfY

Various examples are presented in this report. The first few years also revealed the complexity of data governance, and the internal barriers the State must overcome, whether they stem from a rigid application of the legal secrecy, computer silos, or the weakness of the culture of cooperation among administrations. Whilst identifying the key aspects of the data policy, the 2015 CDO report[2] covered these issues extensively.

As of 2016, the Chief Data Officer—henceforth part of the DINSIC, along with the Etalab mission—was able to follow through an inter-ministerial dynamic which resulted in, the appointment of a number of data administrators who have undertaken a considerable amount of work within their respective administrations; the co-operations with a number of governmental start-ups; as well as the creation of the "Public Interest Entrepreneurs" ("Entrepreneurs d'intérêt general") program, which allowed a certain number of administrations to host data scientists.

As a result, the CDO could gradually devote his resources to new challenges, and particularly the one posed by the issue of data flow, whether by contributing to the Digital Republic Act, or bringing the issue of data to the heart of the "State as a platform" strategy by creating the highly structuring and popular APIs, such as the businesses API[3] – which shares more than 15 million pieces of information per year – or the Geocoding API[4] which represents 1.6 billion addresses per year.

Thanks to the newly announced Digital Republic Act, and because the CDO believed data was an essential infrastructure of the economy and the State's functioning, he was entrusted with creating a "public data service" collecting all the reference data. Such an infrastructure could prove to be very powerful. Indeed, as the success of companies such as Uber and Airbnb has taught us, being able to control the data infrastructure provides the benefits of a physical infrastructure without the need to neither build nor actually own it.

This report explores extensively the issue of data infrastructures, as it goes partly beyond that of open data and calls for larger investments and a much superior operational structure.

In recent years, the subject of data has gradually become a key issue. Greater social vigilance and demands to safeguard privacy have led to renewed European regulations regarding the protection of personal data (2018). At the same time, awareness regarding the need to establish a governance framing the use of algorithms and ensuring that their capabilities meet the requirements of modern day democracy, has emerged. The CDO has been actively examining these issues and some initial responses have been provided by the Digital Republic Act.

2018 will probably mark the start of a new phase. Indeed, the explosion of the number of available data and the rapid increase of computing power has breathed

2 "Leverage data to modernise public action", 2015 Report of the Chief Data Officer to the Prime Minister, available on gouvernement.fr: https://bit.ly/33hpwa1
3 Entreprise.api.gouv.fr
4 Adresse.data.gouv.fr/api

a new life into Artificial Intelligence, a key scientific discipline. Albeit in a form commonly referred to as "weak AI", it is already widely present in our daily lives.

In addition to the industrial competition that will determine which very few countries will possess the actual capacity to develop artificial intelligence resources, a cultural battle centred on the issue of data is already rearing its head. The matter is actually quite simple: which data, thus, which cultural model, will be used to educate the artificial intelligences destined to play a critical economic and social role.

The government's commitment, strongly promoted by the President following the findings of Cedric Villani's report on March 29, 2018, calls for the increase of the State's capacities to harness and use data in the service of public action.

The current revolution will continue, and requires a deep evolution of the public service if it is to succeed in its missions in a cost-effective manner, whilst meeting the highest standards and the users' growing expectations. Most of all, for the State to master these new capabilities and in keeping with the foundation of our Republic: Democracy, the People's government, by the people, for the people.

Henri Verdier
Chief Data Officer

CONTENTS

Introduction

Since the post of Chief Data Officer was created in September 2014, the French Government has become aware of the importance of the data revolution and has conceived a **data policy** focused on three main objectives: to provide high-quality data, particularly through the public data service; to enable **the flow** of data by applying the "access by default" principle to all communicable data and through the development of APIs that promote the exchange of data between administrations and with the civil society; and lastly **to exploit** the data in order to improve the efficiency of public initiatives.

To implement this strategy, the Government relies on several actors: its Chief Data Officer, position filled today by Henri Verdier, accompanied by his team within the Etalab mission, but more generally on the State's Inter-ministerial Direction for Digital Information and Communication Systems, and the network of ministerial data administrators who have gradually been appointed in various ministries, as well as the "Public Interest Entrepreneurs" innovation policy, launched in 2016.

Data is currently **at the centre** of both **public action** and economic activity and it must be seen as an **essential infrastructure** to the functioning of the economy, just as the transport and telecommunications networks. The Government must be the catalyst, encouraging the rest of society.

The Chief Data Officer

Established under Decree No. 2014-1050 of September 16, 2014, the Chief Data Officer (CDO) reports to the Prime Minister and is attached to the State Inter-ministerial Director for Digital Information and Communication Systems.

The CDO coordinates the activities of administrations with regard to inventories, governance, production, circulation and exploitation of data.

In compliance with the imperatives of legal secrecy and the protection of personal information, he arranges for the best use and widest circulation of data, aiming to evaluate public policies and their improvement, to ensure the transparency of public actions as well as to foster research and innovation.

Hence, he encourages and supports the development and use of data science within the administration.

A duty accomplished by several actors

The position of Chief Data Officer is currently held by the State Inter-ministerial Director for Digital Information and Communication Systems. The operational team is part of the Etalab mission, and works closely with the other members of DINSIC and within a network of other administrations.

Thanks to this new organisation, the CDO can implement an approach that combines:

Operational measures and initiatives:

– to open-up public data;
– to develop APIs;
– to advise the Government on the creation of a public data service;
– to analyse the data processing of major projects which require approval from the DINSIC, under Article 3 of Decree No. 2014-879 of August 1, 2014 [1] concerning State information and communication systems;
– as well as its original tasks: to develop in-house data science projects, responding to referrals brought to the CDO, and lending support to the main ministerial administrators.

Opinions and reviews:

– referrals to the Commission in charge of access to administrative documents (Commission d'accès aux documents administratifs CADA);
– audit of the Register of Secure Electronic Titles (SETs) at the request of the Minister of the Interior in cooperation with the French National Cybersecurity Agency (Agence nationale de la sécurité des systèmes d'information, ANSSI);
– Franco-British Mission on the subject of data driven growth, commissioned by the Minister for Economy and Finance, The Secretary of State for Digital Affairs, The Chancellor of the Exchequer and the Minister for Culture, Communications and Creative Industries.

Assistance, supervision and support:

– support to the development of the National Health Data Services;
– supervision of the "Public Interest Entrepreneurs" program;
– assistance in the development of new approaches of network security by the Inter-ministerial Network of the State (RIE) and the National Cybersecurity Agency (ANSSI).
– assistance to Ministries for the organization of ministerial hackathons and their exploitation.
– support to some governmental start-ups.

---

1 See article 3 of Decree No. 2014-879 on legifrance.gouv.fr: https://bit.ly/2GBohJY

In terms of public policy, the existence of multiple levers of action is of great importance. Indeed, the real impact of data science can only be achieved with a continuum from the creation of real data infrastructures to the total transformation of jobs.

The Chief Data Officer's report to the Prime Minister

In accordance with the Decree that created the Chief Data Officer position, he (the CDO) "delivers an annual public report to the Prime Minister on the matters of inventory, governance, production, circulation and use of data by the administrations. This report provides a review of existing data, its quality and of the innovative usages it brings. This report also presents recent developments in the data economy. It holds recommendations for the improvement of the exploitation and exchange of data between administrations".

Thus, it has three distinct objectives: **to provide an overview** of the administrations' practices with regard to data, **to project future developments** in the data economy and to recommend improvements so as to exploit its full potential.

These key aspects are addressed in the three sections of this report:

– the first section provides and explains the outlines of the **data policy.** It recounts the actions undertaken by the administrations and the levers used by the Chief Data Officer since the publication of the first report;

– the second part offers an analysis of a key subject: data as an infrastructure. It recommends the creation of a **real data infrastructure** in which the State has the ability, and the obligation, to have a central role;

– the third section suggests ways **to turn this into a winning strategy** and strengthen governmental measures with regard to data in 2018.

This report, presented under the sole responsibility of the Chief Data Officer, is the result of a collaborative effort. It gathers many contributions of the DINSIC's teams – including the Etalab mission and particularly from de Data scientist Paul-Antoine Chevalier – and their ministerial partners. It owes much to Etalab's strategic consultant Simon Chignard's writing skills and strong commitment.

Part One

Data policy: producing, sharing and exploiting the data

Why should we concern ourselves today with the data produced and exploited by the administrations? The Chief Data Officer's first report has already pointed out the implications of the unrestricted use of data with regard to the current evolutions. The trends that were identified then have developed.

The first trend concerns the shift from a data economy based on scarcity, toward one of **abundance**[1]. Nowadays, data is produced and exploited more easily and for a lesser cost. More than before, it is collected in a systematic manner by the information systems and sensors. The State's data is sometimes **in competition** with that of others, produced following a different process.

This transformation has major consequences on the way the administrations, and even society as a whole, can create economic and social value based on data.

Until recently, the situation was quite paradoxical. On the one hand, larger producers of data such as operators and the central administration services, monetized their data (through license fees), to a handful of economic actors[2]. On the other, the exploitation of the vast majority of the data was reserved to the administration which produced or collected it, which meant a loss of opportunity.

Indeed, by selling the data it should have been giving away, and under-exploiting it when it came to its own operations, the State proved to be a poor manager of the asset. It is because data flows widely when no legal limit opposes it (legal secrecy…), that it can be used to create value, improve public services, kindle new products and services and provide the economic actors with the crucial information they require for their activity.

Hence, a data policy is the answer to the challenges and opportunities of the data revolution for the State and the administrations. It can be summed up in three principles: producing key data, sharing it and promoting its exploitation.

1. Producing key data

The Primary role of the State is the **production of the data critical** to the proper functioning of administrations and the entire economy.

(Fortunately) Public authorities did not wait for the advent of Big Data to start addressing the issue of data. Indeed, the State has long been producing **the registers** required for its operations, such as naming or identifying a

---

1 The consequences of this shift from a situation of scarcity to a situation of abundance is the basis of the Franco-British report on "Data Driven Growth", commissioned in November 2015 by the Minister of the Economy, Industry and the Digital Sector, Emmanuel Macron, and the Chancellor of the Exchequer, Georges Osborne, from Henri Verdier, Chief Data Officer, and Nigel Shadbolt, Dean of Jesus College, Oxford and Co-chairman of the Open Data Institute. The full report is available at the following address: https://bit.ly/2MHpVfY

2 Along the lines of Google's 2012 acquisition of the IGN databases. This operation, which cost several million euros, has not been repeated since Google developed its capacity to update its own data.

location, a business or an individual. Indeed, the administrations manage a wide range of databases, vital to the proper management of the country. The National Register for the Identification of Natural Persons (Répertoire national d'identification des personnes physiques) assigns a unique number to each individual (their registration number, commonly known as their "social security number"). In the economic sphere, the database for the identification of enterprises and establishments (SIRENE database, produced by INSEE) also plays an essential role in the organization of a dialogue, not only with the administrations but also between businesses themselves.

Preserving the sovereignty of information

The State recognizes the **new aspects of sovereignty** with regard to digital technology and to the stakes associated with the competition for the production and the use of the largest registers.

Nowadays and unlike 20 years ago before the widespread use of the Internet and communication networks, several public and private registers are in competition. On this matter, one example is Bloomberg. This company possesses its own system of identification of economic actors, and also references French businesses.

However, in an open digital economy, the notion of ***de facto* standard** prevails. In the field of data, this means that references are no longer proclaimed unilaterally by an actor, rather, they are simply recognized as such by users.

The *de facto* Standard created by a foreign private actor such as Bloomberg can *de facto* compete with a state-owned register such as SIRENE, the database created by the INSEE. In order to remain a standard, the state-produced data must be easily accessible and widely distributed.

