

OPEN DATA GUIDELINE

OPEN DATA GUIDELINE

This Guideline is part of the cooperation project between
São Paulo State Government and the UK Government



SPUK



Improving business environment through transparency in São Paulo State

Melhoria do ambiente de negócios por meio da transparência no Estado de São Paulo

Partners

ceweb.br nic.br cgi.br

SEADE
Fundação Sistema Estadual
de Análise de Dados

Fundap




Embaixada Britânica
Brasília

 **GOVERNO DO ESTADO
SÃO PAULO**
Secretaria de Governo

Year 2015



This content is licensed under Creative Commons.
Attribution-NonCommercial-NoDerivs
CC BY-NC-ND



SPUK



Improving business environment through transparency in São Paulo State

Melhoria do ambiente de negócios por meio da transparência no Estado de São Paulo

EXECUTION

GOVERNMENT OF THE STATE OF SÃO PAULO

Secretariat of Government

- Sub-secretariat for Partnerships and Innovation

Chief of Staff

- Special Advisor for Foreign Affairs

Foundation for Administrative Development - Fundap

State System for Data Analysis Foundation - Seade

Public Administration Transparency Board

THE UK GOVERNMENT

British Embassy in Brasília

THE BRAZILIAN NETWORK INFORMATION CENTER - NIC.br

Web Technologies Study Center - CeWeb.br

AUTHOR

Marco Túlio Pires

COORDINATION

General:

Roberto Agune - iGovSP

Vagner Diniz – CeWeb.br

Executive and Editorial:

Caroline Burle dos Santos Guimarães - CeWeb.br

Helena Pchevuzinske - iGovSP

Sergio Pinto Bolliger - iGovSP

IDEALIZATION

Alvaro Gregório - iGovSP

DESIGN

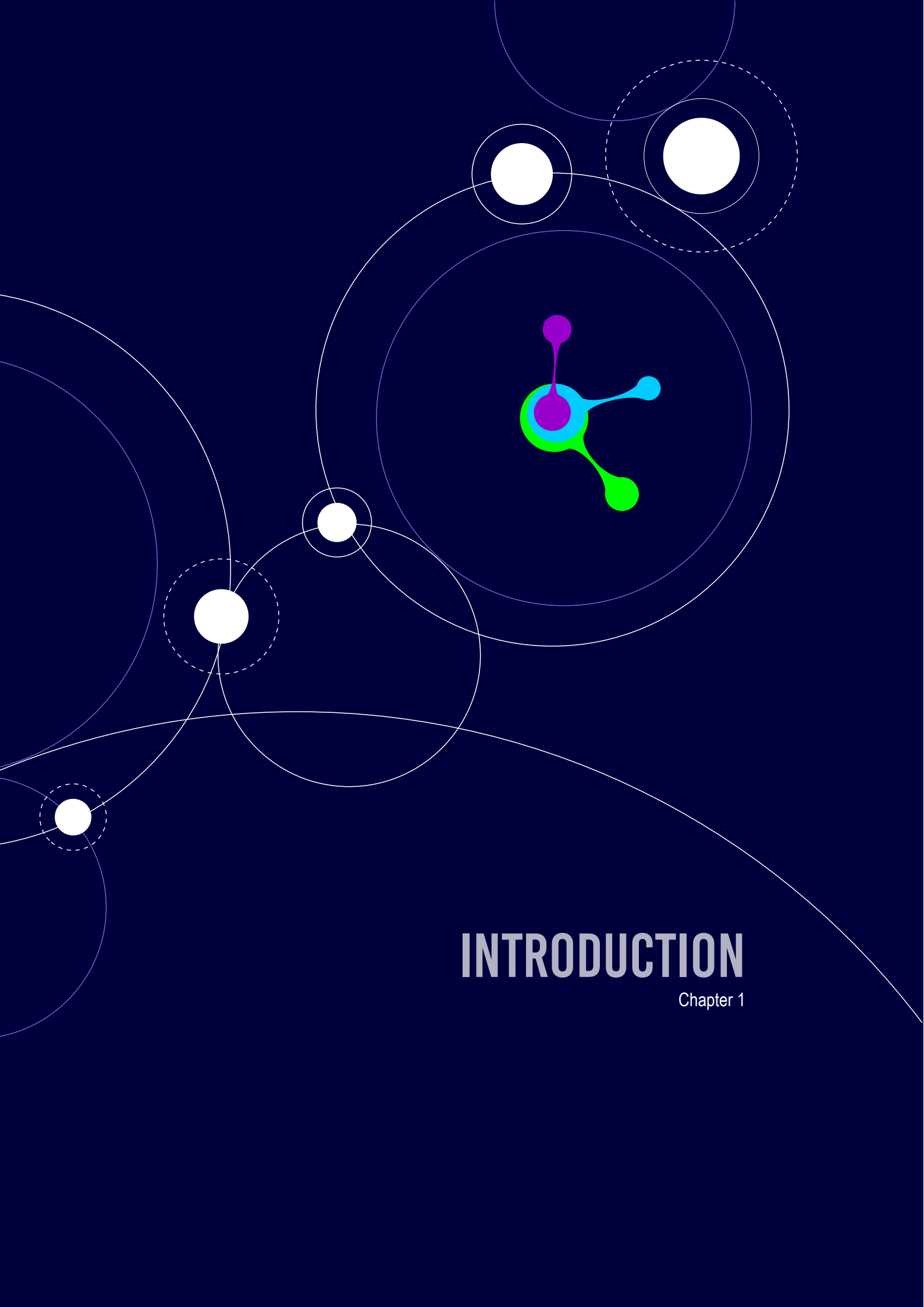
Alcione de Godoy - iGovSP - e-books

Deblyn Pereira Prado - NIC.br - HTML

Ricardo Hurmus - Bulerías Arte & Design - ilustrações

INTRODUCTION.....	06
WHAT IS “OPEN DATA”?.....	09
Can any data be opened?.....	11
BENEFITS OF OPENING UP DATA.....	12
Public Transport (Private Sector Initiative).....	14
Legislative and Executive Participation (Third Sector).....	15
Education and Research (Third Sector).....	18
Health and Public Spending (Government, Citizens).....	19
BRAZILIAN INFORMATION ACT.....	20
Access to information in São Paulo.....	21
Active Transparency.....	22
What are the exceptions to making data open?.....	23
OPEN DATA PLAN.....	25
The five stars of open data: ★★★★★.....	26
★.....	26
★★.....	27
★★★.....	28
★★★★.....	29
★★★★★.....	29
Open Data Teams.....	31
Publishing.....	34
Do you need to design an API?.....	35
Map of technological decisions.....	38
DATACATALOG/REPOSITORY.....	40
TECHNICAL SCENARIOS, TECHNOLOGICAL OPTIONS.....	43

Level 1.....	44
Level 2.....	46
Level 3.....	47
DATA USE LICENSE.....	51
Terms of use of SP Open Government.....	53
FORMATS OF THE DATABASES.....	55
Delimiter-separated formats (CSV).....	57
XMLformat.....	58
KML format.....	61
JSON format.....	62
geoJSON/topoJSON.....	64
SQL format (dump).....	64
Shapefile format.....	65
REFERENCES.....	67



INTRODUCTION

Chapter 1

This Open Data Guide was prepared to expand and contribute to the transparency policy of the State of São Paulo. Contained here is information about the benefits of an open data policy, its challenges and technical characteristics, and a series of recommendations based on international standards and successful experiences around the world in order to make the opening of databases an inspiring and beneficial process.

This guide is not intended to be an exhaustive resource, or purely technical, or an end in itself. It was developed to be housed on the Web and uses web references. Much of the information contained in the following sections can, and should, be supplemented by browsing through the reference links and complementary readings. The information gathered here seeks to take into consideration in the best possible way the complex ecosystem of Public Administration, with its challenges and differences, whether in the size of the teams or in the state of infrastructure in government agencies.

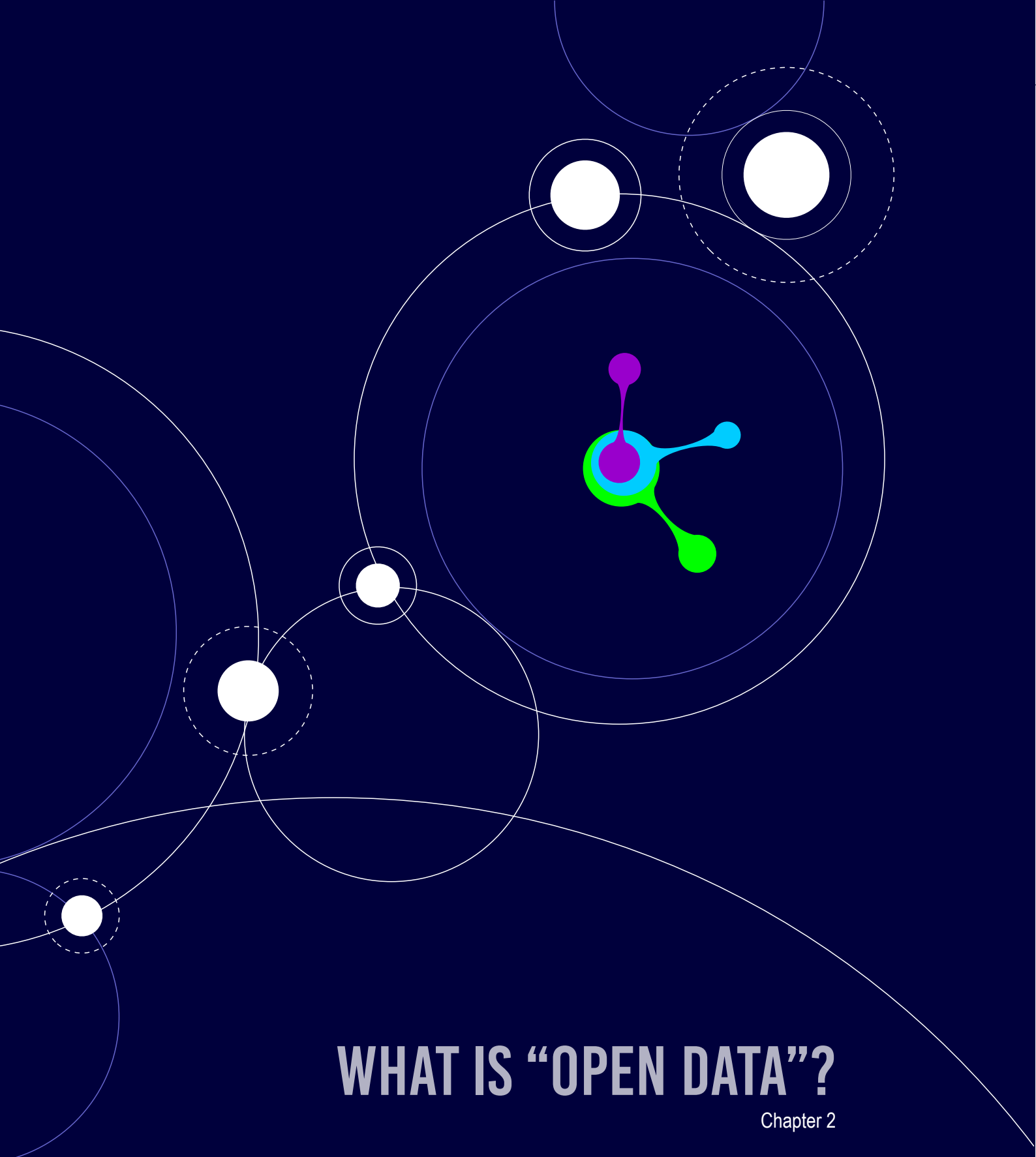
This is a guide that aims to allow technicians, public employees, managers and executives to find out about a movement that is taking root in democracies around the world and is becoming increasingly popular in Brazil. Opening databases presents itself as a point of no return in the context of more transparent governments and fairer societies. It is only a small part of this movement, which also includes “classic” resources for transparency, such as the Brazilian Information Act, as well as relatively new areas such as social participation and control.

An open database policy needs to have short-, medium-, and long-term narratives: Where are we? Where do we want to go? Who will directly benefit from open data? Citizens? Journalists? Public employees? Companies? Scientists? Infomediaries? How are they reached? These are some of the questions that need clear, objective answers within the organization and planning of any effort to open data.