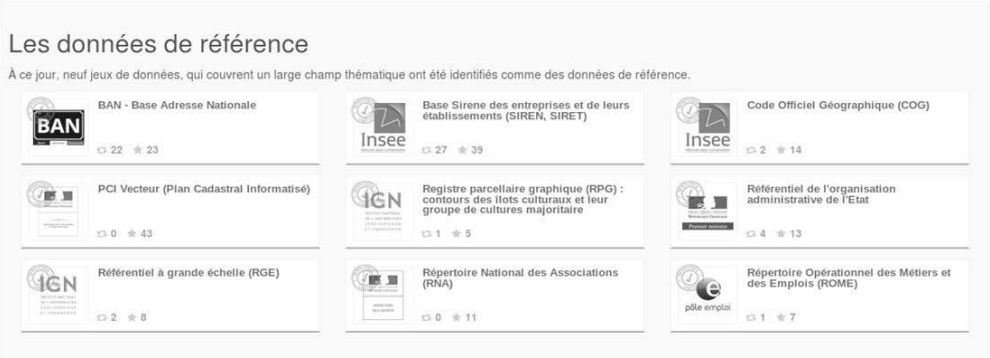
Identifying and giving legal recognition to reference data

It is no secret that not all openly available data has the same value and that some have **more potential**. However, until very recently, the legislation had not assigned a specific status to this type of data. This has now been accomplished thanks to the recognition of the concept of **reference data** by the Digital Republic Act [3].

Reference data, as defined in the French Code of Relations between the Public and the Administration (code des relations entre le public et l'administration or CRPA) must meet three criteria [4]:
i. It is used to identify or name products, services, locations or individuals;
ii. It must be frequently used by public or private actors, outside the administration or organization that owns it;
iii. The quality of its delivery is critical to its exploitation.

---

3 Digital Republic Act 2016-1321 of 7 October 2016: https://bit.ly/2p9KyUk
4 See CRPA Article L-321-4 and the following: https://bit.ly/2YwgsPy



The first criterion (i) relates to the concept of **key data**: the SIRET number, which serves to uniquely identify the establishment of a business or an organization is a case in point. This number serves as a link between several databases: for example, it links the SIRENE database to the health facilities database (Base des établissements de santé or FINESS) and to fiscal and social data.

The second criterion (ii) emphasizes the data's **reusability value**. Reference data is, as the name suggests, data that *refers*.

The third and last criterion (iii) emphasizes the compulsory **high standards** with regard to its availability.

The 9 sources of reference data

	Producer	Domain
The Register of Enterprises and establishments (SIRENE)	The National Institute of Statistics and Economic Studies (INSEE)	Economy
The National Directory of Associations	Ministry of Interior	Associations
The State and Public Services' Administrative Organisation Database	Directorate of Legal and Administrative Information (DILA) (Prime Minister)	Administration
The Operational Repertory of Professions and Employment (ROME)	Pôle Emploi	Economy – employment
The Digitized Cadastral Map	The Directorate General of Public Finances (Ministry of Economy and Finance)	Geographic – property
Official Geographic code	National Institute of Statistics and Economic Studies Geography – territorial organisation	Geographic – territorial organisation
The Rural Land Register	Services and Payments Agency (Ministry of Agriculture)	Geography – agriculture
Large Scale Repository	The National Institute of Geographic and Forestry Information	Geography
The National Address Database	IGN, La Poste, OSM France, Etalab	Geography

Of the nine sources of reference data identified at this stage[5], five constitute a coherent corpus of geographic data, and four allow for the identification of businesses, associations, administrations, professions and jobs. As of now, these databases are accessible via a dedicated space on the data.gouv.fr platform[6].

The Register of Enterprises and Establishments (Répertoire des entreprises et de leurs établissements) is published by the National Institute of Statistics and Economic Studies (INSEE). In accordance with the Digital Republic Act (October 2017,), its digitized version, the SIRENE database, is accessible and free in open data[7] since January 4, 2017. It provides information on all French establishments, regardless of their legal form or field of activity (industrial, trade, crafts, liberal profession, agricultural, territorial community, banking, insurance, non-profit…). To date, SIRENE registers more than 10 million establishments, and its services register around 10 000 modifications daily. The updates are published every day.

The National Directory of Associations (Répertoire national des associations, RNA) is constituted by the Ministry of Interior. It aims to identify with a unique number (RNA number, formerly, the Waldec number) every association registered in a prefecture. For each association, the RNA provides the name, purpose, headquarters and offices' address, duration, corporate object code as well as its legal nature. This database currently lists 1.5 million associations.

The State and Public Services' Administrative Organisation Database (base de l'organisation administrative de l'État et des services publics) is run by the Directorate of legal and administrative information (DILA, services of the Prime Minister). This register is the reference for identifying and contacting the organisations of the central administration.

The Official Geographic Code (Code officiel géographique, COG) regroups the official nomenclature of administrative districts, city districts (arrondissements), municipalities, departments and regions, as well as their names (libellés) and codes. The data is produced by the INSEE since 2003. In order to introduce modifications such as the merger of regional authorities (collectivités), the register is updated yearly. It may be considered as the original index of many geographic data.

The Digitized Cadastral plan (Plan cadastral informatisé) is produced by the General Directorate of Public Finances (Direction générale des finances publiques, DGFIP) of the Ministry of Economy and Finance. The land plots are identified with the unique INSEE code pertaining to the municipality, its section and plot numbers. The French online cadastral plan gives access to approximately 600,000 plan sheets both in image and vector formats.

---

5 Article R. 321-5 of the Code of Relations between the Public and the Administration: https://bit.ly/2YPsuTt

6 https://www.data.gouv.fr/fr/reference

7 See article L. 324-6 of the Code of Relations between the Public and the Administration: https://bit.ly/2OLwhxi.

The Rural land Register (Registre parcellaire graphique RPG), produced by the Services and Payments agency (Agence de services et de paiement, ASP), is a reference database with regard to agricultural land use. It contains 7 million graphical objects.

The Large Scale Repository (Référentiel à grande échelle, RGE), is co-produced by the National Institute for Geographic and Forestry Information (Institut national de l'information géographique et forestière), the La Poste Group, the Openstreetmap France organisation and the Etalab mission (DINSIC). This database classifies all addresses referenced on the French territory, and currently lists the geographical location of over 25 million addresses.

The Operational Repertory of professions and employment (Répertoire opérationnel des métiers et des emplois, ROME), produced by Pôle emploi (the employment agency) is the reference for matters relating to employment and skills evolution. For example, the Bob Emploi service exploits it to calculate similarities between jobs and suggest different professional paths to job seekers.

With the appointment of the Ministerial Chief Data Administrators in the upcoming months, the list of reference data is destined to grow.

Opening up to new ways to collaborate and produce data

The State cannot ignore the evolution in the methods of data production. Indeed, just as the INSEE analyses the receipts from cash registers in order to measure the inflation rate, **data produced by the private sector** can be useful to the public authorities.

Therefore, the Digital Republic Act has introduced the concept of **general interest data**. It pertains to all the data pertinent to the public that provides a general social benefit, even beyond a direct link to a public service initiative. Besides the field of government subsidies and public service delegations, this notion could secure sector-by-sector access to data which would allow consumers to be better informed and able to make their own decisions, and users to benefit from a smoother service (data relating to transport, energy and mobile phone coverage in particular). There is however no doubt that this idea could have a wider application. As a matter of fact, this topic is reviewed in the Public Sector Information Directive, initiated recently by the European Commission.

The digital commons also offer another model of production and data governance, which focuses on collaboration and sharing.

A good example is the National Address Database, introduced in 2015 by the IGN, La Poste, OpenStreetMap France and Etalab, with support from the Chief Data Office. This database is unusual and unique for, not only is it the most exhaustive database of French addresses to date, but also because its governance gathers administrations, public companies as well as members of an association.

Often cited as an example abroad, this model could be reproduced in other sectors such as the health sector, town and country planning, etc. One can also imagine a way in which data of critical importance to several stakeholders is co-maintained or co-governed, when none of them possesses the necessary resources to claim its individual ownership.

Defining new standards for data

The importance of *de facto* standards does not however mean that the State should no longer recommend new ones, especially when standards enable a public policy priority to come to fruition. However, their success, and level of appropriation lies in the Government's capacity to create and rally a solid ecosystem, surrounding it[8].

Thus, not only does the State participate in the conception of general principles ("the transparency of public procurement"), but also in the creation of the data standards required to put them into practice. Some work intend to define norms and format obligations in order to be able to compare, and aggregate data produced by a wide variety of actors. Some examples include data taken from the charging infrastructure for electric vehicles, essential data from public orders as well as data relating to subsidy agreements.

Two European directives, transposed by executive orders in 2016, assert the obligation to give open access to the public order's key data. The texts refer to the common wording "open format and unrestricted usage[9]". Since October 1, 2018, the public buyers are compelled to publish the data pertaining to their purchases exceeding 25 000 euros on the public markets as well as to concessions.

It seemed necessary to standardize the public orders' data, in order to promote the emerging ecosystem surrounding it. An in-depth reflection was devoted to the matter of format, and a referential, set by a decree on April 14, 2017[10] was eventually elaborated following a state-run experimentation involving some pilot administrations. The same standardization process was carried out regarding **data relating to subsidies**[11].

At the same time, and alongside five other countries within the **Open Government Partnership** framework, France made a commitment to the **Open Contracting Partnership**. The latter aims to develop and promote

---

8 This principle also applies to standards originating from the private sector as demonstrated by the case of transport data. Google's active contribution to the creation of the *General Transit Feed Specification* (GTFS) exchange standard has not only helped facilitate the reuse of transport data, but has also guaranteed Google an ideal position to incorporate public transport data into its services (more specifically Google Maps).

9 See articles 107 of Decree No. 2016-360, 34 of Decree No. 2016-86 and 94 of Decree No. 2016-361.

10 http://www.data.gouv.fr/fr/datasets/referentiel-de-donnees-marches-publics/ and https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000034492587&categorieLien=id

11 Decree No. 2017-779 of May 5, 2017 regarding electronic access to key subsidy agreement data: https://www.legifrance.gouv.fr/eli/decret/2017/5/5/PRMJ1636989D/ jo/texte and the order of November 17, 2017 regarding the conditions under which key subsidy agreement data is made available: https://www.legifrance.gouv.fr/eli/arrete/2017/11/17/ PRMJ1713918A/jo/texte

a standard format of open data relating to public tenders (the Open Contracting Data Standard), in order to work jointly to achieve the opening up and the provision of data pertaining to public orders, which will consequently create an international standard. On November 28, 2017, **France took over the presidency** of this organization in order to promote and develop the concrete use of data relating to public orders[12].

Improving the coordination in data production could be a factor of cost saving and efficiency

There are still opportunities for cost-saving and efficiency in the production of reference data. Currently, the production of reference data relies on a few operators and governmental departments. Their know-how and expertise in this field are considerable, but progress nonetheless remains possible with regard to the governance of this data's production.

Indeed, most of these producers act fairly independently, including from ministerial supervision. As a result, and despite being connected[13], the main registers are produced without necessarily sharing a **common strategic framework**. Thus, improving the governance of production would be a factor of cost saving and efficiency.

For instance, the introduction of a unique identification number for associations has long been presented as an element of simplification valuable to the voluntary sector[14].For now, associations that request subsidies, pay taxes or employ staff, must often present two identification numbers: the first, in the SIRENE database and the second in the National Directory of Associations.

2. Improving public data flow

Data flow must be promoted, whilst respecting legal secrecy and people's privacy. Circulation must become the rule, and its restriction must be a justified exception.

The actions undertaken since the creation of the Chief Data Officer task relate to two complementary objectives:

– to adapt and allow an evolution of the legal framework, so as to minimize the technical, economical and legal obstacles to the circulation of data;
– to design and operate the tools and devices (platforms, API, etc.) to facilitate the flow of data, in coherence with the "State as a platform" approach.

---

12 https://www.open-contracting.org/2016/12/07/open-contracting-version-francaise/#eng
13 The address component is one example of the links between databases. It is used by multiple reference databases (including the SIRENE database and the National Directory of Associations).
14 See the report submitted to the Prime Minister by member of Parliament Yves Blein: *50 mesures de simplification pour les associations* (*50 Simplification Measures for Associations*), October 2014.

The opening up of data: a significant progress

Since the publication of the Chief Data Officer's first report1, several extensive key databases have been published in various domains. Although the goal of Open Data by default, mentioned in the Digital Republic Act, has not yet been achieved, there has been significant progress in numerous fields.

In the health sector, the National Health Insurance Fund (Caisse Nationale d'Assurance Maladie) has continued its efforts to publish its data. It began in 2015, with the publication of the Cross-regime Health Insurance Expenditure Database (DAMIR. The CNAM has released the database for hospital medical prescriptions issued in the city (June 2017) and the database for cross-regime medical biology (March 2017).

In the economic sector, the opening up of the SIRENE Directory in January 2017 constituted a major advancement, and resulted in an important volume of data reuse. Regarding the field of geographic data, the publication of the Digitized Cadastral Plan (September 2017) is a good example of the release of a key dataset.

In the Housing sector, the release (December 2017) of the Directory of Public Housing (repertoire des logements locatifs des bailleurs sociaux) containing detailed data of about 4.9 million housing units, is another example of the publication of high value-added granular data.

Adapting and developing the legal framework

Since the publication of the first report of the Chief Data Officer, the Valter law (regarding the free access to public data) and the Digital Republic Act have deeply transformed the legal and regulatory frameworks, with the aim of facilitating the circulation of data.

From reactive disclosure to open by default

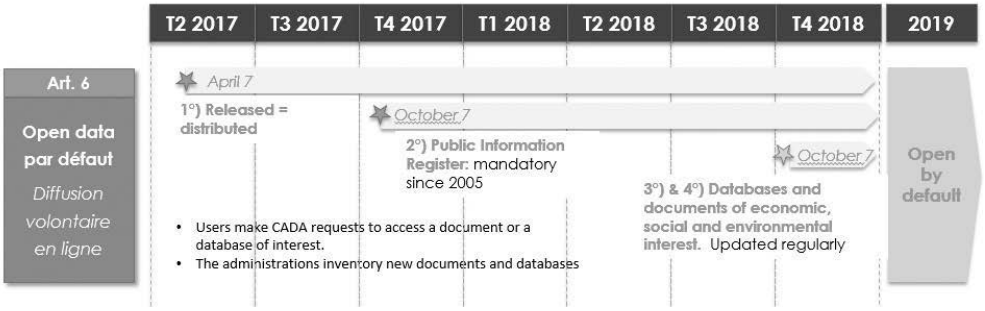
The shift from a reactive disclosure to open by default prompted by the Digital Republic Act has greatly expanded the scope of administrative documents available online.

In fact, according to the law, administrations employing more than 50 agents (except municipalities of less than 3500 inhabitants) are henceforth compelled to publish the following information in the form of open and easy to reuse standard[15]:

– documents released following a request (April 7, 2007);

15 For open data by default, see article L. 12-1-1 of the Code of Relations between the Public and the Administrations (CRPA), and for the dissemination standard, article L. 300-4 CRPA.

– documents appearing in the public information directories (October 7, 2017);
– the databases updated on a regular basis (October 7, 2018);
– the regularly updated data presenting an economic, social, sanitary or environmental interest (October 7, 2018).



In this context, an administration can henceforth provide online access to a document except when the matter is private and must be restricted to a personal access only[16].

Towards a unification of the legal framework for all "administrations"

The second major development involves the unification of the legal framework for all administrations. Indeed, the concept of "administration" (with respect to the French Code of Relations between the Public and the Administration) is particularly broad. It covers the scope of all administrations (public administrative service, as well as industrial and commercial), but also every legal entity under private law, in charge with a public service mission, and all data produced or collected within this context. Some examples include the operators of the transport network which provide a public service delegated to them by an administration.

Whenever public data is concerned, all these administrations, **without exception**, are bound by the obligation to distribute and provide access to it. They are also subject to the same rule regarding the reuse of the data since the **end of the exception** regarding the data of the industrial public and commercial services, as well as that of the derogation extended to cultural services which allowed them to freely define the conditions for the reuse of their data. Let's note that the reuse is not restricted, and that exploitation for commercial purposes is permitted.

Moreover, the legislation eliminates the possibility for the administrations to lay claim of intellectual propriety in an attempt to block the free reuse

16 See article L. 311-9, CRPA.

of their databases (the *sui generis* right of the database producer), with the exception of those created within the frame of an industrial or commercial public service mission open to competition.

In conclusion, regardless of its category, every administration is subjected to the same rules regarding the accessibility, provision and exploitation of its data, with the notable exception of the EPICs, a category of French public undertakings, (établissement privés à caractère industriel et commercial), which are open to competition.

The existence of a single framework brings **a much welcomed clarification** for the users who reuse the data, as well as for the producing administration themselves.

Transitioning from licence fees to free and open public data

Free access to public data has now become the rule. Exceptions to this principle are strictly regulated and must be substantiated [17].

Only two categories of administrations are still allowed to charge for the reuse of their data:

– those mainly engaged in collecting, producing and providing or publishing public information, and whose costs are covered up to 75% by the fiscal profit, subsidies or allocations. This concerns the Hydrographic and Oceanographic Service (SHOM), Météo France and the IGN.
– cultural institutions, regarding data derived solely from operations of digitization.

On top of that, **the charges are capped** and must match the costs of production and provision of the data to the public.

Also, the law has declared all the data produced by the INSEE and ministerial statistics services free of charge.

This follows the framework of recommendations found in the Trojette Report. This document put into perspective a slight decrease in incomes, of which a significant portion originates from public actors (around 15% of the total fees perceived), as well as an excessively high marketing organization cost with respects to the generated value.

Free access removes a significant economic hurdle to the issue of data reuse. Indeed, because of an insufficient budget, some administrations decided not to acquire data that was nonetheless vital for the accomplishment of their duty. In a way, it is also a process **simplification** since it puts an end to the need to sign a licence, negotiate the fees and invest in the commercialization.

17 The "Valter" law, which implements European Parliament and Council directive 2013/37/EU of 26 June 2013; see article L. 324-1 et seq. CRPA.

A licensing policy to improve data flow

The Digital Republic Act stipulates that in the case a free reuse results in the establishment of a licence, the latter must be **chosen from a specific list**, fixed by decree, updated every five years[18] and available via this link www.data.gouv.fr/fr/licences. This aims to avoid the multiplication of licences, guarantee the unimpeded flow of data (open data), and particularly allow cross-referencing.

If an administration wishes to issue an unlisted licence, it must first be approved by the State, under conditions fixed by decree.

Additionally, the evolution of the legislative corpus has been the occasion to introduce a new version of the "Open Licence". By authorizing the reproduction, redistribution, adaptation and commercial exploitation of data, it offers the greatest freedom with respect to data reuse. This new version is adapted to the international framework, since it is compatible with other open data licences standards developed abroad, such as the British government (Open Government Licence), among others (ODC-BY, CC-BY 3.0). The only requirement is the mention of the authorship of the data, as well as the potential presence of personal data. It also ensures that intellectual property rights will not put a strain on the freedom to reuse the data.

Towards an increase in data flow between administrations

A twofold approach is applied to pay particular attention to the flow of data between administrations:

- **the pooling** and reuse of data particularly where major databases are concerned, aiming to enhance the efficiency of public policy;
- **the simplification** of administrative procedures for the benefit of the users, following the "Dites-le-nous une fois" principle ("once only principle") .

Article 1 of the Digital Republic Act provides the administrations with a right to access and publish any public data possessed by another administration, with the exception of the secrets protected by law, necessary for the accomplishment its public service mission. This information can be reused by any administration for accomplishing a public service purpose, other than the one for which the data was produced or received. In a way, this article aims to **extend to the administrations the rights** already enjoyed by the citizens.

Moreover, since January 1, 2017 the sharing of public information within this frame, between multiple administrations of the State, or between the latter and public administrative establishments and amongst public administrative establishments themselves is free of charge (with the exception of the aforementioned situation). This is in continuity with the recommendations found in the Fouilleron report and presented to the Prime Minister in December

18 See article. D. 323-2-1 CRPA.

2015. This report put into perspective the perverse effects caused by the billing of data sharing between administrations. It also underlined the fact that, half of the transactions involving data cost less than €500 per unit, however it is easy to imagine the costs involved in the organization of such a transaction. In this sense, the sale of data between administrations was not a zero-sum game. It undermined the efficiency and quality of public services by generating transaction costs and negative effects such as the avoidance of data usage due to budgetary constraints or the development of circumventing strategies by the purchasing administrations.

At the same time, in an effort to simplify procedures by avoiding the redundancy of requests for documentary evidence, the administrations can exchange information pertaining to their users' administrative procedures [19].

Controlling open data and data flow

The renewed missions of the Commission on Access to Administrative Documents (Commission d'accès aux documents administratifs, CADA).

Thanks to the new legal framework, the Commission is henceforth competent to examine requests for counsel or guidance on the matter of online publication of administrative documents and beyond the issues of release and reuse, including the matter of licenses and potential fees. In its latest annual report, it also asserts its crucial role in the "regulation of open data".

Referrals to the Chief Data Officer

With the objective of improving the data flow, the decree establishing the position of Chief Data Officer has also established a right for individuals to seize its authority. Local authorities, legal persons of public law as well as legal person of private law conducting a public service mission can refer to the CDO on matters pertaining to data usage by their respective organizations [20].

The system for submitting a request to the CDO completes the procedure of the CADA. After having identified the document he wants to access, the citizen should first submit a request to access it to the relevant administration. As a second step, and if needed, a request can be made to the CADA. Support from the CDO can intervene in case technical difficulties impede the provision of the data.

19 See articles L. 114-8 to 10 and L. 113-12 and 13 of the CRPA.
20 The form required for submitting a formal request to the AGD is available on the AGD's blog: https://agd.data.gouv.fr/ saisines-de-lagd/formulaire-de-saisine/

Various sources of referral:

The Chief Data Officer can be seized by administrations, companies, and any individual, regarding any subject pertaining to the circulation and/or the use of public data. The diversity of the sources is illustrated by the following requests submitted in 2017:

– The Commission of Access to Administrative Documents has asked the CDO to convey an opinion regarding a demand to access a real-time data stream. The CDO provided technical advice on the conditions needed in order to consider this kind of stream as published "in an open standard which is easily usable and reusable by an automated processing system", in respect to the law.

– A private business has submitted a request to the CDO regarding the availability of information relating to the state of the real estate market. The CDO was able to immediately provide all the data sources and the legal framework permitting its accessibility to the public, all the while taking into consideration the future legal evolutions.

– A citizen requested access to case law documents. The CDO undertook a project to anonymize the information, together with the National Commission on Informatics and Liberty (CNIL) and the Justice Ministry.

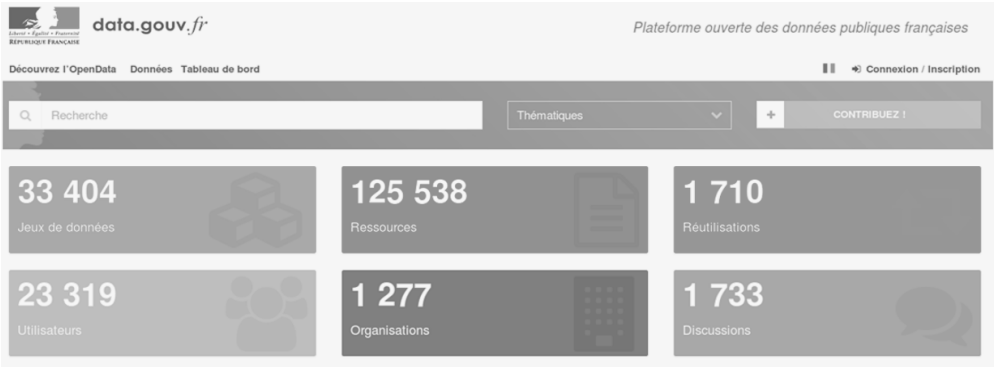
– The Housing Ministry requested access to notarial records. The CDO provided his judicial expertise in the preparation of a decree destined to be presented in front of the State Council.

On December 31, 2017, the CDO had two dozen current cases and around forty closed ones. This sample demonstrates the usefulness and efficiency of the process, although it remains largely unknown. The forthcoming publication of recurrent referrals and the development of the ministerial administrators' network will enable the possibility to consider an increase in the volume of referrals.