Governments around the world are already seeing good results from opening their databases. One of the most emblematic examples is the British government. Because of open data, it was discovered that several IT departments in public administration were buying the same consulting services from the same company. By analyzing the data, administration realized it could save six million pounds by decreasing the number of contracted consulting hours. This is equivalent to the entire financial resources required to fund the transparency program of the British government. Opening up the

data also led to the emergence of a number of companies and services. Some of these cases are included in this guide.

Examples such as these, which represent the best efforts of open data, are waiting to be discovered in São Paulo and Brazil.



WHAT IS “OPEN DATA”?

Chapter 2

“Open data” is a term that has gained popularity in the transparency and open government movement around the world, but it is not always clearly defined. Opening up data follows the same principle as Open Government: Treat access to public information as a rule, not an exception. In this guide, “data” refers to information generated by all public agencies in the state of São Paulo that come from administrative activities of government management: contracts, functions, projects, policies, and partnerships with other sectors. In short, all the data that is in the custody of the state government or entities linked to it.

Opening this data means that government information may be freely used, reused, and redistributed by anyone without any restrictions. This requires at the most that the database’s source be provided and that this information be redistributed under the same conditions or licenses in which they were originally acquired.

For a data set to be considered “open,” it must present at least the three following characteristics:

Availability and access: The data must be available as a whole and in a way that does not create complicated processes for the interested party to copy it. The best-case scenario is to provide for the data to be downloaded over the Internet. The data also needs to be available in a convenient and modifiable format.

Reuse and Redistribution: The data must be provided under terms that permit reuse and redistribution, including combining this data with other databases.

Universal Participation: Everyone must be able to use, reuse, and redistribute the data. There should be no discrimination against fields of endeavor, persons or groups. For example, “non-commercial” restrictions that would prevent “commercial” use of the data, or restrictions on use for certain purposes (e.g., only for personal research), are not allowed.

These three characteristics are summarized in three “laws” suggested by the open data activist David Eaves:

1. If the data cannot be spidered or indexed, it does not exist.
2. If the data is not available in open and machine readable format, it cannot engage.

3. If a legal framework does not allow it to be repurposed, it does not empower.

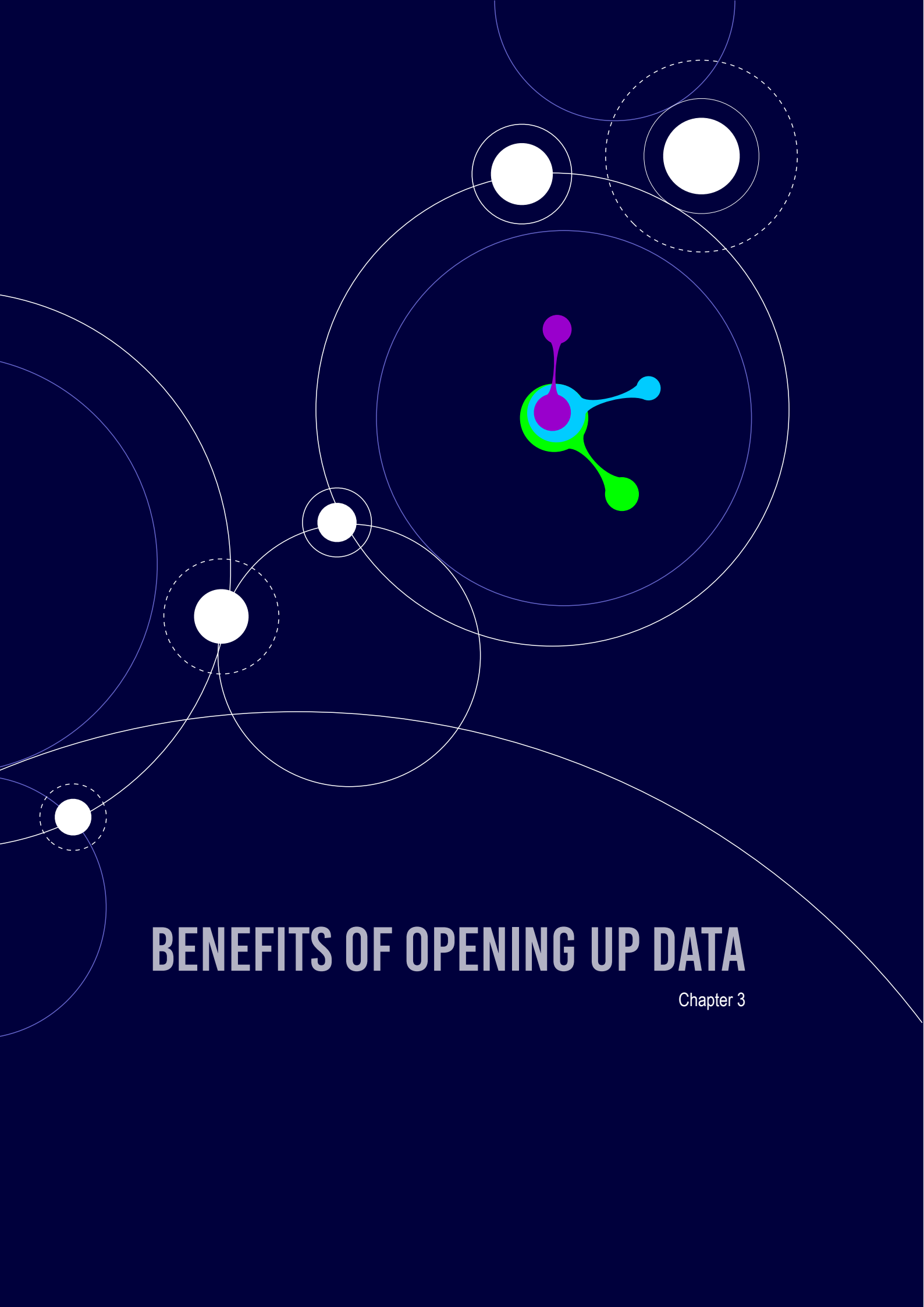
There are many reasons to be very clear about the meaning of “open data.” Since the term has such a broad meaning, it is necessary to define exactly what characteristics are considered ideal in the scope of this guide, so that the information published by the government can be used by all sectors of society in a compatible manner.

Opening up databases with the characteristics described above is important because it makes interoperability possible; this exists when different actors in society work with different databases together. Building better and better systems and solutions, whether developed in the spheres of government, private, academic, or civil society, depends on the interoperability of databases.

Clarity about what “open data” is also ensures that two or more databases coming from different sources can be combined without major technical obstacles. Among other things, this avoids having government be a large warehouse of “closed” databases that serve only for utilization by individuals and are useless for larger, complex systems that are capable of providing solutions, views, services, or value to citizens or groups in society.

Can any data be opened?

No. All public data should be open, but not all data is public. Brazilian legislation does not allow opening up private data that identifies individuals, violates their right to privacy or honor, or discloses confidential data or data that may compromise national security.



BENEFITS OF OPENING UP DATA

Chapter 3

Opening up government databases can bring a series of benefits to different sectors of society by creating a cycle of mutual benefit. One of the first beneficiaries is the government agency that decides to open its databases. The simple structuring and execution of organizational mechanisms that allow these databases to be systematically published and opened can contribute to a significant increase in the quality of teams and services, as well as knowledge of internal bottlenecks and obstacles. It is not possible to manage or know that which you cannot measure. Opening government data enables you to know and measure these activities.

Open data may also contribute to the advancement of science. Opening government data enables independent researchers or those associated with institutions to access a valuable layer of information about the formulation and implementation of public policies, how society resources are directed, and the impact of programs on improving people's quality of life. Well-designed research contributes to raising the level of knowledge that a given society has about itself and even about other societies in international research, as well as providing conditions for its actors to make increasingly justifiable decisions.

It is also beneficial to the private sector. Responsible and consistent opening of government data allows entrepreneurs or groups of entrepreneurs to use their creativity to build tools, solutions, and technological advances that often escape the notice of expert teams that work with this data within the governmental structure. This means that opening up data can contribute to generating jobs and wealth, creating healthy interdependence between the government and the private sector.

Civil society organizations can also take advantage of open data. Opening up databases can raise the quality of services provided by non-governmental organizations that are often complementary to those offered by the state. The government has a plethora of information about the public and services of interest to these organizations. Free and unrestricted access to this data enables these organizations to put their services to the test and measure their results, continually increasing the quality of their activities.

Finally, everyone can benefit from open databases by having facilitated, free access to data generated in the governmental sector. Society as a whole benefits. At the individual level, the benefit is exercising the fundamental human rights of freedom and access to information. Open data, which is part of open and transparent government, is one of the pillars that supports

societies that want to be more free and just. It contributes to strengthening democratic processes, opening paths for greater citizen participation in public administration, and promoting social control of government activities. Access to government data allows anyone to monitor the implementation of public policies and measure their effectiveness. The availability of tools that allow citizens to know what actions the government is taking strengthens the state's legitimacy, enhancing its role as a key player in the pursuit of social welfare.

Successful initiatives in this area that serve as inspiring examples for opening data are not lacking around the world, including in Brazil. Discover some of them below.

PUBLIC TRANSPORT (PRIVATE SECTOR INITIATIVE)

- [Citymapper](#)

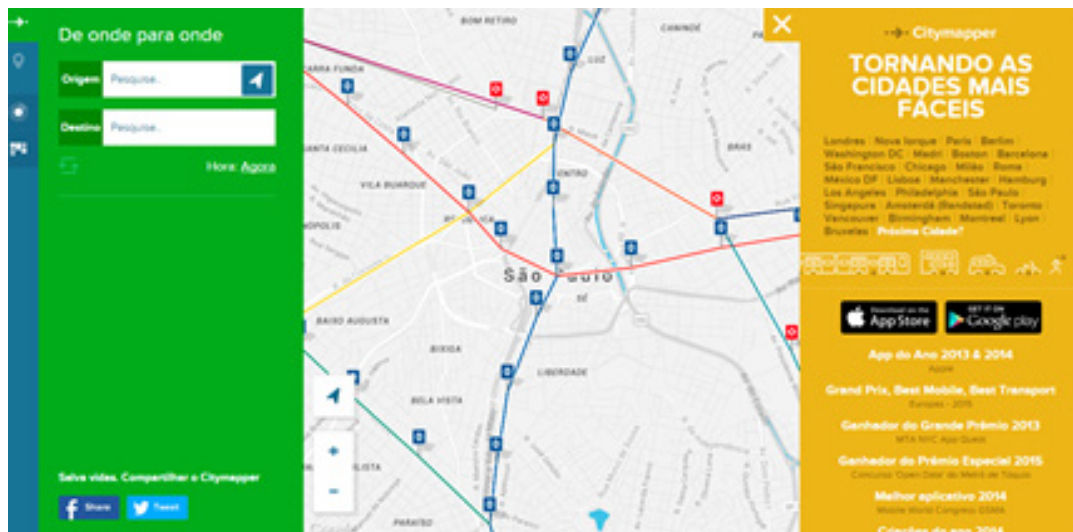


The Citymapper is a good example of the potential of open data to foster the generation of new business. It is a free smartphone application created by a former Google employee in 2011 in London with the objective of improving the experience of millions of people in the city who depend on public transportation every day.

The developers of Citymapper took advantage of data about buses, trains, and subways published in real time by the Transport for London to formulate an algorithm that always shows the best travel time from one point to another, when the next ride will be, and how many calories you spent walking. The information is updated in real time and show traffic conditions, weather, and technical problems of vehicles.

The plan worked and the company expanded its services to other cities. From 2011 to 2015, it reached seven countries and 13 cities: London, New York, Paris, Berlin, Washington DC, Madrid, Boston, Barcelona, San Francisco,

Chicago, Milan, Rome, and Mexico City. The application can be connected to any city that offers real-time data on public transport services.