Designing and implementing tools and measures to improve data flow

Within the scope of the "State as a platform" approach, the Etalab mission at the DINSIC, operates a whole set of tools and measures (platforms, APIs, services) which facilitate data flow.

The French open data platform



The strategy for the distribution of open data relies on the data.gouv.fr platform. This platform currently includes over 33,000 open datasets from over 1,200 organizations, among which: all the ministries and most of the agencies they supervise; the National Assembly and the Senate; administrative and legislative authorities such as the Constitutional Council, the Court of Auditors, the High Authority for Transparency in Public life (Haute autorité pour la transparence de la vie publique), The National Commission on Information Technology and Civil Liberties (Commission nationale informatique et libertés) and also the Commission on Access to Administrative documents (Commission d'accès aux documents administratifs); as well as local and regional authorities representing all territorial levels, from the smallest municipality to the largest region.

Over the last two years, the number of visitors on data.gouv.fr has increased on a regular basis, reaching over 185,000 single visitors in December 2017.

Trends in the number of visits to data.gouv.fr

	Number of single visits per month	Annual increase
December 2013	47,000	
December 2014	53,000	12%
December 2015	77,000	47%
December 2016	127,000	63%
December 2017	185,000	46%

Verticals: spaces with specific themes

It has become clear in recent years that simply listing thousands of datasets did not allow administrations – and, more generally, society — to harness

their full potential. In fact, optimizing the reuse of this data requires not only that it be discoverable but also to lift any obstacle impeding its exploitation. This should facilitate the appropriation as well as initiate and promote dialogue between producers and potential reusers in order to promote positive feedback loops.

The initiatives, methods and tools used to achieve this are specific to the data in question.

For instance, information required by a potential reuser of a geographical dataset (scale, projection, coordinates system, etc.), has little relevance for a reuser of accounting data, who will need to be aware of the chat of accounts applied. Differences in the nature of the data in question have been at the root of the emergence of various user communities grouped around thematic reference data, formats, tools and practices, resulting in a vertical structuration of the data system. For example, the sphere of geomatics brings together producers and reusers of geographical data under the umbrella of the INSPIRE Directive.

By leveraging this *de facto* "vertical" thematic organization of the data ecosystem, Etalab is now able to enrich the data.gouv.fr platform. The first three initiatives of this kind correspond to existing well-structure ecosystems: those centred on geographical data, transport, and businesses. These verticals allow the creation of a range of specific services corresponding to this type of data and bring together communities of specialists.

The Geographical data vertical

Up until 2016, data.gouv.fr's "gateway INSPIRE" (named after the INSPIRE directive) was only mandated to harvest geographic open data and list it on data.gouv.fr. In 2017 this catalogue became geo.data.gouv.fr, data.gouv.fr's platform for the distribution of geographical data. By taking full advantage of the INSPIRE directive's richness in geographical metadata, this specialized platform allows for a more detailed presentation of the data than the one, more generic, available on data.gouv.fr. It currently references over 100,000 geographical datasets, of which 20,000 are downloadable as open data from 118 local and national portals.

Thanks to the dynamic triggered by the launch of the public data service, three sub-domains have been particularly focused on: administrative geography (regions, departments, and municipalities), addresses and the cadastral map. Two platforms are dedicated to the latter two, adresse.data.gouv.fr and cadastre.data.gouv.fr

At the same time, an applicative programming interface, geo.api.gouv.fr, was launched. It aims at making the access to reference data as simple as possible to reusers. It provides access to updated versions of administrative geography (on geo.api. gouv.fr), as well as geocoding, checking and standardising any postal address (on api-adresse.data.gouv.fr). Illustrating the reusers' interest in the API format of accessibility in complement of the possibility to download raw data, more than a billion calls were made via these APIs in 2017.

Using APIs to facilitate the geographical data integration

Since 2014, Etalab has been developing, in partnership with the National Institute for Geographic and Forestry Information (IGN), La Poste and the OpenStreetMap France Association, the National Address Database, a geo-database of French addresses.

In addition to providing data, Etalab has been developing, since 2014, a geocoding engine, Addok, specifically optimized for address searches, and a free access geocoding API. Measuring the usage of these tools is a good way to prove their utility and pertinence. Indeed, in 2017, the number of users of the National Address Database's geocoding API has significantly increased, reaching over a billion requests and 11.8 million single visits in the first ten months, whereas only 443 million requests and 5.5 million single visits were registered in 2016.

Major e-commerce websites use this online service to improve the quality of the delivery data.

As an integral part of the geo vertical, the Geo API gives access to data from the official geographic code (Code officiel géographique) as well as the administrative boundaries of the municipalities, departments and regions. Similarly to the geocoding API, the Geo API is widely used and has received almost 200 million requests in 2017.

The business data vertical

Following the opening of the Register of Enterprises and Establishments (SIRENE, INSEE), published on data.gouv.fr since January 4, 2017, and the subsequent registration of this business identification reference dataset within the Public data service on April 1, 2017, Etalab has initiated the establishment of a data.gouv.fr vertical dedicated to business reference data and modelled on the geographical data vertical.

For this purpose, there was a collaboration with a section of the "Dites-le-nous une fois" team, which already operates an API meant solely for the benefits of the administrations and aims to disseminate information regarding business identities.

The size of this "business API" was technically redesigned to accommodate any reusers, and became in 2017 entreprise.api.gouv.fr, the first stone of the new business vertical. By the end of 2017, it was used by about a hundred administrations aiming to simplify administrative processes and avoid having to request supporting documents. As a result, in December 2017, more than 1.2 million pieces of information were obtained via the API and without the need to request them a second time from the businesses.

The transport data vertical

In the summer of 2007, as part of a governmental start-up, the Ministry of Transport and the DINSIC Incubator of Digital Services agreed to jointly develop a platform granting access to the transport open data covered by Article L. 1115-1 of the Transport Code. This anticipated the National Access Point, which was mentioned in the EU Regulation adopted on May 31, 2017: transport.data.gouv.fr. By the end of 2017, the stations, stops and hypothetical public transport timetables for the conurbations of Brest, Toulouse and Grenoble had been incorporated into the pilot platform.

Support for the public data ecosystem

The public data ecosystem encompasses both producers (the State, local and regional governments, stakeholders from the private sector and associations) and reusers of data.

In 2017, the Etalab mission participated in funding the Opendatalocale experimentation conducted by Opendatafrance. Nine testing territories were offered support in the implementation of the "open by default" principle. This provision of the Digital Republic Act currently involves all local and regional governments with over 3 500 inhabitants and over fifty employees.

The Opendatalocale framework provides multiple tools to the ecosystem:

– A common core of local data: in order to promote a homogenous opening up of the data throughout the country, data standards have been established in cooperation with reusers and publishers of IT solutions;
– tools to raise awareness and provide support in the form of documents, such as model clauses for public procurement;
– a serious game to train open data instructors.

The foreshadowing of the ministerial data administrators network

In compliance with the recommendations expressed in the Chief Data Officer's previous report, a number of central administrations – particularly the Ministry of Interior, the Ministry for the Ecological and Inclusive Transition and also the Bercy General Directorate of Public Finance – have appointed a ministerial data administrator (MDA).

The members of the network meet under the Chief Data Officer's leadership.

In early 2016, the Ecological and Inclusive Transition Ministry created the position of data supervisor. It was assigned to Laurence Monnoyer-Smith, General Commissioner for Sustainable Development. Three arguments justify this choice: The general commissioner's statistic department possesses the necessary capacities, her mission to provide geographic information and her mandate to coordinate the scientific and technical networks.

On July 1, 2016, Daniel Ansellem took office as ministerial data administrator within the Ministry of Interior. The latter also designated data officers within each business directorate and each operator supervised by its authority (The National Agency for Secure Documents or ANTS, and the National Agency for the Automated Processing of Offences or ANTAI). In 2017, the ministry's data administrator supports two projects within the "Public Interest Entrepreneurs" program (Carte AV and MatchID), and two additional are planned for 2018.

Lionel Ploquin was designated data administrator at the General Directorate of Public Finances in august 2016. He works alongside the director of Cap numérique, a nationwide competent service, in charge of the digital transformation of the directorate. This collaboration reflects the fact that the DGFIP is aware of the crucial role of data management and valorisation with regards to the digital transformation of the public sector.

The Ministry of Agriculture has created the position of general delegate to digital affairs and data within the General Secretary, in December 2017.

3. Making use of data to improve public policy

Beyond the production of essential data or optimizing data flow, it is essential for the administration to be able to harness the data's full potential in order to improve the efficiency of the public action.

In addition to official statistics, which use data to generate knowledge, data science within the administration uses it to develop new operational services aiming to improve the work processes.

Since the publication of the first Chief Data Officer report, various achievements of the in-house data scientists' team, such as fighting unemployment, identifying the business experiencing difficulties at an early stage or providing tools to the services in charge of combatting auto theft and burglaries, have illustrated the benefits of this approach in support of public policies.

The "Public Interest Entrepreneurs" project has also contributed to create new data usage within the ministries.

Fighting unemployment by providing new services for jobseekers

As far as employment policies are concerned, the **La Bonne Boîte**, la Bonne Formation and Bob-emploi projects demonstrate how using administrative data can help develop new digital services to support jobseekers.

Developed by Pôle Emploi in collaboration with the Incubator of Digital Services and the data scientists of the Etalab mission of the DINSIC, La Bonne Boîte gives jobseekers access to the hidden job market, and lets them send targeted spontaneous applications to the companies most likely to employ them.

The service relies on an algorithm which predicts the recruitment probability within a company of a given sector and region. Very easy to use, it currently registers almost 70,000 single visits per month and is used by 9,000 Pôle Emploi advisers. **70% of its users claimed to have found at least one relevant company to contact**[21]**.**

Bob-Emploi, a digital coach that assists jobseekers, was developed by the Bayes Impact Association. This tool uses recommendation algorithms developed in collaboration with Etalab. One year after its launch in November 2016, Bob-Emploi has more than 115,000 registered users of which 86% declare they have found Bob-Emploi useful[22].

Early identification of businesses likely to encounter difficulties

It is acknowledged that public involvement in businesses experiencing difficulties is most effective when it occurs at an early stage. A number of public stakeholders in Bourgogne-Franche-Comté have contributed to the development of **Signaux Faibles**, a tool which aims to help identify businesses that are likely to encounter difficulties[23].

Thanks to an **early identification**, the Commissioner for Economic Recovery (CER) can provide assistance to these businesses, before their situation becomes critical. The algorithm developed by combining the data from URSSAF (Organizations for the Collection of Social Security and Family Benefit Contributions) and DIRECCTE (Directorate for Enterprises, Competition, Policy, Consumer Affairs, Labour and Employment) has already triggered **twenty-five visits of the CER to the businesses, out of which sixteen were deemed useful**.

---

21 https://labonneboite.pole-emploi.fr/stats
22 https://www.bob-emploi.fr/transparence
23 In partnership with the Bourgogne-Franche-Comté DIRECCTE (Regional Directorate for Enterprises, Competition, Policy, Consumer Affairs, Labor and Employment), and the Bourgogne and Franche-Comté branches of URSSAF (Organizations for the Collection of Social Security and Family Benefit Contributions).

Developing decision-making tools for the National Security Services

To prevent and combat auto theft

In collaboration with the Ministry of Interior, Etalab has developed **MapVHL**, a decision-making tool which gives access to records of auto theft and recovery.

Initially, the aim was to develop a predictive model which would allow the identification of potential high-risk areas. However, the collaboration with in field enforcement agencies has shown that the primary requirement, before investing in a predictive model, was to acquire a **detailed knowledge of theft records** via a simple and intuitive interface. Thanks to the feedback of field officers, a mapping of the recovery locations has also been made[24].

After a first demonstration with the Compiègne (Oise) Police force, it was tested by the law enforcement officers of the Beauvais Departmental Security Services. In combination with the use of tablets, it was tested in the field by the anti-crime brigade.

Improving road safety

In order to have a better understanding of road accident cartography and to be able to provide guidance to the security forces in their activities, the Mission for Data Valorisation within the Ministry of Interior's mission on the Ministerial Governance of Information and Communication Systems has developed the **CarteAV** tool (cartography—accidents—verbalizations) within the frame of the "Public Interest Entrepreneurs" 2017" program[25].

Comparing traffic accident data and data containing records of fines is a way of **identifying** dangerous spots, where the security forces rarely intervene, and conversely, the areas where a significant **number of offences are reported** although there appears to be fewer dangers.

CarteAV was developed in 10 months' time by a team of two data-scientists/developers and was tested in short cycle by around 20 police services.

Facilitating the work of the administration via the automatic comparison of databases

The Central Office for Combating Illegal Employment (Office central de lutte contre le travail illegal, OCLTI) is mandated to combat illegal employment and especially the fraudulent intra-European posting of workers. Its mandate requires a frequent comparison of databases in order to identify the victims of transnational fraud. In practice, this means manually examining if the lists of employees match the databases of the national social security in order to confirm that they are indeed registered.

---

24 For detailed feedback, see the CDO's blog post: https://agd.data.gouv.fr/2018/01/12/predire-les-vols-de-voitures

25 The source code is available on Github: https://github.com/eig-2017/cartav

By using fuzzy matching methods and existing free software [26], the team of data-scientists has demonstrated the possibility to obtain very satisfying results with automated methods [27]. The systematic implementation of these methods would save days of human labour and make for greater efficiency in combating the fraudulent posting of workers.

The "Public Interest Entrepreneurs" Program

Launched in late 2016, the "Public Interest Entrepreneurs" program is an innovative and original initiative. Each administration can propose a challenge that will be solved in 10 months by a 1 to 3-person team.

The winners, selected by a jury of experts in the digital field as well as administrative agents are hosted by the administration. They are given access to the databases and supervised by high-ranking mentors as well as the Etalab team.

Twelve challenges were selected after the second round:

– SocialConnect, presented by the General Commission for Territorial Equality (Commissariat général à l'égalité des territoires),to develop a collaborative platform to network and showcase the social innovation ecosystem in the territories;

– Signaux Faibles, presented by the Bourgogne-Franche-Comté DIRECCTE, to develop tools to identify businesses in difficulty;

– PréviSecours, presented by the Ministry of Interior, to develop a first aid predictive model to help firefighters make better use of their resources;

– Lab Santé, proposed by the Ministry of Solidarities and Health, to analyse data from the National Health Data System (SNDS);

– Hopkins, proposed by the Ministry of Action and Public Accounts, to detect financial fraud;

26 See the github dedupe python library: https://github.com/dedupeio/dedupe
27 http///agd.data.gouv.fr/2016/11/22/rapprocher-deux-bases-donnees/

– Gobelins, proposed by the Ministry of Culture, to reveal the richness of the National Furniture (Mobilier national);

– PrédiSauvetage, proposed by the Ministry for the Ecological and Inclusive Transition, to determine the causes of sea accidents in order to prevent them;

– DataESR, proposed by the Ministry of Higher Education, Research and Innovation, to develop a platform and analyse data originating from the higher education and research sectors;

– CoachÉlèves/Assistprof, proposed by the Ministry of Education, to create digital coaches that will provide learning assistance for students and teachers;

– Brigade numérique, proposed by the Ministry of the Interior, to develop a service of digital reception for police stations for the benefit of citizens;

– B@liseNAV, proposed by the Ministry of the Army, to create an enhanced nautical chart to make navigation safer;

– ArchiFiltre, proposed by the Ministry of Solidarities and Health, to develop an automatic filtering method for unstructured data, destined to be archived.

A group of scientific experts to assist the administrations in their choice of service providers

The administration doesn't always have the ability to assess the scientific relevance of an initiative proposed by a commercial service provider with regard to data-science solutions. Therefore, it may prove difficult to assess its pertinence.

To assist the administrations and ensure that they are not forced to use algorithms they do not understand, Etalab offers to convene a jury of scientists composed of researchers and distinguished scholars to help them better understand and evaluate the pertinence of a solution suggested by a provider.

Concretely, the jury may question the provider on various aspects of the algorithm, evaluate its pertinence in the context of its expected application, or ask questions regarding the choice of the knowledge base to train it.

The jury met for the first time in December 2017, to advise the General Directorate of the Judiciary Police (DCPJ) in its choice of a commercial solution regarding a geographical targeting algorithm to guide investigators who specialise in serial crimes*.

* https://agd.data.gouv.fr/2018/01/19/un-appui-scientifique-aux-administrations/

Part Two

Data as an essential infrastructure

According to the decree creating the Chief Data Officer position, the report submitted to the Prime Minister should, in addition to presenting the state of play as regards to data policy, detail the **"developments in the data economy"** and contribute to planning the State's action in this domain. Hence, the second part of this report is devoted to an analysis of the central theme: **data as an infrastructure**. By offering a comparative approach of the initiatives undertaken by various European countries, it advocates the creation of a fully-fledged data infrastructure. The State can and must play a central role in **a "State as a platform" approach**.

1. Data should be considered as an infrastructure

In their respective fields (transport, tourism and the hotel industry), **Uber, Airbnb and Booking**: demonstrate the **shift towards new business models**. Possessing material assets has long constituted a key factor in the structuring of these markets. For example, becoming a stakeholder in the hotel industry used to require the capacity to invest in real estate. However, the dominant position occupied by an actor such as booking.com today, illustrates the fact that this paradigm of resting on material assets is widely called into question.

In economic environments dominated by the concept of platforms, it is **the possession and use of intangible assets** that makes the difference. Foremost among them and well ahead of the brand is data and the ability to process it. Some of these actors owe their considerable market power to the volume of data they were able to collect within their brief existence span.

It is clear that there is a **widening gap** between the intangible nature of the resources available to offer a service and the concrete nature of its impact.

These online activities have major impacts on reality, to which most major cities are already confronted. New York, Barcelona and Paris are introducing regulations to address the unavailability of residential long-term rental offers in touristic areas caused by the activities of short-term rental platforms. In San Francisco, Uber transports as many passengers as the city's public transport network, and in some cities, it even advertises itself as a complement to the public offer. A major part of these private initiatives' success – which challenge and in some cases pose a serious threat to public initiatives – is their use of data.

Two elements define **the role of data** in these platforms' strategies. The first involves the gathering of data: "if it can be transformed into data, it is". **Data collection is massive and uninterrupted**. Every interaction with a client, every request, every click is turned into data. Records of use, clients' locations and also their ratings provide the algorithms with a constant supply of data. The second distinctive trait concerns the exploitation of the data, which is also constant and massive. In a data-driven enterprise, data is found at each stage of design and provision of a service. More importantly, the use of data makes it increasingly difficult to separate the stages of design and production as shown by the A/B tests which optimize webpages (or article titles) in real-time.

In many ways, the State is involved in the debates surrounding the platforms and data usage. First, as a regulator: in early 2018, the Parliament will be reviewing the "Data Protection Act" (loi "Informatique et libertés") to adapt it to the new measures introduced by the European Data Protection Regulations (GDPR). Furthermore, France is to play a leading role in Europe by spearheading discussions on the taxation of digital enterprises.

The State can also take **an offensive rather than defensive approach** in the face of the challenges and opportunities created by these platforms, through the creation of a public data infrastructure.

The public infrastructure of the 21st century

According to Nigel Shadbolt, Vice-Chairman of the Open Data Institute: "Data is a new class of public infrastructure for the 21st century. It is all around us and easy to miss. We need to view it as an infrastructure that is as fundamental to modern society as power and transport, and which requires investment, curation and protection."

A country's development is closely tied to the existence of efficient and high quality infrastructures, whether roads, railroads, energy networks or telecommunications. That is why the State has long spent a significant portion of its investments for the maintenance and construction of this infrastructure. Indeed, on average, members of the European Union allocate over **one third of their public investments to this spending**[1]. By way of illustration, 1 km of freeway represents an investment of €6 million, and 1 km of high-speed rail link represents €16 million. These infrastructures contribute to the land planning and development, facilitate trade as well as the movement of goods and people. The externalities generated are predominantly positive. Conversely, slow rates of development in certain countries can be explained by the existence of deficiencies in their infrastructures.

Nowadays, data must be seen as one of the **essential and critical infrastructures**. Essential, because, in an information-based economy, access to reliable and up-to-date reference data is a **precondition for the development** of digital services. Critical, because significant care must be taken to ensure that the supply of this data remains uninterrupted, neither by an involuntary failure nor malicious acts. In this sense, these infrastructures may be considered "activities of vital importance"[2]. The development of smart cities or smart grids relies heavily on securing data availability and access.

Many European countries have grasped the fact that data must be regarded as a crucial public infrastructure, on the same level as physical infrastructures.

In 2013, the German Federal Government created a ministry in charge of transport and digital infrastructures. In matters of physical and informational infrastructures such as the connected cars, its investments are managed through an integrated approach[3]. Across the Channel, the National Infrastructure Commission, which formalises the long term national strategy and channels the investments in regard to the country's crucial infrastructures, has integrated the issue of data into its field of activity. Also, in 2015 the OECD stressed (in its "Data-driven innovation report[4]") the importance of such infrastructures in terms of economic and social development.

---

1 Infrastructure in the EU: Developments and Impact on Growth, European Commission, 2014.

2 According to the definition provided by the General Secretariat for Defence and National Security (SGDSN): "Since they assist in the production and distribution of goods or services that are essential to the exercise of the State's authority, to the functioning of the economy, to the maintenance of defence capabilities or to national security, certain activities are considered to be of vital importance."

3 "Policies for mobility and modernity", Federal Ministry of Transport and Digital Infrastructure (BMVI), 2014.

4 http://www.oecd.org/innovation/data-driven-innovation-9789264229358-en.htm

2. The purpose of a data infrastructure

Regardless of the different approaches used in Europe (below we will outline the initiatives undertaken by the United Kingdom, Denmark and Estonia so as to compare them with the French initiatives), it is striking to see **they share the same observations and goals**.

Data which is underused or does not circulate is not exploited to its full **usage value**. If it is not made accessible to users and does not benefit from the feedback to enhance it, its quality can **deteriorate** rapidly and it will tend to lose its value over time.

The absence of a high-quality data infrastructure has very real financial consequences. For example, mistakes in the transcription of an address causing the non-delivery of mail represents an incremental cost of about €300 million per year.

The Fouilleron report has identified a number of cases where administrations are duplicating or reproducing already existing databases, if not actually **creating their own databases**, because they have been unable to openly, freely and securely access data produced by other administrations.

The costs of poor data circulation and quality may be described as follows:

– The direct and indirect losses associated with the use of inaccurate date;
– the maintenance of redundant databases and the costs of duplicate data entries when a reference database is not made available to all public and private actors, who need to access it. This applies for example to local and regional administrations (collectivités), which until recently did not have access to the official database of associations, despite being the primary source of financing of the voluntary sector;
– the transaction costs associated with research and the acquisition and processing of public data can be very substantial for industries with a critical need of data.