LEGISLATIVE AND EXECUTIVE PARTICIPATION (THIRD SECTOR)

- [Sunlight Foundation](#)



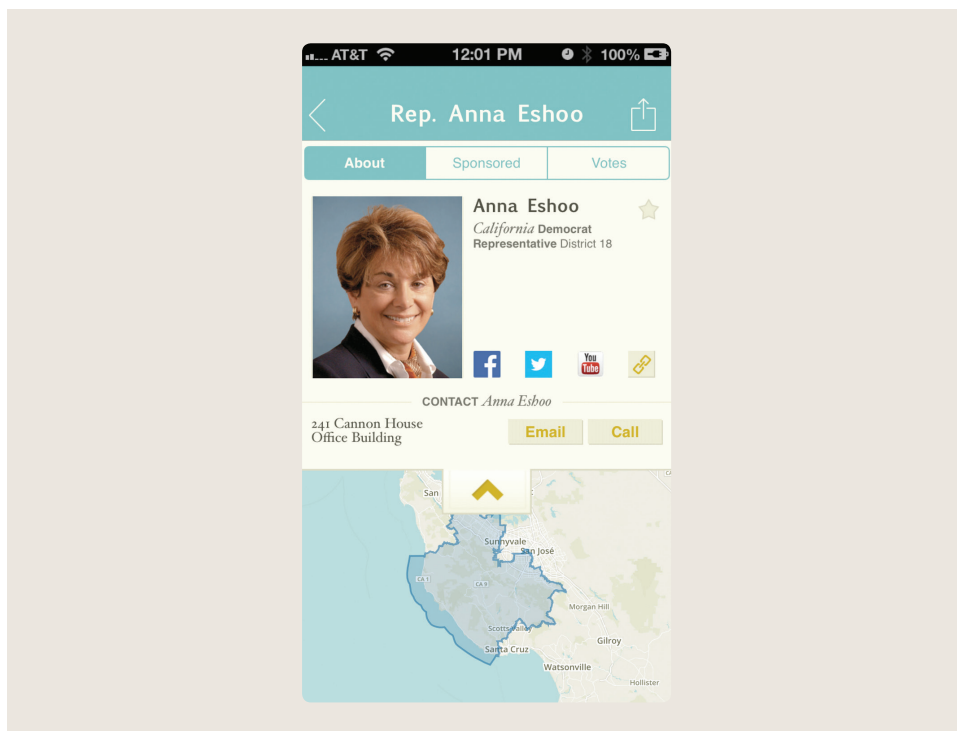
Sunlight Foundation is a successful example of how the third sector can work with the government to encourage popular participation and social control by using technology. Created in April 2006 in the United States, the foundation's objective is to increase the transparency and accountability of America's legislative and executive levels. Its initial focus was to monitor the role of money in American politics and propose changes aiming toward open government.

Today the foundation also contributes with international forums that discuss government transparency. Sunlight has three main fronts: a political team that coordinates actions within and outside the US Congress to propose changes in legislation; a team of journalists specializing in data who generate and publish reports on the state of American transparency; and a technology laboratory

that develops tools from open data aiming to increase the participation of citizens in government and enable other developers to build applications with public information.

The Sunlight Foundation carries out the systematic organization of various US government databases. Because the databases are sure to be published consistently and regularly in an open format, this has allowed the Foundation to develop a number of tools that bring citizens closer to public administration:

- Congress



This is a smartphone application that displays information about congressional representatives and senators, allowing citizens to contact them and track their activities. You can read the latest laws passed, view the list of activities in the House, navigate through votes, and stay on top of upcoming committee hearings and their audiences. The application was developed using databases

opened by the US government and is freely distributed. Its source code is open and anyone can download it.

- [Influence Explorer](#)

INFLUENCE EXPLORER

This is a website that tracks political donations at the state and federal level, allowing anyone to track the level of influence per politician, business, or individual. The tool provides an overview of campaign finance, [lobbying](#), parliamentary funds, irregularities in hiring, and federal government spending.

- [Capitol Words](#)

capitolwords

a project of the Sunlight Foundation

This tool lets you explore the contents of the speeches of all US senators and congressional representatives since 1996. You can search by state, date, or legislator. The service allows you to compare terms and phrases, displaying the results with graphs, and rankings of the legislators and their parties. Capitol Words also serves as a large database so that other developers can create applications that depend on it.

Take a look at [other tools](#) (English website) developed by the Sunlight Foundation.

Among the companies and organizations that use Sunlight Foundation services are [Wikimedia Foundation](#), which manages [Wikipédia](#); [Greenpeace](#); the [If This Then That](#), quesite that lets you create, among other things, instant cellphone notifications on the progress of laws in the US Congress; and [Barack Obama](#) campaign team.

EDUCATION AND RESEARCH (THIRD SECTOR)

- [QEdu](#)



QEdu is a good Brazilian example demonstrating the benefits of the third sector, academia, and government working together to better understand basic education in Brazil and provide subsidies for public policies. This is a free portal developed in partnership with the [Lemann Foundation](#) and [Meritt](#) that allows anyone to obtain information about the quality of learning in Brazilian schools with data on public and private schools. The views are generated using databases from the Brazilian government, such as [data from the National High School Exam](#) (Enem), [Prova Brasil](#) (“Test of Brazil”), School Census, and special indicators of the [National Institute of Studies and Research](#) (INEP).

The tool shows how well 5th and 9th grade students learned mathematics and Portuguese; the profiles of students, teachers, and school principals who took the Test of Brazil; enrollments for each stage in school; pass, dropout, and failure rates, socioeconomic levels; school infrastructure; age-grade distortion; and more. You can search and compare by schools, cities, or states. The portal also allows anyone to filter and download data need in an open format.



HEALTH AND PUBLIC SPENDING (GOVERNMENT, CITIZENS)

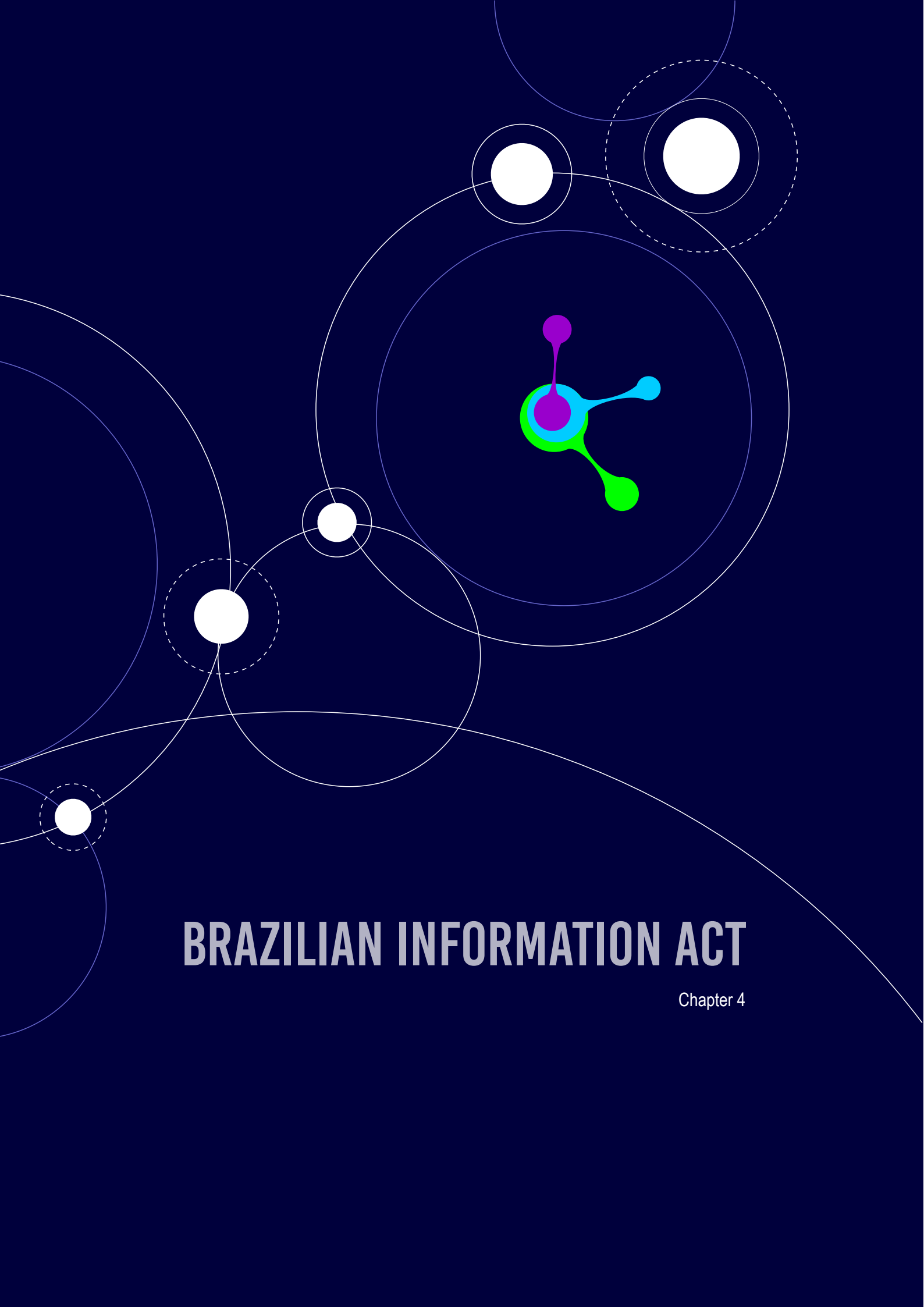
- [Prescribing Analytics](#)



Prescribing Analytics shows how government officials and independent professionals can work together to find ways to save public money. It is a data analysis tool about British government expenditure on a specific group of drugs called statins. This drug helps to combat high cholesterol in patients with health problems.

The service was created by a group of independent developers and physicians who work in the UK's public sector. The British health system was given the task of saving [20 billion pounds](#) by 2015, being that half of this is spent on medications. One of the substances most prescribed by physicians in the UK's public health system is statins. The tool accesses the open database of physicians' prescriptions in the country and analyzes which types of statins were prescribed by physicians in the public system: more expensive variants or usually cheaper "generics."

The data showed that between September 2011 and May 2012, the British government could have saved [22.96 million pounds](#) per month if the physicians had prescribed generic statins that have the same efficacy as the more expensive ones.



BRAZILIAN INFORMATION ACT

Chapter 4

There is great motivation for opening up government databases in Brazilian and international law. Brazil is a signatory to several international agreements that treat access to information as a right of every citizen, including the Universal Declaration of Human Rights, adopted by the United Nations General Assembly in 1948. The right to information (below in bold) is provided for in Article 19 of the declaration:

“Everyone has the right to freedom of opinion and expression; this right includes freedom to hold opinions without interference and to seek, receive, and impart information and ideas through any media and regardless of frontiers.”

It is understood that any information that is under the state’s custody is public, while respecting certain restrictions. This is the spirit of most of the access to information laws around the world, and it is no different in Brazil. As a rule, government databases are considered open. The cases in which opening them is not considered suitable are exceptions. That is why it is recommended that you familiarize yourself with the [Brazilian Access to Information Law](#) and its [regulations in the State of São Paulo](#) to promote opening up governmental databases consistently and in accordance with legal dictates.

ACCESS TO INFORMATION IN SÃO PAULO

The government of São Paulo has been preparing since the mid-1980s to organize the information within the state. That has included instituting the File System of the State of São Paulo and a number of laws and decrees during the 1990s and 2000s that deal with the ecosystem of responsibilities and data management in São Paulo. One of the most important steps was taken in 2010 with a decree that dealt specifically with how the state should publish and organize the data considered open on the World Wide Web, the Internet.

[Decree No. 55.559 of March 12, 2010](#) created the open data portal of São Paulo entitled “[SP Open Government](#)”. This portal concentrates efforts to

¹Leis estaduais nº 10.177, de 30 de dezembro de 1998, que regula o processo administrativo, nº 10.294, de 20 de abril de 1999, que dispõe sobre proteção e defesa do usuário de serviços públicos; decretos estaduais nº 22.789, de 19 de outubro de 1984, que institui o Sistema de Arquivos do Estado de São Paulo - SAESP, nº 44.074, de 1º de julho de 1999, que regulamenta a composição e estabelece a competência das Ouvidorias, nº 54.276, de 27 de abril de 2009, que reorganiza a Unidade do Arquivo Público do Estado, da Casa Civil, nº 55.479, de 25 de fevereiro de 2010, que institui na Casa Civil o Comitê Gestor do Sistema Informatizado Unificado de Gestão Arquivística de Documentos e Informações - SPdoc, alterado pelo de nº 56.260 de 6 de outubro de 2010.

publish non-confidential databases and provide unrestricted access to the São Paulo government. The decree provides that these databases shall be published in an “open format” and makes due consideration in relation to confidential data that should not be published. According to the decree, each public agency is responsible for selecting, publishing, and updating these databases under the coordination of the Secretariat of Public Management, which is responsible for maintaining the portal. Another decree that deals with access to information in the State of São Paulo is [No. 58.052 of May 16, 2012](#) that regulates Federal Law No. 12.527 of November 18, 2011, which is the current Brazilian [Information Act in force in Brazil](#).

ACTIVE TRANSPARENCY

State [decree No. 58.052 of May 16, 2012](#), which regulates the federal Brazilian Information Act in the state of São Paulo, incorporates in Article 23 the principle of active transparency, and outlines the following requirements:

Article 23 - It is the duty of the agencies and entities of the State Public Administration to promote, regardless of requirements, the disclosure, in an easily accessed location, of information that includes their competencies, documents, data, and information of collective or general interest produced by them or in their custody.

§ 2º - In compliance with the heading of this article, the state agencies and entities shall use all legitimate means and tools available to them, and the information must be published in official sites on the World Wide Web (Internet).

The publication of information on the Internet, according to the state decree, must meet certain criteria, such as offering a content search tool that allows access to information that is objective, transparent, clear, and in easy-to-understand language. The agencies must also allow the reports to be saved in various electronic formats, including open and non-proprietary formats such as text spreadsheets, in order to facilitate the analysis of the information. It is important to note that although the text of the decree does not establish in detail what an “open and non-proprietary” electronic format is, the characteristics of such formats are already well established. The Open Data Guide has a dedicated section describing a number of open electronic formats,

most of them non-proprietary.

Regarding active publication of information on the Internet, the decree states that the agencies should provide the information in a manner that enables automated access by external systems in open, structured, and machine-readable formats. A [PDF](#) file, for example, it is not sufficient to meet what the decree requires. PDF is not considered an open technology under the terms of this guide because it is not structured, much less machine-readable. Its function is to generate documents that will be printed exactly as they appear on the computer screen. It is not a technology that aims to facilitate automated access by external systems. The format section of this Open Data Guide presents a number of technologies that better fit the requirements of the state decree.

The decree also provides that the agencies shall disclose in detail the formats used for structuring the published information, as well as ensure its authenticity and integrity and keep it up-to-date. The agencies should also take all necessary measures to ensure that the content is accessible to people with disabilities.

WHAT ARE THE EXCEPTIONS TO MAKING DATA OPEN?

Free and unrestricted access is considered a rule in the [Brazilian Information Act](#). There are a few exceptions, and all of them are described in the text of the law and state decree. They are the following:

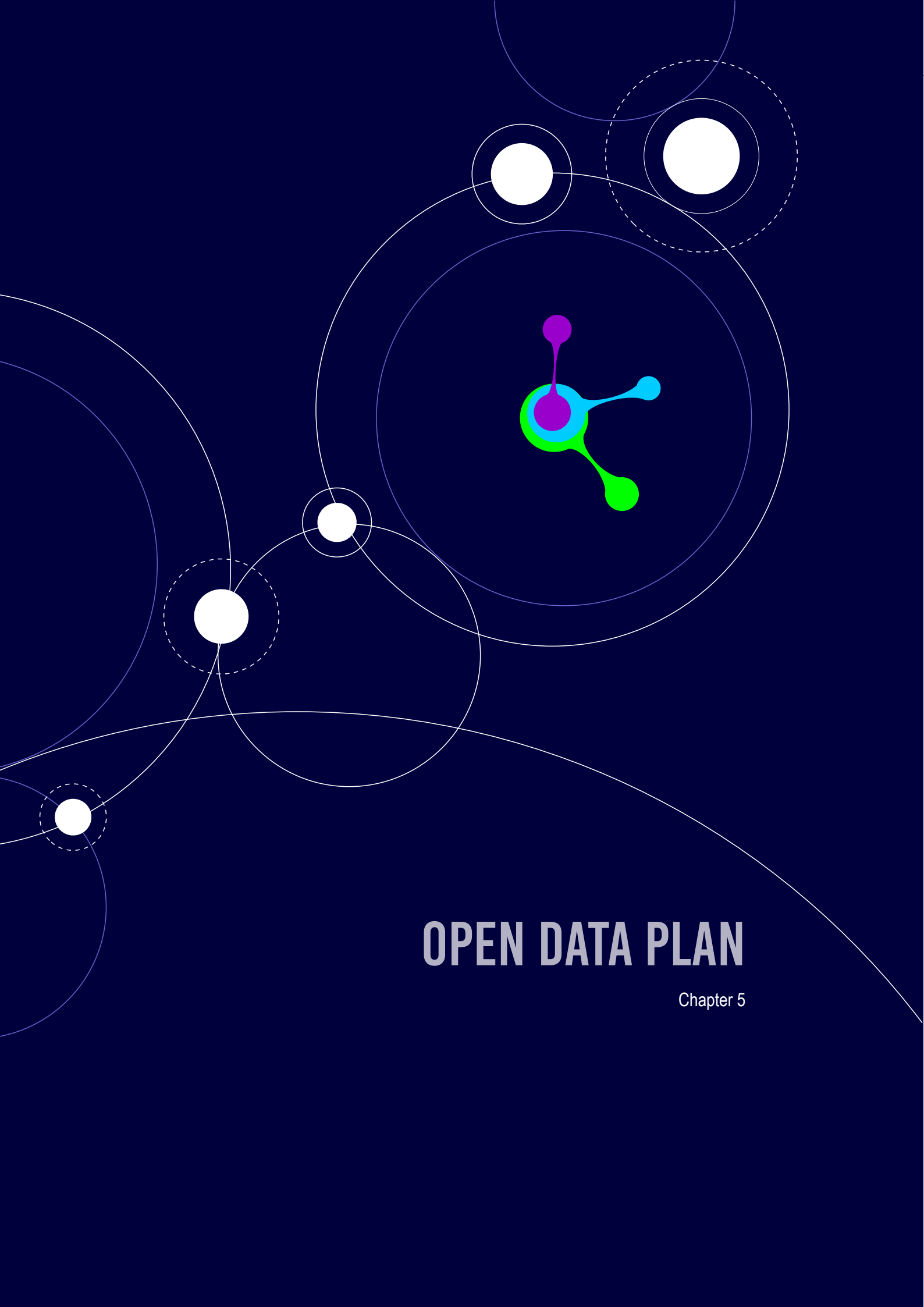
Safety of society and the State

- Information for which confidentiality is considered essential for the safety of society and the State are considered as exceptions, as well as scientific and technological research for this purpose, and State secrets. It is data that, if disclosed, might:
- Endanger national defense and sovereignty or the integrity of national territory.
- Harm Brazil's international relations or disclose information of a confidential nature that has been provided to the government by other countries.

- Jeopardize the life, safety, or health of the population.
- Create a high risk of financial, economic, or monetary instability in Brazil.
- Put at risk the safety of foreign authorities in the country.
- Compromise ongoing research or inspection activities related to prevention or prosecution of offenses.

Personal information

This is information related to an identified or identifiable natural person concerning intimacy, privacy, honor, and image, as well as individual freedoms and guarantees. Telephone numbers, identity cards, driver's licenses, taxpayer numbers, bank statements, etc. are information that can identify people and is therefore confidential.



OPEN DATA PLAN

Chapter 5

Opening data in any public agency is a process that should be planned carefully and with the involvement of various actors: keepers of the databases, Information Technology (IT) professionals, public managers, and technicians in the areas of interest. As a rule, all databases that the agency already owns can be published. However, we must ensure that the databases selected for publication do not fit into the exception scenarios provided for in the Brazilian Information Act. If these databases contain any confidential information, that data must be hidden before it is published.

This section of the guide suggests creating an “Open Data Plan.” This is a strategy for starting the opening data processes or assessing the open data situation at a given agency. However, each government institution is free to decide how to open their databases in accordance with state and federal legislation on access to information.

THE FIVE STARS OF OPEN DATA ★★★★★

In 2010, the British scientist Tim Berners-Lee, inventor of the World Wide Web, proposed a [star rating](#) system to encourage society, especially government data custodians, to open their data. The system helps to diagnose the level of open data in public agencies and provides achievable steps to reach increasingly higher levels of open data.



The first stage of the five-star system only requires that your data be available on the Web in any format (a PDF document, a Word file, or any other type, proprietary or not, open or not). In addition to being accessible on the Web, the data should be provided under an open license. An open license gives the authorization for this data to be used by anyone without restriction and for any purpose, including commercial. Several open licenses describe the use of data. This guide is not an exhaustive analysis of existing open licenses, but it gives guidelines so that each agency can formulate open licenses that are compatible with international principles of open data.

In order to meet the requirements of the first star, a file just needs to be posted on the Web in any format with an open use license specified. If

your agency already publishes data on the Web in any format, this means that you just need to add an open license to be awarded the first star.

WHAT ARE THE BENEFITS OF ★

The first star brings a number of benefits for the consumer of the data as well as for the publisher. Consumers can view and print data or download data to a personal computer or USB flash drive, enter the data into any other system, perform analysis and share it with anyone they want in the way they want.

For the publisher, it is simple because there is no need to stick to the formats used. In addition, it will not be necessary to explain repeatedly to all those interested in the data that it can be used for any purpose, as long as it is distributed using an open license.

Publishing data on the Web in any format under an open license is a major step in opening data. However, data consumers might find it difficult to extract the data from these documents without having to manually type them into other systems.



The second stage of the five-star system requires that the information be published on the Web under an open license, but it also requires that these databases be made available in a structured format that allows easy handling of its rows and columns.

In other words, while the first star allows scanned images of tables and PDF files of reports to be published, in order to obtain the second star it is necessary that the file be made available in a manner that allows consumers to use data analysis and structuring applications without needing to enter the information manually. Excel files (.xlsx), for example, are structured files.

The second star does not have any requirements in regard to the application that generated the file or document format, provided it is structured.

WHAT ARE THE BENEFITS OF ★★

In addition to all the benefits of the first star, two stars allow consumers to use proprietary applications, such as Excel, to aggregate, perform calculations, visualize, and carry out other operations with the available data. Furthermore, the data can be exported into another structured format with ease.

The publisher can still publish data simply, just paying attention to the fact that the published file must be in a structured format, and respecting that it must be on the Web and distributed under an open license.

It may not seem like much, but two stars represent a major breakthrough in open data, since the information is available on the Web in a structured format and under an open license. However, the data is still locked up in a document. To extract the data, depending on the format used, consumers must use proprietary software (which can cost money), which is an obstacle for many people.



The third stage of the five-star system is similar to the second, but includes another requirement: The structured files available on the Web under an open license must be created in a non-proprietary format. Instead of a document in .xlsx format created by Excel proprietary software, for example, the option would be to use CSV, a structured format that does not depend on manipulation of proprietary software.

WHAT ARE THE BENEFITS OF ★★★

Agencies that reach three-star open data will offer citizens all the benefits of two stars, as well as allow anyone to download and manipulate the data in the most convenient way without requiring a specific application.

The publisher should pay attention to the necessary converters and plugins to export the data from the proprietary format into an open format. The entire publishing process remains fairly simple, since it only means providing documents in an open format on the Web, distributed with an open license.

With three stars, anyone can easily use the data, but this information can be made available in formats that allow even more interaction between systems and facilitate sharing.



To better understand how the fourth and fifth stars work, it is important to read the Semantic Web Guide. While the first three stars give guidelines for publishing data on the Web using open document formats and under an open license, the fourth and the fifth stars introduce the concept of linked data link for reference to the Semantic Web Guide

To achieve four stars, you must do all that the previous stars advise, but in addition to using documents in an open format, the data must also be published on the web page itself using URIs to describe each piece of the data, so that anyone can point to the elements in a standardized format in the published document.

WHAT ARE THE BENEFITS OF

Data published following the four-star guidelines can be linked to from systems available anywhere else on the Web. Consumers can reuse parts of the data and combine it with other data.

The publisher begins to have fine control of each cell in the database and can optimize database access, load balancing, caching, and more. Other agencies that publish data can link their databases to yours using the same URI scheme.



O último estágio do sistema de estrelas requer que seus dados publicados no esquema de URIs estejam conectados a outras bases de dados publicadas sob as mesmas condições. Mais informações sobre como os dados conectados funcionam pode ser encontradas no Guia de Web Semântica.

WHAT ARE THE BENEFITS OF ★★★★★

Data published following five-star guidelines allows anyone to discover more data as they browse through it. Consumers can also directly learn about the data publishing scheme by just studying its structure.

The publisher allows the data to be discovered, increasing its value. The publishing agency will gain these same benefits, since the resources will be available for anyone.

The five stars of open data serve as a serious and objective guideline in relation to the objectives of each public agency for opening its databases. Summarizing:



Publish databases on the Web (whatever format) under an open license



Publish databases in a structured format under an open license (e.g., Excel file instead of a scanned image)



Use non-proprietary formats and an open license (e.g., CSV file instead of Excel)



Use URIs to denote things, so that anyone can point to them



Link your data to other databases to provide context

In view of the recommendations of the system developed by Tim Berners-Lee, the “Open Data Plan” can be developed taking into account the following:

Scope

When preparing to open its databases, the agency must choose which databases, or which parts of them, should be opened. If deployment is guided by the five star scheme for open data, the technical and human challenges to achieving the various star levels should be taken into account.