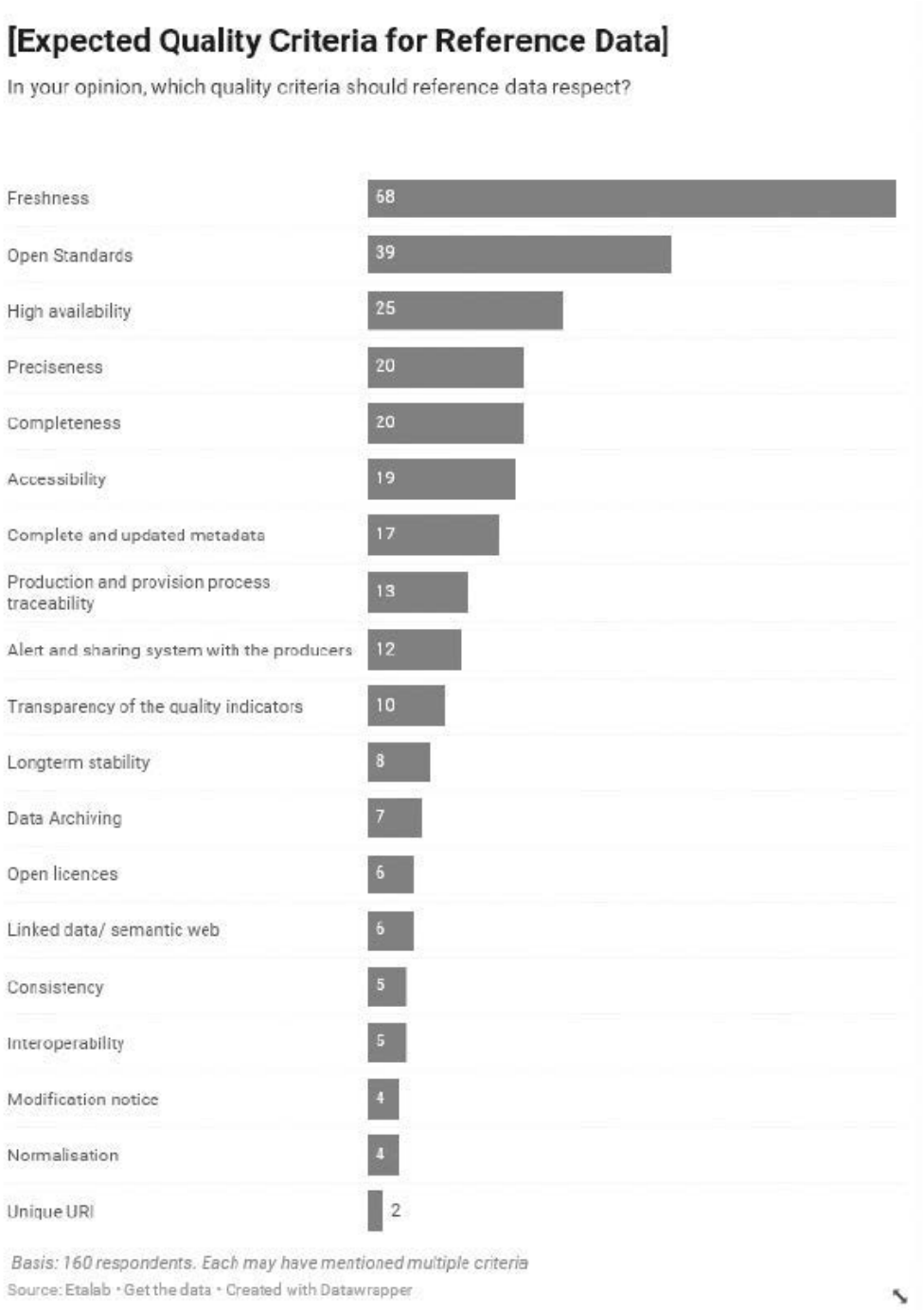
An infrastructure that ensures the optimum use of data

Roads are designed to facilitate trade and the transport of goods and people. Telecommunications networks aim to spread information and facilitate the cooperation between individuals. A data infrastructure also has a purpose: to ensure the optimum use of data.

Obviously, such an infrastructure would primarily be intended for the administration, but it ultimately concerns all those (businesses, associations and civil society) who use public data. All the EU members who have made a commitment to build a data infrastructure **share this same objective**.

What should be done in order to provide all stakeholders with an access to the data? Which **commitments** should be made by the public authorities?

To answer these questions, Etalab carried out a public consultation [5] amongst potential users of reference data. Thanks to the participation of 160 respondents (stakeholders from the public and private sectors and associations), it was possible to identify specific expectations, particularly regarding the standards required for reference data.



5 The consultation took place from September, 29, 2016, to October, 20, 2016; a summary of the results is available online: www.etalab.gouv.fr/consultation-spd

Up-to-date, available and easily reusable data

It seems therefore very clear that the primary expectation is the **"freshness"** of the data (data updates and the time lapse between an event, e.g. between the registration of an association and its mention in the published database). Another expectation is the **high availability** of the infrastructure (99.5% on a monthly basis). The use of **open standards** (second most frequently cited criteria) was introduced into the Code of Relations between the Public and the Administration under the Digital Republic Act.

Completeness, accuracy and freshness are the usual requirements in terms of **data quality**. But users also expect traceability during the production and publication processes, as well as the possibility to interact with the producer (to report errors or suggest improvements), in addition to transparency regarding the benchmarks used to control the quality and publication of data.

"Data you can rely on"

Invisibility is one of the characteristics of an infrastructure.

Anyone turning on a tap expects an instant supply of clean water. It is only when there is a temporary interruption in the water supply (or in the provision of electricity/telecommunications/transport) that we are reminded of the full extent of the efforts and resources that have been put in place in order to provide us with an instant supply.

For now, the supply of public data has not reached this standard. Some of the data is not updated frequently enough. Other data is inadequately or poorly documented. Producers may modify data patterns for their own internal purposes without consulting, or even informing, reusers of forthcoming changes.

This is prejudicial since the data infrastructure relies on the same characteristics as the others. Its users voice the same need to trust it: "**I need to trust** the quality of the data supplied in order to use it to provide a service or elaborate an analysis based on its elements" [6].

If the service is interrupted or the standards deteriorated, then the users' **trust** would be broken and the infrastructure wouldn't have served its purpose.

3. A Benchmark of European initiatives

The countries committed to this path share the same idea regarding the **purpose** of the data infrastructure and the expectations associated with it. However, the **means to reach this goal** vary, so do the methods **to build** the infrastructure. In order to put into perspective the various approaches,

---

6 One particular user compares the provision of reference data to the standardisation which exists in the hospitality sector "When I choose a hotel belonging to the Novotel chain, I am confident that my experience as a customer will be identical in all countries. There are no surprises and I know exactly what I am entitled to expect."

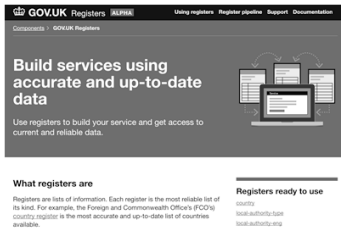
we will analyse the initiatives undertaken by the **United Kingdom**, **Denmark** and **Estonia** as compared to the French.

GOV.UK Registers (the United Kingdom)

In 2013, the British Government launched the **National Information Infrastructure** project (the NII) in response to the recommendations made by the Shakespeare Review. From 2013 to 2015, the Cabinet Office established the general guidelines of the NII[7] and proceeded to identify the datasets concerned. In 2015, the Chair of the Open Data User Group (ODUG-UK, the organisation representing the users of open data[8]) submitted a report that criticised the implementation of the NII, highlighting the lack of prioritisation in its efforts. The first list of data published on data.gov.uk consisted of 233 datasets, and their quality was not satisfactory (some stopped being updated a few months after their initial publication).

Amongst other recommendations, the Open Data User Group suggested the Government should first focus on a few core datasets, namely **nomenclatures**.

Being less ambitious than the NII, GOV.UK Registers falls within this perspective[9].



GOV.UK Registers is promoted by the Government Digital Services (GDS). The United Kingdom presents it as an element of the Government-as-a-platform approach, just as GOV.UK Notify (which manages notifications), Pay (which managements payments) and Verify (an ID Verification System)

"The most reliable source of data in its field"

Fourteen registers are currently available online and forty-five others have already been identified as potential future registers. Each one must be "the most reliable source of data" [10] in its particular field.

On average, each dataset contains several dozens or hundreds of entries mainly concerning **nomenclatures** relating to administrative and territorial organisation (from countries to counties), as well as to some facilities (prisons) and public infrastructure (organisations in charge of the drainage network system).These registers can be downloaded in **various formats** via APIs. The pattern of the Register Platform API is common to all the registers, which facilitates their integration by third party developers.

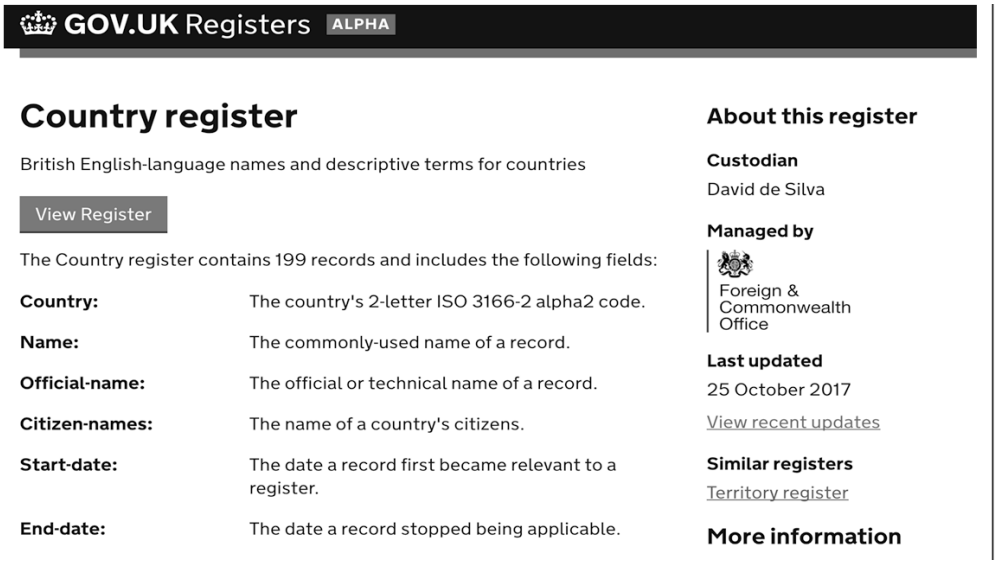
---

7 See https://data.gov.uk/consultation/national-information-infrastructure-prototype-document/what-national-information

8 The Open Data User Group organisation has since been dissolved.

9 Today, according to various contacts we were able to interview in the UK, the National Information Infrastructure project is on standby.

10 "Each register is the most reliable list of its kind" – What registers are? https://registers.cloudapps.digital/



A centralized governance and a high level of accountability on the part of producers

The Government Digital Service (GDS) is responsible for maintaining the registers' platform. In this capacity, it has developed a single approach to ensure the recognition and inclusion of new registers. Each ministry may **apply for an official recognition of its data as a register** by the GDS. The latter has established a list of three **eligibility criteria**: the register can only contain raw data that is not derived from statistics; it mustn't contain personal or private data and finally, there should be no objection to its publication as open data.

The **final decision on whether to include** a new register in the official list rests with the **GDS**. It takes into consideration not only the users' requests, but also the producer's profile [11].

The register's producer must appoint a custodian chosen among his agents. This is a personal commitment: the agent's name will appear on the register's page. The agent must update the data, provide answers to the users' questions and act as the GDS' contact person. In the event of an absence, temporary or permanent, the custodian must designate his replacement. All custodians are trained by the GDS team in charge of registers.

11 In particular, the GDS reserves the right to determine which administration is the most pertinent to maintain a register, in case several administrations declare themselves competent.

Basic Data – Grunddata (Denmark)

The Grunddata program is part of Denmark's **digital strategy** which was initiated in 2012 by the Danish government and the local municipalities, joined by the regional governments, in 2013. It is currently run by the National Digitisation Agency attached to the Ministry of Finance.

Denmark has a long tradition of digitising registers and administrative databases. Its legislation **facilitates the cross-referencing** between databases thanks to a **unique individual identification number**.

By way of comparison: in France, the use of the National Register for the Identification of Natural Persons' identification number (or social security number) to cross-reference databases is strictly regulated by the law.

The Grunddata program is very broad and covers the complete life cycle of reference data, starting from its production to its circulation and funding. Its key initiatives illustrate this ambition:

- the identification and **quality improvement** of the main registers;
- the **convergence** of the main reference databases, with extensive modelling of the data and their connexions;
- the implementation of a **Data Distributor** to replace the current dissemination systems;
- the implementation of a strong governance, led by the Ministry of Finance, with its **own budgetary lever**.