Maybe some of the databases are already ready to be published on the Internet in a structured and open format under an open license (three stars); others might only be in closed and not structured formats but can be placed on the website under an open license (one star).

The scope will help map which databases can be opened and which star category they will fall under. The number of stars will depend on the technical and human resources available to the agency.

Prioritization

After defining the open data scope in relation to the number of stars you want each of the databases to meet, the team responsible for opening the databases must ask some questions in order to establish publishing priority for these databases.

- Which databases can be published immediately? With how many stars?
- Which databases will need to have some work done before they are published?
- Which databases will be published in the long run?
- What is the minimum star category that the agency wants all its published databases to meet?

The interested parties

You should also consider which areas of the agency (boards, coordinating committees, departments, etc.) will be part of the opening data process, considering the number of stars you want the databases to achieve. Each area should be responsible for providing the data it produces in formats that conform to the number of stars defined in the scope of the open plan.

OPEN DATA TEAMS

To open the databases of your agency, you need to think about which actors will be directly involved in the workflow of publishing these databases on the Web such that they are always updated and follow a consistent standard of disclosure. The number of people and the profile of the professionals involved will depend on the number of stars that your agency wants the open data

to achieve. Below is information on the minimum required profile of the professionals for each of the stars.

★ Publish databases on the Web (whatever format) under an open license

Staff:

- Technician of the public sector, custodian or person responsible for the database that should be opened.
- IT professional who will be responsible for putting the documents on the website, together with the open license.

Streamlined workflow: The technician of the public sector sends the database (in any format) to the IT professional, who in turn makes the database available on the agency's website with an open license.

★★ Publish databases in a structured format

Staff:

- Technician of the public sector, custodian or person responsible for the database that should be opened, paying attention to the fact that the database should be in a structured format (e.g., Excel file instead of a scanned image).
- IT professional, who will also make sure that the database is in a structured format and will be responsible for putting the documents on the website, together with the open license.

Streamlined workflow: Technician of the public sector, custodian or person responsible for the database that should be opened will, if necessary, convert the database into a structured format, getting help from the IT professional. This professional sends the document, together with its open license, to the website of the public agency.

★★★★ Use non-proprietary formats

Staff:

- Technician of the public sector, custodian or person responsible for the database that should be opened, paying attention to the fact that the database should at least be in a structured and non-proprietary format (e.g., CSV file instead of Excel).
- IT professional, who will also make sure that the database is in a structured format and, if necessary, convert the database into an open format and will be responsible for putting the documents on the website, together with the open license.

Streamlined workflow: Technician of the public sector, custodian or person responsible for the database that should be opened will, if necessary, convert the database into an open structured format, getting help from the IT professional. This professional sends the document, together with its open license, to the website of the public agency.

★★★★ Use URIs to denote things, so that anyone can point to them

Staff:

- Technician of the public sector, custodian or person responsible for the database that should be opened, paying attention to the fact that the database should be in a structured and open format.
- IT professional who will help prepare the database so that it complies with the standards described in the Semantic Web Guide.
- IT professional who will be responsible for maintaining these databases on the Web, together with the open license.

Streamlined workflow: Technician of the public sector, custodian or person responsible for the database that should be opened gets help from the IT professional to make the database comply with the standards described in the Semantic Web Guide. The database is then published (or updated) on the Website with adequate infrastructure to support [Linked data](#).

★★★★★ Link your data to other databases to provide context

The staff and the streamlined workflow for five stars is almost the same as the four stars.

PUBLISHING

With data in hand in an open format and under a license that allows it to be reused freely, now it is time to publish the databases on the Web. This is an important time for the Open Data Plan and should be planned calmly and diligently. Publication depends on how many stars the agency concerned wants to achieve with the open data. Ideally, a publication that aims to achieve five stars in the long run should prepare from the beginning so that adjustments along the way are smooth and predictable.

The best way to achieve openness and interoperability of the data, with regard to feasibility of access on the Web, is using structured and planned repositories. This does not necessarily mean that the agency concerned must acquire complex data storage systems. A simple web page with a well-structured list of documents can serve as a good data repository or catalog, provided certain precautions are taken. The complexity of the system will depend on the number of open databases and the technical and human resources of each agency.

The format in which this information is organized must be agreed on in advance with the participation of all the actors involved. Preferably, this should include agencies that publish correlated data and can publish databases that have the potential to “talk” to each other. This is important so that everyone has a common understanding of the meaning of the data that will be shared.

The objective of this organization to make sure everyone who has access to the information will be able to interpret the data in a uniform way, using data exchange systems and platforms. This prior standardization among the parties takes shape when you publish the names and definitions of the elements used on the Web in a sharable and referenceable format, regardless of the degree of support obtained.

This planning paves the way for developing five-star databases, regardless of the applications used to organize and publish them. Whatever the adopted

strategy, it is important to include [the concepts of a few standards](#) in the planning:

- A URI is a resource identifier that is used to identify or point to something on the Web.
- A URL is a URI that identifies a resource and provides the means of acting upon it, obtaining and/or representing this resource, specifying its primary access mechanism or the “network” location. For example, the URL <http://www.w3c.br/> is a URI that identifies a resource (W3C Brazil Site) and represents this resource (HTML of the page for example) and is available via the HTTP of a network host called <http://www.w3c.br>.
- RDF/XML: XML is a [W3C](#) standard format for creating documents with hierarchically organized data, as is often seen in formatted text documents, vector images, and databases.
- SPARQL: the “sparkle”, also recommended by W3C and under the care of W3C Semantic Web groups, is used to search information independent of the format of the results.

There are standards for publishing data in an open format. It is imperative that these standards also be specified and regulated in norms or any other government recommendations to enable an interoperable environment in all of the e-Gov domains.

DO YOU NEED TO DESIGN AN API?

An important issue to be taken into account when opening databases is designing an API (application programming interface) to provide information on the Web. In the scope of this guide, an API can be summarized as a layer of interaction between a database and an application that feeds on the data. The API provides to interested developers and entrepreneurs a set of standard Web calls for extracting data from a certain database. Designing an API requires refined technical knowledge and, if the database is going to be public, arbitrary standards must be defined, trying to predict the cases in which developers and entrepreneurs will need the data. An API provides a number of advantages, such as easier and faster access to databases. Instead of downloading the entire

database, programmers will only need to make a simple call on the Web to extract the section that interests them at that moment. It also facilitates real-time access to specific parts of the database, allowing the development of applications that rely on rapidly updated data.

An API can be private when a developer has control over the database and creates it to facilitate access to the data, or public when the custodian of a database designs an API to serve a community of developers and entrepreneurs, trying to foresee what kinds of database calls will be useful and generic enough to provide for the greatest number of possible applications. Services such as [Facebook](#) and [Twitter](#) have public APIs that allow programmers from around the world to interact to a limited extent with the immense amount of the data involved.

Despite its advantages, designing an API within the government can bring about uncomfortable situations, depending on the case. It is necessary to carefully think about whether designing an API is the best way to go, since there are alternatives that can better suit both developers interested in government data and teams of public employees or professionals contracted by the state to keep the APIs working stably and reliably.

A HYPOTHETICAL CASE

Imagine that the Department of Logistics and Transport of the State of São Paulo designed a public API so that any developer could access information about the maintenance conditions of São Paulo roads. One day the API server was flooded and the database server crashed. State services that depended on this database stopped working. The logs showed that there was a sharp increase in traffic between eight and nine o'clock in the morning and loads of API calls were made from many different places. After nine o'clock, the server load decreased and everything returned to normal.

² Adaptado de: <https://www.peterkrantz.com/2012/publishing-open-data-api-design/>

WHAT HAPPENED?

Continuing with the fictional scenario, the year before, the Department of Logistics and Transport started to make their data available as part of the state's transparency policy. They were in a rush and, with reduced staff, decided to create an API for the road data by setting up an Internet-facing API server. The API design took into consideration potential use cases that application developers might have, but it was hard to know what people wanted. The staff of the department settled on three generic API calls.

A year after the problem with the servers, the department learned that an entrepreneur had developed a very successful mobile app used by several hundred thousand people. Every morning before the users went to work, the app showed the maintenance situation of the roads in São Paulo. To download this data, each application installed on each mobile device had to make two API calls. That promptly crashed the department's servers because the infrastructure was not designed to cope with the load.

ALTERNATIVE

An alternative to the model presented above is to publish data dumps in files. In this model, data from the database is exported and transformed into an open file format, such as CSV. After that, the files are properly named and stored on a web page server. This means that any developer can download all the data, load it into their own system and design their API (in this case private) according to their planned use of the data. Then high load will hit their own servers without affecting the operation of other government services. Another advantage is that it is very simple to publish data dumps on a web page server. If files and URLs are named consistently, it is easy for developers to pick up data over time (e.g., <http://exemplo.com/estradas/2015-01-30.csv>).

CONSIDERATIONS

- Do you really need an API? Designing an API can become an expensive project that competes with other IT projects with higher priority. In addition, this type of project involves making decisions about which calls will be made. Do you know how your consumers

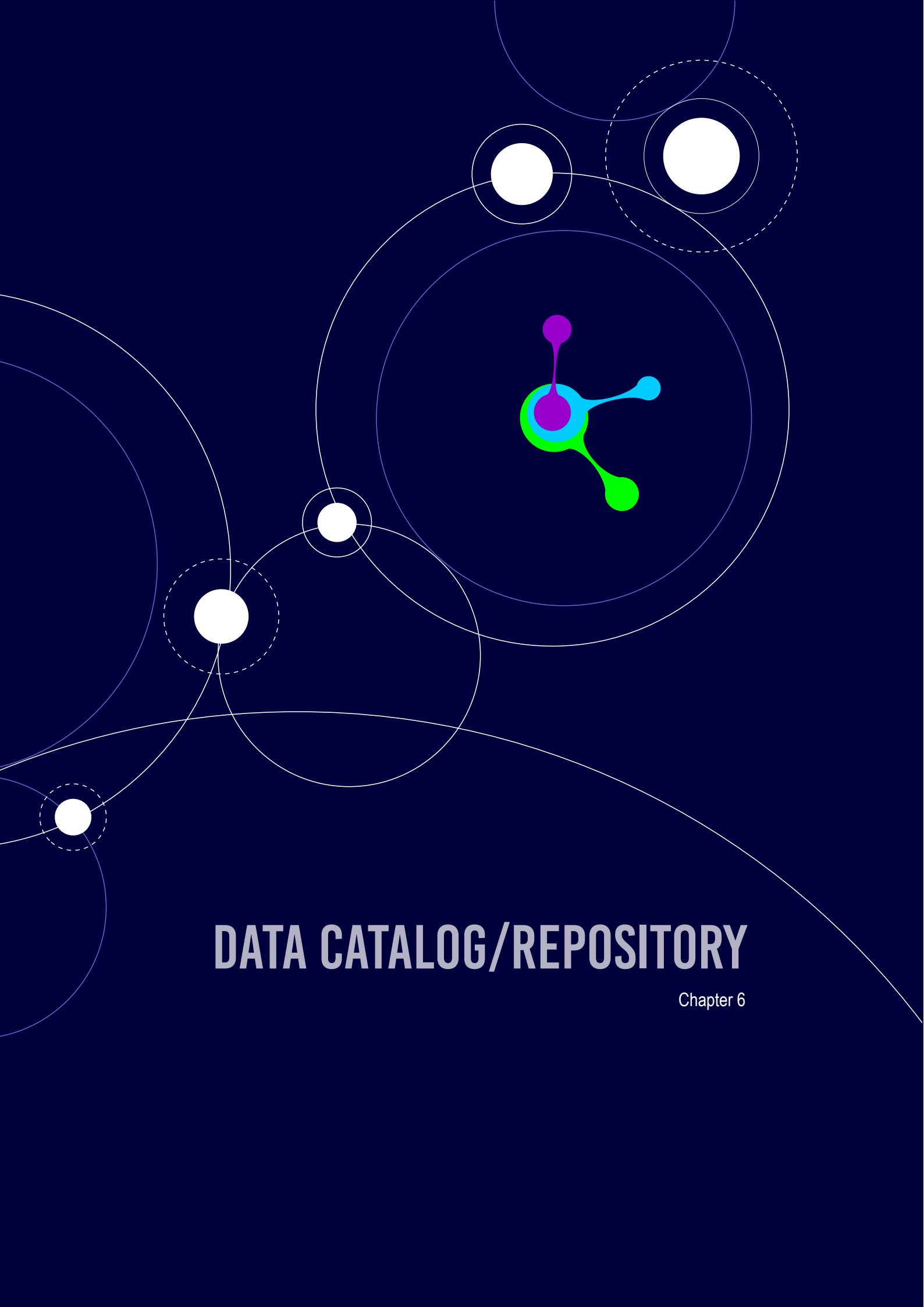
will use your data? Will your API help consumers use the data in the best way possible? What is your plan for coping with increased load?

- Make it easy for developers to keep a local copy of your data up-to-date. Providing consistently named data dumps makes this simple.
- Isolate internal systems from the effects of external data publishing. Take proper care so that the load coming from the Web does not interfere with internal databases, affecting other services of the government.
- Make sure you can change your systems without breaking URLs. Developers will build apps that depend on your URLs. Do not force them to rewrite their software just because you are switching to a new platform. Signs that things can be better designed include platform-specific fragments like “aspx” or “jsp” in your URLs. Get rid of those.

MAP OF TECHNOLOGICAL DECISIONS

The table below was adapted from the [open data kit](#) of the federal government. It shows a number of solutions for publishing open data, the most used technologies, and the expected average time for deployment. The estimates depend on the technological and human resources available to each public agency that wants to open their databases. The star system assumes that the data is published under an open license.

Solução	Pré-requisitos	Prazo	Estrelas
Publicar dump da base de dados	Acesso à base de dados	Curto	★★★★
	Servidor web para arquivos		
Publicar dados em arquivos CSV	Mecanismo de ETL (caso esteja em banco relacional)	Curto	★★★★
	Servidor web para arquivos		
Publicar dados em arquivos JSON / XML	Mecanismo de ETL (caso esteja em banco relacional)	Médio	★★★★
	Serviço de desenvolvimento		
	Servidor web para arquivos		
Desenvolver módulo de dados abertos em sistema existente	Serviço de desenvolvimento	Longo	★★★★
	Servidor web para rodar nova solução		
Desenvolver API RESTful de dados abertos desacoplada da solução (você precisa mesmo de uma API?)	Mecanismo de ETL	Longo	★★★★
	Serviço de desenvolvimento		
	Servidor web para rodar nova solução		
Novo Sistema, com a gestão de dados incorporados em sua arquitetura	Mecanismo de ETL	Longo	★★★★
	Serviço de desenvolvimento		
	Servidor web para rodar nova solução		
Publicar dados em arquivos RDF	Ontologia da área do conhecimento do sistema	Longo	★★★★★
	Mecanismo de ETL		
	Servidor web para arquivos		
Disponibilizar dados por Endpoint SPARQL	Ontologia da área do conhecimento do sistema	Mais Longo	★★★★★
	Mecanismo de ETL		
	Banco de dados de triplas		
Publicar dados em API de dados ligados (Linked Data)	Ontologia da área do conhecimento do sistema	Mais Longo	★★★★★
	Banco de dados de triplas		
	Serviço de desenvolvimento		
	Mecanismo de ETL		
	Servidor web para rodar nova solução		



DATA CATALOG/REPOSITORY

Chapter 6

São Paulo already has a centralized data catalog called [Governo Aberto SP](#) (SP Open Government). It represents the efforts of the state's Public Administration to gather into one location information about its public databases, their custodians and features, how to download them from the Web, and their formats. A central data repository is recommended so that citizens do not have to spend hours “gathering” databases from different sites of state agencies. Successful initiatives around the world show that grouping databases in a central catalog is not only recommended from a convenience point of view for citizens, but is also a smart way to measure and monitor the health of public databases available to society.

SP Open Government can stop being a final product that serves as query tool, and start being a platform. We should think about what types of information are entered and what kinds of products may come from this portal, so that backstage integration (e.g., designing scripts to automate the periodic process of publishing databases) becomes part of the server's workflow without disrupting its routine. The portal could also be a means for the government to track which open data is being published and keep it organized and updated.

Among the products that can rely on a well-managed centralized data catalog are dashboards, or dynamic panels. These tools could show the state's performance, using several indicators built from the databases available on the portal. This would be an interesting strategy, since the data used must be open and up-to-date for the dynamic panel to work, creating a virtuous cycle.

The panels can be created according to executive, department, and citizen demands in areas such as crime, finance, health, environment, and transportation. The choice can be made by brainstorming with public employees, managers, and civil society or by a public survey. This information would consist of the argument for development, which could be done through competitions. The government would make the databases and APIs available to programmers and companies, and they would develop the prototype. This relationship between the government and private sectors could stimulate the generation of new business using open data.

There are several tools on the market that can help the Public Administration implement a centralized data catalog. Currently, two that stand out are [Socrata](#), developed by an American company, and [CKAN](#), a free and open source tool maintained by [Open Knowledge](#) and a community of developers. Both tools are used around the world in government open data portals and

have advantages and disadvantages. It is up to the public manager to examine the technical characteristics to choose one that will best meet the needs and context of São Paulo.

The background is a solid dark blue. It features several overlapping circles of varying sizes. Some circles have white outlines, while others are dashed. In the center, there is a stylized molecular or network structure with a central purple circle, a cyan circle, and a green circle, all connected by thin lines. There are also four white circles: one in the top left, one in the top right, one in the middle left, and one in the bottom left. The title 'TECHNICAL SCENARIOS, TECHNOLOGICAL OPTIONS' is written in a bold, white, sans-serif font at the bottom center.

TECHNICAL SCENARIOS, TECHNOLOGICAL OPTIONS

Chapter 7

Opening databases and making them available on the Web also requires contemplating technical and infrastructure scenarios to keep these databases accessible and up-to-date. This guide is based on three different scenarios, as suggested by the Open Government Data [Toolkit of the World Bank](#).

These decisions should also be taken into account when compared to the agency's open data goals and how they fit into the five-star system. For comparison, the World Bank's technical recommendations only consider databases with, at most, three stars.

Three levels of complexity are considered, based on the number of databases available and the update frequency for each:

Level 1: less than 100 databases with less than 10 databases updated each week

Level 2: 100 to 1,000 databases with 10 to 100 databases updated each week

Level 3: more than 1,000 databases with 100 or more databases updated each week

This guide provides technical solutions as recommendations only without endorsing any software or technology solution from a specific company.

LEVEL 1

The first level of complexity is for agencies that want to start their open data efforts and do not have a very large number of databases. In such cases, a single IT professional is capable of managing the location where the databases will be made available.

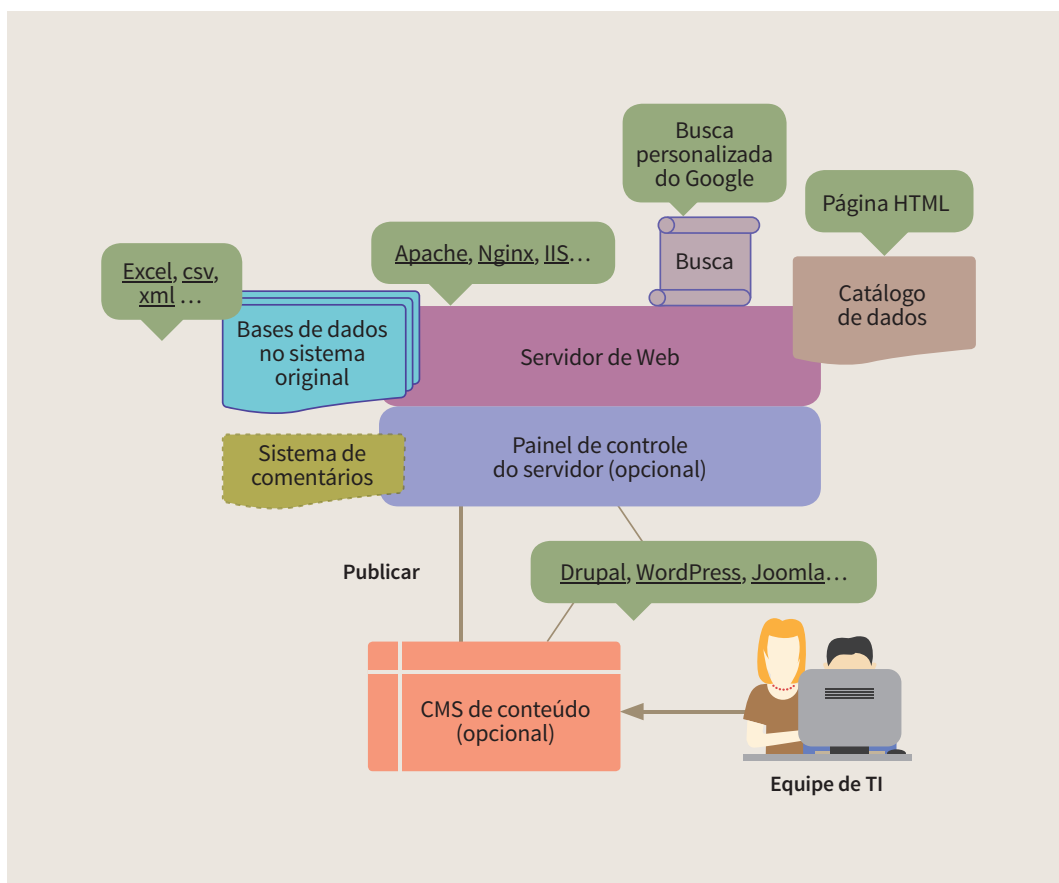
An agency that has less than 100 databases to publish on the Web can configure a standard web server, hosting the documents on the server itself or a cloud storage service. The databases must have their description (also known as “metadata”) embedded in the page where they are published and allow the use of existing search tools such as [Google's custom search](#).

It is not necessary to configure an automatic update routine for the databases; the data can be manually sent with the support of the agency's IT staff. It is

recommended that a comment system be included on the database page in order to encourage suggestions and criticisms from consumers who access the databases.

Summary:

- Set up a site using a standard web server.
- Host data directly on the server (or using a cloud hosting service).
- Metadata embedded in the page itself where the databases are displayed.
- Set up a search on the site with existing tools such as Google's custom search.
- Manual updating of data, metadata, and content with support of the IT staff.
- Include a comment system on the database pages to receive suggestions.



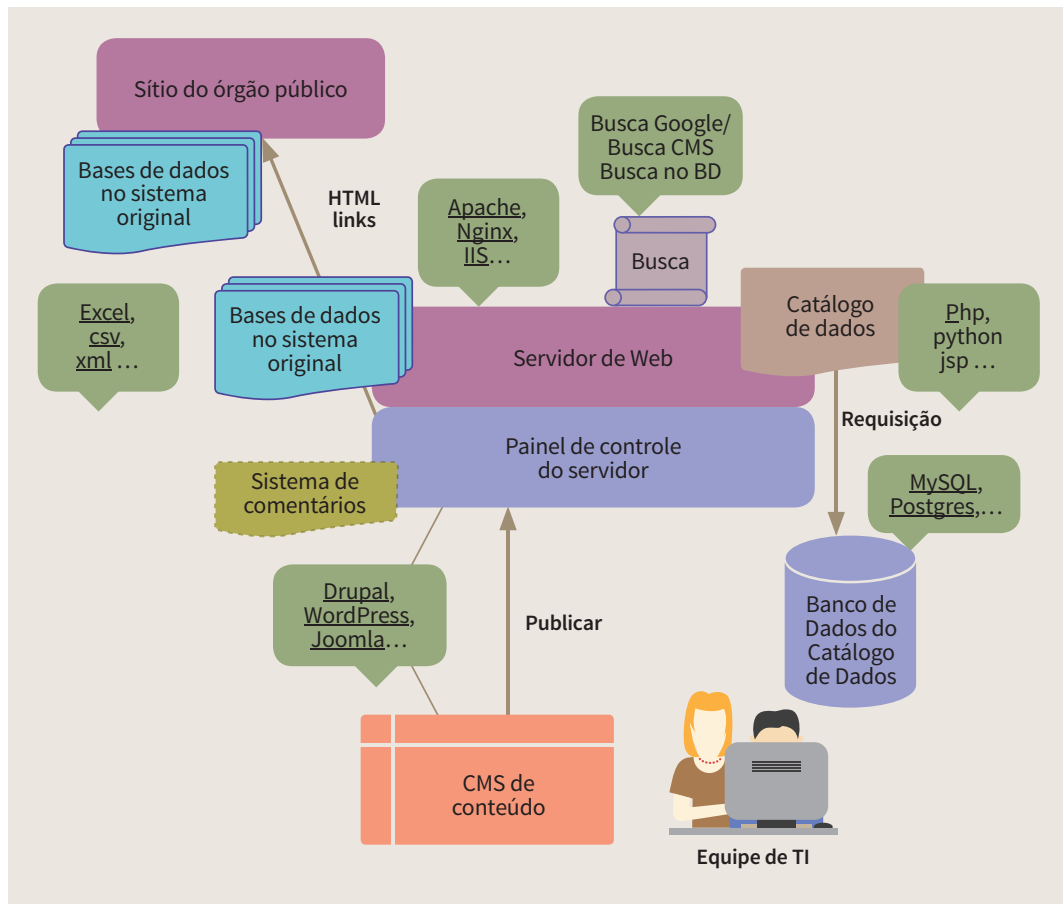
LEVEL 2

The second level of complexity describes a scenario where a moderate number of databases must be published on the Web. The solution presented can be shared among various agencies depending on the demand. At this level of complexity, it is recommended that the IT staff have specialists in optimized configuration of servers and databases.