A thematic structure

Grunddata is structured around six theme-oriented data categories:

- **real-estate and land property**: the land register but also deeds of land property;
- **individuals**: Data from the Civil Register (personal identification) including the name, address, civil status, filiation (children and parents), citizenship rights.
- **enterprises**: registration data but also data regarding their financial results and workforce;
- **addresses**, roads and administrative divisions;
- **maps** and geographic data;
- **water** and **climate**: maps and hydrographic models, meteorological data.

Each category has been **assigned to one or more producers**, under the **Basic Data Board** (see below on governance) supervision. Since 2013, efforts have been made to identify the initial situation and the areas of improvements.

For example, in the addresses' category, a number of initiatives have been deemed necessary: such as the numbering of public and private spaces who did not have an address (gardens, a number of industrial zones) and the creation of a unique and permanent identification number allowing the identification of premises, even in the event of changes in the street name or a renumbering.

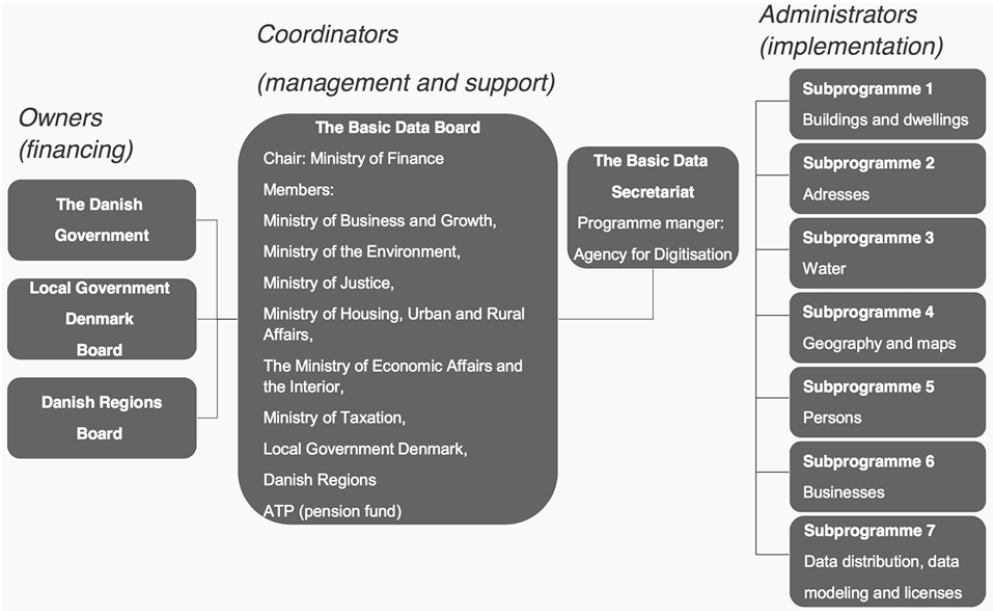
Two projects involving all categories were launched:

- the definition and implementation of an **inter-register data model**, aiming to create a stable and permanent link between different databases (for example using a single identification number for all the registers);
- the implementation of a single national[12] **Data distributor**, aiming to replace the existing distributors and provide a high level availability (for example by updating the most dynamic registers daily).

A strong governance that includes the funding of reference data

The **funding of producers** of basic data was promptly identified as vital to the program's success. When the Grunddata program was launched, a number of producers still relied heavily on fees, which could represent as much as 80% of their budget. Furthermore, they acted as both the project manager and product owner.

The Danish Government created a governance specifically dedicated to basic data, and provided it with its own budgetary lever. **The Grunddata Board** regroups all the basic data producers and the main financial contributors (the Government, the municipalities and the regions) is chaired by the Finance Minister.



12 See http://datafordeler.dk/

This structure manages all the budgets associated with the **production and distribution of basic data**. Consequently, the producers have lost the relative autonomy they previously enjoyed concerning part of their budget (around 20%) in the interest of a **single and centralised management**. In this context, every ministry or state agency in charge of an area has a specific budget at its disposal, for the implementation of initiatives adopted jointly with the Board and the Digitisation Agency (operator of the Grunddata program).

Because it was **very ambitious**, involved multiples actors – Government, municipalities, private sector – and covered various aspects of issues surrounding the data, the Grunddata program has experienced **delays** in its production and distribution.

A longer, and more complex implementation than anticipated

In November 2017, five years after the launch of the program, the Data Distributor has only just published online its first register: a list of the names of Danish locations. The revised calendar plans a progressive online publication over a period running up until 2020 for some registers. In order to finalize the project, a request was made to the Danish Parliament for the allocation of **additional funds**. The implementation of the centralised governance and the work on converging databases have proven to be more complex than initially anticipated.

X-Road (Estonia)

X-Road, presented as the backbone of Estonia's e-government strategy, is a textbook case on the European level. Initiated in 2000, X-Road aims to **connect administrative databases** with the perspective of simplifying administrative procedures. The system relies on a mandatory digital identity issued for all Estonians above 15. Today, this system records over 95% of the Estonian population.

The data producers attached to X-Road offer upon request services **to registered users only**. The producers dictate the access conditions to their services. Hence, the tax department determines who will be given access to the tax certificate delivery service. The Information Systems Authority (RIA) 'regulates the access to X-Road by examining the requests to join the system both by data producers and service users (administrations and businesses).

To date, X-Road provides access to over 1,200 services offered by around 150 data producers and 950 service users. Most of the latter are administrations or businesses that require access to the identity data recorded in the databases of the Estonian administrations.

In 2017, the most used services (in terms of number of requests) were [13]:

13 Source: http://x-road.eu/xtee-stats/

– **tax statement**, certifying that an individual is up to date on his tax payments;
– **medical insurance entitlement document**, to ascertain whether a person is insured or otherwise give him access to new rights;
– information regarding the **professional status** (active, inactive, jobseeker etc.);
– access to a patient's **medical file** (including the history of medical prescriptions).

The main "consumers" of the services provided via X-Road (in terms of numbers of request in 2017) are the **banking and financial sectors** (especially when studying a loan attribution), the **medical sector**, the fiscal Estonian administration, the police and national defence services.

It differs from the previously described initiatives (Denmark and the United Kingdom) on multiple aspects:
– The purpose: X-Road was developed within an e-administration framework, and for the purpose of simplifying administrative procedures rather than the broader perspective of data exploitation.
– Access regulation: contrary to the British and Danish initiatives, data access is given for the completion of a specific task for which a new access permit should be requested each time.
– The type of data provided: most of the databases interconnected via X-Road regroup personal information data such as records of taxpayers and landowners, medical files etc.

In this sense, the Estonian approach is comparable to the administrative simplification programs (such as "dites-le-nous une fois"), rather than to an open data infrastructure such as the one envisioned in the UK and Denmark.

The private sector pays to access reference data (the cadastral survey and the business register)

Alongside the X-Road system, several registers – including the cadastral survey and the business register – are available online.

In exchange for a fee, private stakeholders can access these registers and download data. For information, in 2015 the revenue generated by these fees exceeded **€3 million**[14].

Comparing with the situation in France

The idea of an "**energy-related data public service**" was first mentioned in 2014 within the context of the Energy Transition Act[15]. At the time, the aim was to promote the flow of data produced by the energy network operators.

14 This revenue covers approximately 70% of these registers' production costs. Source: http://www.rik.ee/en/e-land-register/service-fee-rates
15 http://www.assemblee-nationale.fr/14/pdf/cr-cstransenerg/13-14/C1314005.pdf

The following year, the preparation for the Digital Republic Act was an opportunity to lay the foundations for the public data service. Announced in October 2016, the Digital Republic Act has therefore created a new public service in "charge of making reference data available, in order to facilitate its reuse" [16].

Focusing on the dissemination of data: governance through objectives

According to the legislator's intentions, the **public data service** should focus on the **distribution** of reference data, rather than on the preconditions of its production. Indeed, most of the criteria established through the legal texts concern primarily the provision of the data: availability, freshness, metadata quality, etc.

At the same time, the choice was made to opt for a governance based on the **objectives** rather than one based on assigning responsibilities on the producer and the publisher. Contrary to the Danish model where data is ultimately exclusively published via the national Data Distributor, in France, producers **are free to publish reference data themselves**, or to **entrust its publication to a third party.** Either way, the technical and organisational availability regulations hereunder must be respected.

The technical and organizational availability regulations with regard to reference data

The main rules established by the June 14, 2017 decree involve:

– the documentation of data: metadata;

– the information supplied to users regarding the reference data creation process;

– the producers' responsibilities regarding the frequency with which each reference database should be updated (from annual to a daily basis depending on the producer);

– the availability rate: 99% monthly for download, 99.5% for programing interfaces;

– the conditions of provision, which must guarantee the authenticity of the data;

– the procedures for reporting to the producer errors and incompleteness in reference data, or in some information associated with it;

– the observance of a three months maximum time frame to inform users of any significant modification in the reference data's characteristics, conditions of availability and the structure of the database.

16 For a detailed description of the concept of reference data, see part 1 of this report.

Building up

The decision to build the data public service through **the dissemination** of reference data, rather than through its production along the lines of the Danish example, has had several impacts. First, it allowed the beginning of the open data distribution of nine reference datasets, only a year after the promulgation of the Digital Republic Act.

In our opinion, this **agile approach**, which is gradually being developed in interaction with the users, is the most efficient in terms of rapidity of service delivery.

However, building the infrastructure by **starting downstream** rather than using the Danish upstream approach (the dissemination follows an important research regarding the conditions and criteria of production) also implies a **set of challenges**.

The first challenge involves the creation of a suitable governance. Dividing the responsibilities between the producer and the publisher[17] is also a crucial point. Moreover, users should be involved in the evolution of the reference data public service as well as in the governance. Their contribution is essential, particularly when it comes to fostering a **gradual quality increase of the reference datasets**.

Summary Table

	UKRegisters	**Danish Grunddata**	**X-Road Estonia**	**FR Public Data Service**
Data	nomenclatures	registers	identity Registers (civil status, taxes, health)	registers and administrative databases
Volumetric measurement of data (order of magnitude)	several dozens to several thousands of entries	several million entries	several million entries	several million entries
Personnal data	explicitely excluded	included	included and predominent	intented to be included
Approach	by dissemination	by production and dissemination	through the interconnexion of databases via digital identity	by dissemination
Governance centralisation	limited	significant	mixed: a central authority (link/ connection) and producers (access rights)	weak/limited
Method of cooperation with the producers	via certification and personal accountability of the administrators	via the budget allocation	by a central authority	by following the objectives and the monitoring of public engagements

17 For example, the publisher cannot commit to the frequency of updates. On the contrary, an updated reference data rendered unavailable by a service interruption in the programming interface (API), is useless.

Lessons to be learned from the European initiatives

The analysis of the Danish, British and Estonian initiatives is very instructive in the perspective of creating a French data infrastructure of which the data public service is the first draft.

The first lesson to be learned is that creating a complete data infrastructure capable of meeting the challenges at hand requires not only time and investments, but also a **strong and sustained political commitment over several years**. One cannot build an informational infrastructure – let alone a physical one – in two, or even five, years.

France can draw on the expertise of the major public data producers (including INSEE, IGN, Météo France and the DGFiP). But the construction of a data infrastructure must be regarded as a fully-fledged public investment for which the funding must be permanent.

The second lesson is that the creation of a data infrastructure depends on multiple levers, not all of which are technical:

- **A budgetary lever**: Denmark chose to centralise the funding of reference data production within an inter-ministerial body, resulting in a partial loss of financial autonomy on the part of the producers.
- **A contractual lever**: objectives established in ministries, contracts relating to operator performance and resources must mention the contribution of each producer of the data infrastructure.
- **A legal lever**: the efforts leading to the opening up of public data (based on a free access principle, reuse licenses, open standards) are catalysts for the construction of the data infrastructure.

Finally, **the choice of a governance model** appears to be a structuring element of a data infrastructure. Centralisation, to some extent, is necessary, if only to establish a minimal set of rules and standards, shared by all reference databases.

A road map for the data infrastructure

A data infrastructure is composed of:

– high quality data and particularly data that meets the standard of reference data;

– access infrastructure, via APIs and download;

– identification, security and control mechanisms;

– mechanisms allowing users to participate in the improvement of the data standard.