Publishing 100 to 1,000 databases may occur on a central server that acts as a database repository for the agency's site. The access is done using a CMS (content management system) such as [WordPress](#) or [Drupal](#) and data is managed manually in its original format on the same CMS server, or stored in the cloud. The metadata of the databases is stored in a [SQL](#) database, dynamically displayed on the pages generated. The system automatically checks broken links, and the search can be done directly in the metadata database or by using the CMS options or Google custom search. It is recommended that caching be used to deal with high loads, and that the metadata not be stored in the CMS.

Summary:

- Standard front end using CMS (WordPress, Drupal, etc);
- Databases are hosted in their original format on the application server itself or on the websites of the respective agencies in the case of a shared activity.
- Metadata is hosted in SQL databases and calls are done via dynamic page generation.
- Automatic check for broken links.
- Search via text scan in SQL database, CMS search, or Google custom search.
- Perform caching to balance load.
- Do not save metadata in the CMS.



LEVEL 3

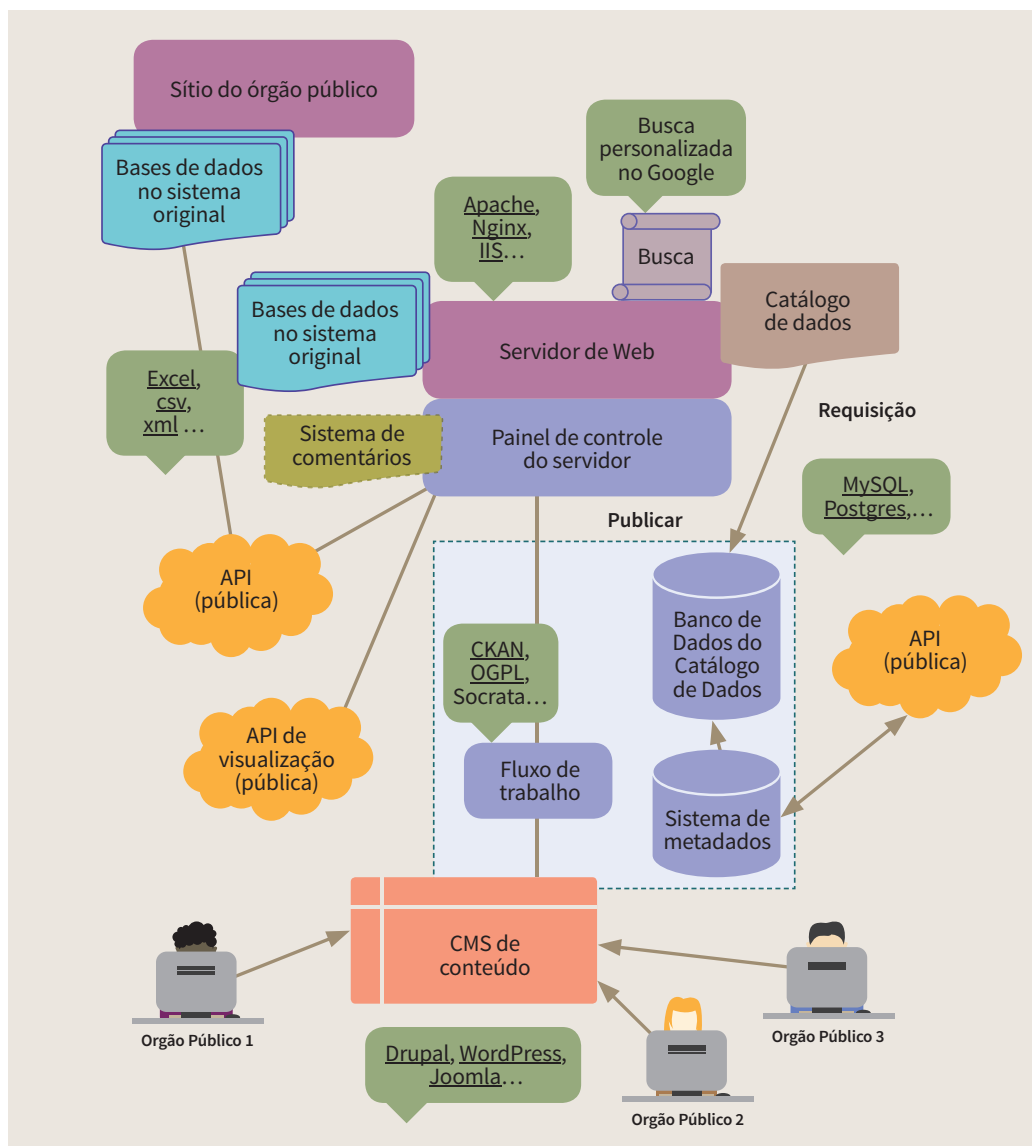
The third level of complexity describes a scenario where more than 1,000 databases are published and 10% are frequently updated. In such cases, it is common for the platform to manage databases from various agencies with specialized IT staff that can meet the demands of various sectors in a timely manner.

One or more servers can be used, depending on the preference of the infrastructure's management team. The front-end part must integrate different web services, preferably with an API. The database documents should be managed automatically, possibly with a cloud storage solution. The metadata is stored in an optimized repository. Sending and updating databases is delegated to each public agency that shares the platform, with automatic validation of databases and access control levels. Broken links are checked automatically. A system should be provided so that consumers can report errors to the person

responsible for each database, via a web form, for example. A search can be implemented by structured queries in the metadata repository in the CMS or using Google custom search on the main site.

Summary:

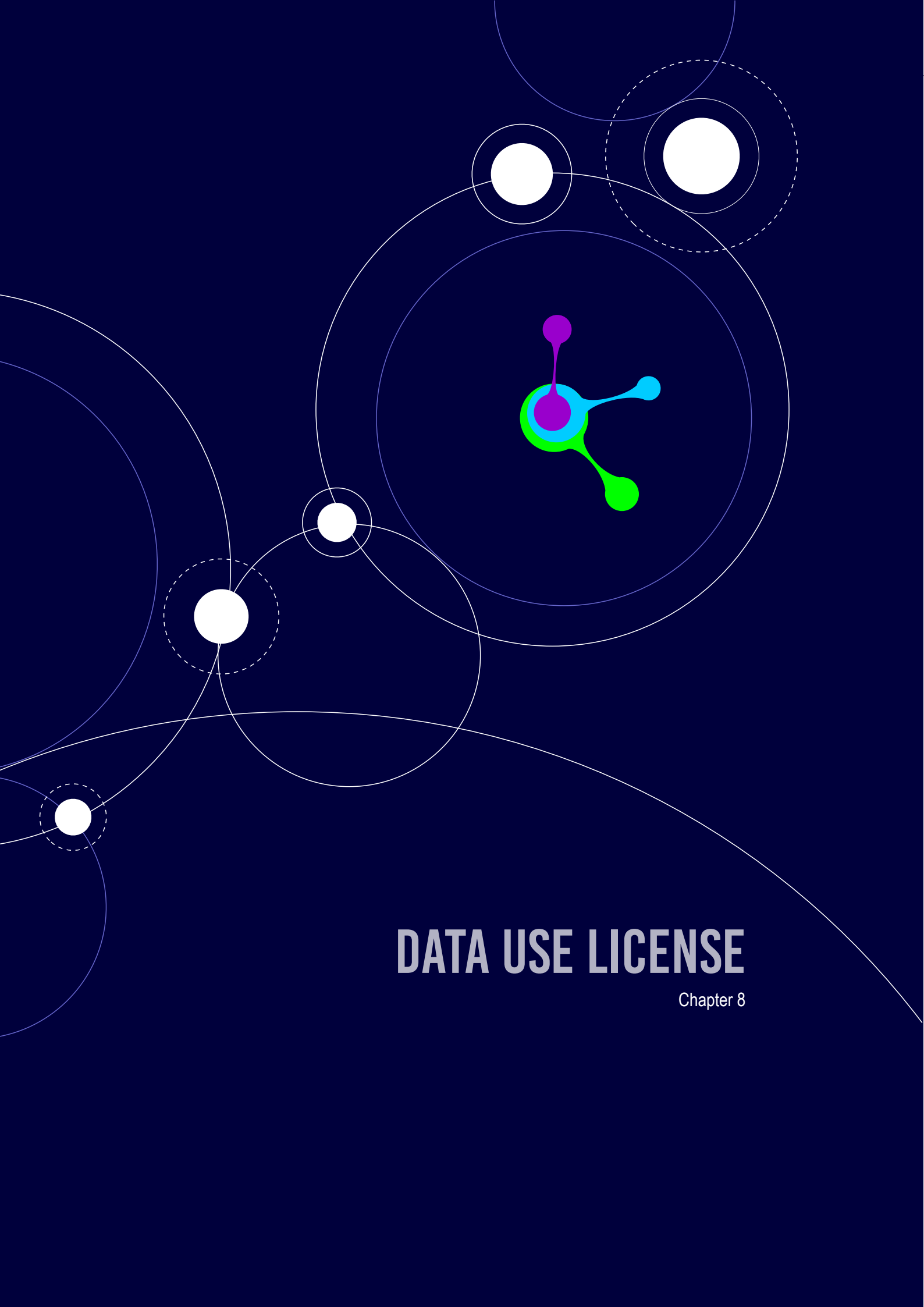
- The front end integrates various web services, preferably by API.
- Automatic management of the original data files stored in a cloud solution.
- Metadata is stored in an optimized repository.
- Sending and updating databases is delegated to each public agency, with automatic validation of databases and access control levels.
- Automatic check for broken links.
- System allowing consumers can report errors to the person responsible for each database via a web form or similar way.
- Structured search in the metadata repository in the CMS or using Google custom search on the main site.



BEST INFORMATION SECURITY PRACTICES

Just like any other initiative involving information technology, opening databases should follow the strict criteria for best information security practices. Please note:

- Data governance to ensure:
- Authority of the source
- Rules of engagement
- Sustainability
- Classification of public vs. confidential data to ensure:
- That private or confidential data cannot be accessed externally.
- Information security controls to ensure:
- Confidentiality and data integrity
- Protection against denial-of-service attacks (DoS)



DATA USE LICENSE

Chapter 8

Opening databases in the terms suggested by this guide involves providing complete information on the Web in non-proprietary formats with unrestricted, free access for anyone, such that anyone can reuse the data for any purpose without restrictions. This paragraph is a brief summary of what can be considered an “open license”, and the conditions under which the custodian of the data grants the consumer permission to use it. Without this license, the data cannot be considered “open,” because there is no guarantee that its use is kept under open data principles.

An “open license” is also important in respecting the Brazilian Information Act. [The National Infrastructure of Open Data](#), of the federal government [has not yet reached a conclusion](#) as to whether the Brazilian legal system is sufficient to deal with open government data without the need for licenses, if other existing licenses fit into the Brazilian context, or if it will be necessary to create a specific license for Brazil. For example, in the [Copyright Act](#), Brazilian legislation explicitly provides for the protection of databases that constitute “intellectual creation.” On the other hand, the Brazilian Information Act states that the state must provide access to all information that it creates or holds, provided that data is not considered confidential. The interaction of these laws is still an object of study.