Some **software components** of this infrastructure are already in place: for example, the data.gouv.fr portal offers the possibility to download content, as well as the opportunity to interact with the data producers (with the possibility to alert them and via discussion groups dedicated to each dataset). Furthermore, some reference data is currently displayed via APIs, namely geographical data or data relating to enterprises (business APIs). These APIs are referenced in a catalogue: **api.gouv.fr**, where users can locate it and contact its producers.

An earmarked funding from the **Investment for the Future Program** (Programme d'investissements d'avenir, PIA) reinforces the existing processes (especially in terms of securing projects and performance) and supplies progressively the missing elements. The data infrastructure is intended to maximise data flow. Concretely, this means that when nothing impedes its publication, data should be made available on the basis of the open data principle.

However, in the case of databases – or certain sections, e.g. such as the one containing the identification elements of an association's manager – containing **personal data or data covered by legal secrecy, the data infrastructure must ensure that it is only accessible by the authorized users**. In order to guarantee this secrecy, **identification and access control mechanisms** will progressively be implemented, thanks to the progress already made by FranceConnect and FranceConnect Agents.

The users of this infrastructure play an active part in its success and governance. More particularly, they participate in improving the quality of the data by reporting errors and suggesting updates. This aspect of the data infrastructure is vital. Conversely, a broader data flow cannot be wished-for without introducing a possibility to interact with the producer because it would result in the propagation of defective data and an increase in the monitoring and correction costs.

Part Three

Moving to action

Since its creation in 2014, the Chief Data Officer has participated in the implementation of a fully-fledged **data policy**, with its several aspects: technical, economic and legal. The laws have evolved. Tools, platforms and APIs have been developed and are now used by numerous users. Essential data is now available in open data. Thanks to several very concrete achievements, **the benefits brought by data science** to public policy have been confirmed. The mind-sets have also changed, and several ministries are starting to integrate data into their strategy and actions.

We must now move into action. Now is no longer the time to prove the value of a better data exploitation. **It is time to start implementing** the change within the administrations, starting with the ministries and major operators. Because it is attached to the Prime Minister's Office and the DINSIC, the CDO must guide the administrations in endorsing the data revolution.

The Chief Data Officer's road map for 2018 covers five areas:

- To provide access to data and shared infrastructures, and scale them.
- To develop a doctrine regarding the data flow within the public sphere.
- To strengthen the ministerial data administrators' network and use it as leverage for the data policy.
- To develop an expertise in the use of artificial intelligence in public initiatives, thus making the State one of the first users of these tools.
- To support the ecosystem of users exploiting data produced by the administration, measure its social and economic impact, as well as its influence on the transformation of public action.

1. Making data, resources and infrastructures available

Data with major impact on the economy and society

The last couple of years have witnessed the provision of high-impact datasets: business databases, cadastre, national address database etc. Meanwhile, sectoral approaches have been developed to facilitate the reuse of geographical data and data pertaining to businesses.

In 2018, in cooperation with the ministerial data officers (see below) we will identify the data that may qualify as reference under the French Code of Relations between the Public and the Administration, within each ministry and sphere of public policy. Some of it is already available online, but may fall short of the high standards set by the data public service. Others however are not yet published. In both instances, working in collaboration with the producers will make the data more easily discoverable and reusable. This could for example include the development of APIs or any other necessary resources.

Particular attention will be paid to the quality of the data published online via the collaboration with other initiatives such as the Qualidata project, winner of the Investments for the Future Program (Programme d'investissements d'avenir, PIA). Data users will be encouraged to participate in the quality improvement of the data through the development of an error reporting tool and by suggesting enhancements.

Data standards and infrastructures

In 2018, the first step will be to implement the defined standards in the public orders and the subsidies agreements. It will first require to rally the relevant ecosystems: the administrations (State, regional governments) but also the software solutions editors and all those who use the data. This effort will be organized internationally within the framework of the Open Contracting Partnership, presided by France since the end of 2017 (see above). Amongst the first potential applications to be considered within the context of this partnership, is the fight against corruption through the analysis of the public orders' data.

In addition, we will sustain our effort to set new standards for data. Translating legal rules and obligations into data standards will not only simplify their application, but also facilitate the emergence of ecosystems that reuse this data.

Regarding the infrastructures, the DINSIC will continue to provide access to shared tools that facilitate the flow of data: the open public data platform, theme-based verticals (geography, business and transport), APIs, and France Connect Identité system. These tools facilitate the circulation and exploitation of data. They contribute to the implementation of a real data infrastructure as outlined in the second part of this report.

2. Developing the concept of data flow within the public sphere

The principle of the opening up of public data by default is now enshrined in the law. The tools (APIs, platforms – exist to allow the broadest possible distribution of data, that is not covered by secrecy, and can therefore be shared freely and to the largest number of people. However, this does not provide a complete response to the issue of data flow.

Giving the right people the right data and managing the right to know

In 2018 the Chief Data Officer will participate in the development of **the data flow doctrine**, including data protected by secrecy and principally the right to privacy.

As it was already stressed in the Chief Data Officer's previous report, secrecy does not mean destruction of information. On the contrary, a secret is an information made available to a restricted number of persons. More importantly, in cases of legal secrecy, it is crucial to determine precisely who is excluded from access and under which conditions. In other words, in order to distribute data, it is essential to know who has **the right to know** and to be able to manage this right.

Managing the right to know begins with an investigation: who is requesting access? What is his mission? Which information does he want to access? Based on this analysis, we should be able to give access to data, including the one protected by secrets. Managing the right to know also means taking responsibility for processing a request and guaranteeing its traceability.

Our renewed ambition is to be able to supply **the right people with the right data**, in accordance with the right to know.

To make this theory operational, we will concentrate our efforts on two levels. Regarding the data of which the circulation is subject to control: we will supervise the scaling of API Particulier and France Connect Identité. For data that must be anonymised or that requires the use of a pseudonym: we will develop, with the help of the Etalab data-scientists, an expertise and the tools necessary to support the administrations in the process of publishing this data.

3. Strengthening the network of ministerial data administrators

Some ministries and ministerial directorates have appointed a data administrator modelled on the Chief Data Officer. In 2018, a priority will be to **strengthen this emerging network** and turning it into an effective tool for the advancement of data policy.

This mission starts with the appointment of a ministerial data administrator within each ministry. His task will be to implement the data policy within the context of his home administration, with regard to 4 aspects: inventory and mapping of existing data, production of essential data, optimal data flow, data exploitation including through data science.

The administrator must also act as guarantor of the good knowledge of legal issues, particularly since 2018 will see the promulgation of the new General Data Protection Regulation (GDPR) and the spread of the principle of open data by default.

Considering their importance in the **production of essential data**, it is advisable that the administrator also appoints a data manager within the **operators**' ranks and in accordance with the line pursued by the administrator within the supervising ministry.

The inter-ministerial position of **the Chief Data Officer** confers him the responsibility to organise and strengthen this network. The mutualisation of methods, shared difficulties and solutions should enable new skill acquisition in the administrations. The CDO will also guarantee the coherence of the initiatives undertaken by the ministries, by keeping in mind the objective of the pooling of resources.

4. Developing a centre of expertise in Artificial Intelligence

The technological field surrounding data and its usage is in constant and rapid evolution, as illustrated by the fast spread of artificial intelligence including in the form of machine learning algorithms (and deep learning). Artificial Intelligence, its potential and risks, challenges the State in its responsibilities as regulator and operator of public policies.

In 2018, under the leadership of the Chief Data Officer, the DINSIC will develop its expertise and capacities in the field of Artificial Intelligence. The data scientists of the Etalab mission have already successfully implemented such approaches, including in the frame of the OpenSolarMap project (see below).

OpenSolarMap: combining human intelligence with artificial intelligence

What if we were able to quickly and simply identify the photovoltaic potential of a building? This capacity would allow us to evaluate the suitability of installing solar panels. It is the objective of OpenSolarMap, which combines crowdsourcing and machine learning to produce new data.

The method applied by OpenSolarMap consists in the deduction of the shape of a roof by analysing satellite imagery, provided in open access by Spot Satellites. This provides us with the most important elements to analyse the opportunity of installing solar panels: the slope orientation and surface area of a roof. The first graphical interface was developed using this method. It is in the form a game in which each user is asked to indicate the slope orientation of a roof. Thus, in less than a month, the platform has collected around 100 000 quality analysis. By crosschecking various analysis of one building, around 10 000 roofs have been successfully categorised.

On the basis of this sample of categorised roofs, an automated classifier was developed using conventional image processing techniques (logistic regression and deep learning). The resulting algorithm is only mistaken in 20% of the cases, which is good enough for the considered application*.

* The data calculated by the algorithm is published on data.gouv.fr. It is also published in map form on cadastre.opensolarmap.org

The aim is twofold: on the one hand, to be capable of guiding ministries in their use of artificial intelligence technology, assessing existing tools and programmes, and to conduct projects related to representative usages. On the other, remain vigilant on matters of ethics and implementation responsibility in the implementation of these processes, even more so when dealing with learning systems that are not always fully comprehensible.

Defining the preconditions for ethical and responsible use

The issues of responsibility, transparency and consistency between the law and information technology (code is law) have already been addressed in 2017 with regard to the post-baccalaureate admissions system and its replacement, Parcoursup. The Digital Republic Act has introduced

provisions relating to the transparency of the algorithms and the opening of the source codes.

In 2018, we will be defining, within the framework of France's commitments in the Open Government Partnership, **the preconditions for the ethical and responsible use of algorithms** (machine learning or "conventional" use) in public policy.

5. Supporting the ecosystem of public data users

The public data users' ecosystem is dynamic and rich. There are over **185,000 single visits per month** to the data.gouv.fr platform. An even greater number of businesses, associations and individuals have access to services which are only available thanks to openly accessible data and, in certain cases, via controlled access (Enterprise API and Individual API).

Supporting this ecosystem is one of the prerequisite for a full exploitation of the data's potential. For example, this will require the participation to, and the organisation of, public events, as well as the acknowledgement of the most striking initiatives emanating from the public, voluntary or private sector, and even providing financial backing for some.

In 2018, the Chief Data Officer will endeavour **to document the social and economic impact** of improved data flow, following up on the first CDO report which provided an analysis of the mechanisms of creating value through data.

GLOSSARY

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

CDO: the Chief Data Officer. The position was created by decree of the Prime Minister on September 16, 2014. The CDO coordinates the activities of the administrations with regard to inventories, governance, production, circulation and use of data by the administrations.

Anonymization: the anonymization of data is a technique for modifying its structure so as to make it very difficult or impossible to "re-identify" the natural or legal persons or entities concerned (source: Wikipedia).

API: Applications Programming Interface allows a software to provide services or data to another, simply. For example, the geocoding API located on data.gouv.fr can transform a postal address into geographical coordinates (latitude and longitude).

Big data: Big data refers both to data with certain characteristics — voluminous or diverse — and also by extension, to its several possible applications.

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

Reference data: reference data is data that is frequently used by several stakeholders from the public and private sectors. Its quality and availability are critical to its uses, e.g. geographical data in state registers.

Key data: is data that links several datasets, e.g. a company's SIRET number.

Data governance: a set of principles and practices designed to harness the data's full potential.

Register: in an administration, a register is a book in which administrative information is recorded, e.g. the register of trade and companies administered by court registries (source: Wikipedia).

Machine learning: originating from Artificial Intelligence, machine learning is a set of techniques allowing algorithms to "learn", e.g. they improve and upgrade themselves as they process new data.

Data has become a crucial tool for the transformation of public action, and more generally, of the economy. Guaranteeing its quality and dissemination is necessary to ensure its best exploitation. Just as the transportation, energy and telecommunication networks, a new type of infrastructure has increasingly become essential in our modern society: that of data.

In his second report to the Prime Minister, the Chief Data Officer advocates the construction of this infrastructure on the basis of a comparative analysis of the methods already applied in some European countries. He also suggests some improvements aiming to let the State fully play its part in this new data landscape.

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE
Premier ministre
Direction interministérielle du numérique et du système d'information et de communication de l'État

Direction de l'information légale et administrative