The Brazilian Information Act, however, does not determine which license should be used when publishing data; but it lists principles that may very well direct the preparation of an open license or terms of use compatible with the open data concepts explained in this guide:

Art. 3. The procedures provided for in this Act are intended to ensure the fundamental right of access to information and must be carried out in accordance with the basic principles of public administration and with the following guidelines:

I - Publishing as a general default rule and confidentiality as an exception.

Art. 8º ...

...

§ 3. The sites mentioned in § 2 must meet the following requirements under the regulations:

...

II - Allow reports to be saved in various electronic formats, including open and non-proprietary formats such as spreadsheets and text, in order to facilitate the analysis of information.

III - Enable automated access by external systems in open, structured and machine-readable formats.

IV - Disclose in detail the formats used for structuring the information.

V - Ensure the authenticity and integrity of information available for access.

VI - Keep the information available for access up-to-date.

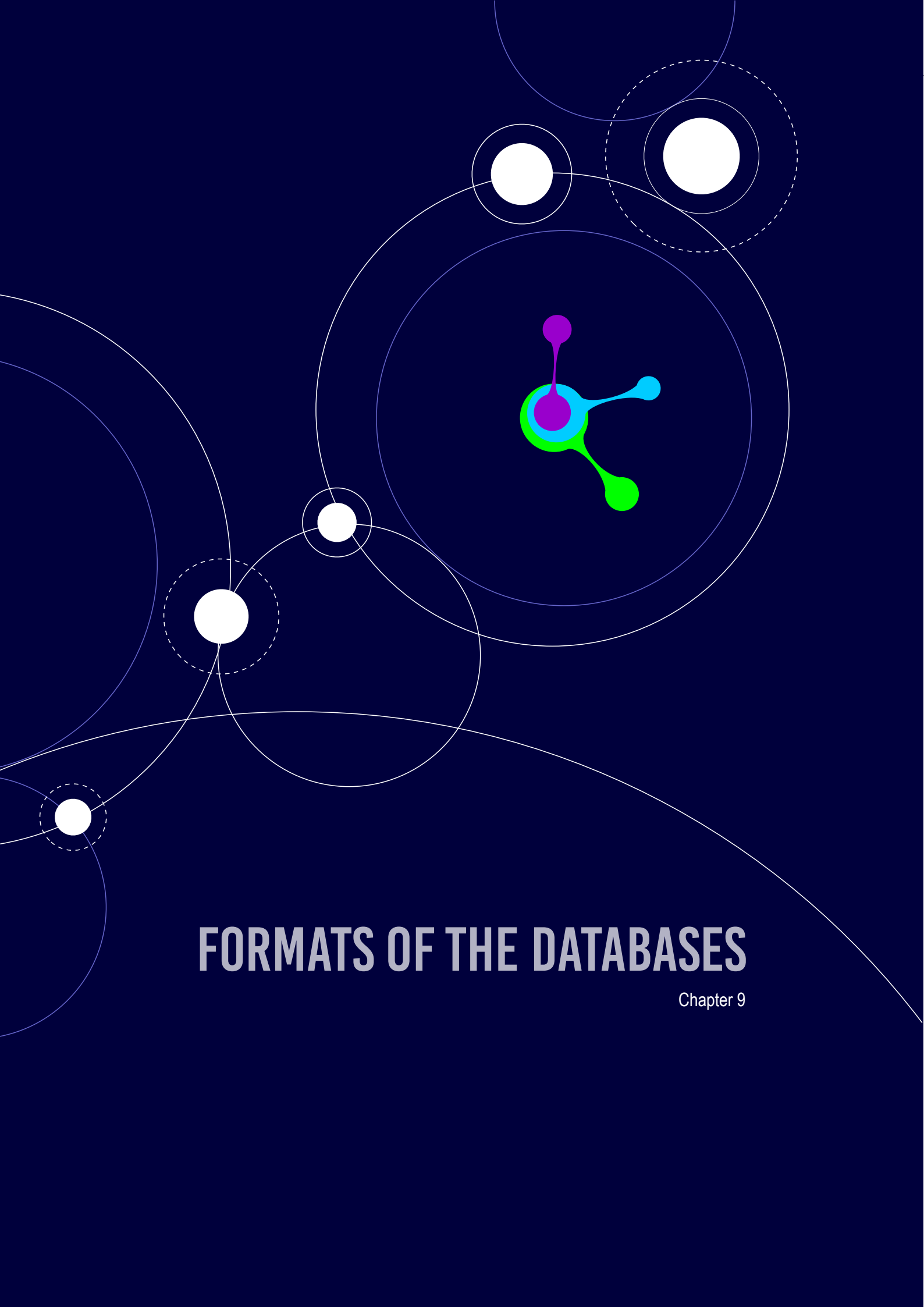
When preparing terms of use or an open license for use of data, it is your responsibility to consider whether the text is compatible with other open data licenses. This consideration is vital for it to be legally possible to perform aggregations, statistics, analysis, and crossing of data from different sources. The results of these activities are what add value to open data and make it useful for society. If the state government uses a license incompatible with the license used by municipal administrations, it is not legally possible to cross the database of hospitals, for example, to build an application that displays city and state health facilities. An open data license or terms of use should be prepared to maximize the degree of compatibility with the licenses used, or likely to be used, by other spheres and branches of the state.

In Brazil, most of the public agencies have not specified any license for publishing data on the Internet. There are exceptions, however, where they are using the [ODbL](#) and [DbCL](#) licenses.

TERMS OF USE OF SP OPEN GOVERNMENT

The government of São Paulo already has a central data portal called [Governo Aberto SP](#) (SP Open Government). The terms of use are based on open principles and can serve as a model for preparing open licenses, with minor adjustments depending on the needs of each agency and their suitability to the open data precepts of this guide. Briefly, [the terms of use of SP Open Government](#) are the following:

1. Any person or company can use the data.
2. There are no established restrictions on the use of data.
3. The government ensures the authenticity and integrity of the data, and that it is up-to-date, only for data downloaded in the SP Open Government portal.
4. In order to reuse the data, consumers must quote the original source of the data (public agency), disclosing that the data was accessed using SP Open Government. Also, the following declaration must be added: “SP Open Government and the agency or entity from which the data were extracted do not guarantee its authenticity, quality, integrity, or whether it is up-to-date after being made available for secondary use”.
5. Consumers are responsible for the secondary use of data, exempting the state.
6. A condition for accessing the data is knowing and accepting the rules.



FORMATS OF THE DATABASES

Chapter 9

In the scope of this guide, a database is nothing more than a computer file built in a structured way in order to store information for subsequent consultation and analysis. Your database can be built manually, provided you define a structure for organizing the data and maintaining consistency. This is important to ensure that queries performed on this database find what they are looking for. A database can be a text file, for example, with a list of all the cities in the state of São Paulo, or a list of the hospitals in the city of São Paulo showing the district where each is located:

```
Hospital Municipal Infantil Menino Jesus, Bela Vista
Hospital do Servidor Público Municipal, Aclimação
Pronto-Socorro Municipal Barra Funda, Barra Funda
Hospital Municipal Cidade Tiradentes, Cidade Tiradentes
```

In this case, the structure is defined by putting two names (hospital and district) in each new row of the file, separated by a delimiter, a comma. Two hospitals never appear in the same row, for example. Roughly speaking, what defines the integrity of a database, is the elements used in order to give predictability to queries on that database: In the above example, all the rows have the name of the hospital first and the district where it is located second. If any row of this database is different from the model “Hospital name, District,” the integrity of the database will be compromised and it will lose its usefulness:

```
Hospital Municipal Infantil Menino Jesus, Bela Vista
Aclimação, Hospital do Servidor Público Municipal
Pronto-Socorro Municipal Barra Funda, Barra Funda
Hospital Municipal Cidade Tiradentes, Cidade Tiradentes
```

In most cases, however, the right tools on the computer are able to automatically create or convert structured files that serve as databases. One of the most common examples is the Excel spreadsheet, computer files ending in “.xls” or “.xlsx.” These documents have rows and columns and enable subsequent analysis and comparison. However, the format of native Excel files uses a proprietary, closed technology; such technologies often cost money and are not widely and freely available to everyone.

The list below suggests a number of open and non-proprietary formats that best fit into the open data principles presented in this guide and provides a brief introduction to each. One format is not recommended over the others. Each team should think about the formats that the databases are currently in (Excel files, for example) and whether there is a means of converting them into any of the formats suggested below, depending on the application.

DELIMITER-SEPARATED FORMATS (CSV)

CSV files (Comma-separated values) are used to store tabular data (numbers and text) in plain text. “Plain text” means that the file is a pure string of characters without any hidden information that the computer has to process.

A CSV file stores data without a “record” number, separated by line breaks (each line of the file is a data “record”). Each record has one or more “fields” separated by a delimiter, most commonly a comma (“,”), semicolon (“;”) or the “invisible” character that appears when you press the “tab” key. Files separated by commas and semi-colons usually receive the “CSV” extension and files separated by a “tab” the “TSV” extension. There are also databases in these formats that receive the “TXT” extension. CSV files are simple and work in most applications that deal with structured data.

Making a comparison with rows and columns in a spreadsheet, the “records” in a CSV file are the rows and the “fields” are the columns. The first “record,” which is the first line, usually contains column names for each of the “fields.” Although an international standard does not exist for CSV, its variations are simple enough so that compatible applications can easily fix the differences. Typically, this is how a CSV file is displayed when opened in a text editor:

```
Continente;País;Capital
África;Angola;Luanda
América do Norte;Estados Unidos;Washington DC
América Central;México;Cidade do México
América do Sul;Brasil;Brasília
Europa;Espanha;Madri
Europa;Alemanha;Berlim
Oceania;Austrália;Camberra
Ásia;Japão;Tóquio
```

This file contains three columns separated by the semicolon (“;”) delimiter: Continent, Country and Capital, as described in the first line. In all there are eight records. The first triad is Africa-Angola-Luanda and the last is Asia-Japan-Tokyo. There is no practical limit to the number of lines or columns in a CSV file. This number can reach millions or tens of millions, depending only on the processing power of the computer that will be used in querying. If the same CSV file was opened in a spreadsheet processor, it would be displayed like this:

Continente	País	Capital
África	Angola	Luanda
América do Norte	Estados Unidos	Washington DC
América Central	México	Cidade do México
América do Sul	Brasil	Brasília
Europa	Espanha	Madri
Europa	Alemanha	Berlim
Oceania	Austrália	Camberra
Ásia	Japão	Tóquio

XML FORMAT

XML is a markup language, similar to HTML (used to build web pages) that is defined and maintained by the World Wide Web Consortium (W3C). The goals of XML emphasize simplicity, generality, and usability across the Internet. Although XML focuses on creating documents, it is also used to represent arbitrary data structures, for integration between computer systems. A typical XML file has the following structure:

```

<?xml version="1.0" encoding="UTF-8"?>
<Exemplo>
  <Localidade número="1">
    <Continente>África</Continente>
    <País>Angola</País>
    <Capital>Luanda</Capital>
  </Localidade>
  <Localidade número="2">
    <Continente>América do Norte</Continente>
    <País>Estados Unidos</País>
    <Capital>Washington DC</Capital>
  </Localidade>
  <Localidade número="3">
    <Continente>América Central</Continente>
    <País>México</País>
    <Capital>Cidade do México</Capital>
  </Localidade>
  <Localidade número="4">
    <Continente>América do Sul</Continente>
    <País>Brasil</País>
    <Capital>Brasília</Capital>
  </Localidade>
  <Localidade número="5">
    <Continente>Europa</Continente>
    <País>Espanha</País>
    <Capital>Madri</Capital>
  </Localidade>
  <Localidade número="6">
    <Continente>Europa</Continente>
    <País>Alemanha</País>
    <Capital>Berlim</Capital>
  </Localidade>
  <Localidade número="7">
    <Continente>Oceania</Continente>
    <País>Austrália</País>
    <Capital>Camberra</Capital>
  </Localidade>
  <Localidade número="8">
    <Continente>Ásia</Continente>
    <País>Japão</País>
    <Capital>Tóquio</Capital>
  </Localidade>
</Exemplo>

```

Markup and content

An XML file has two main features: markup and content. Generally, strings that constitute “markup” either begin with the character < and end with >, or they begin with the character & and end with ;. Strings of characters that are not markup are “content.” In the example above, and are markup. The names of countries, continents and capitals are the “content.”

Tags

Tags are markup that begins with “<” and ends with “>.” There are three types of tags:

- Start-tags; for example: <Location>
- End-tags; for example: </Location>
- Empty-element tags; for example: <line break />

Elements

Elements are XML components that begin with a start-tag and end with a corresponding end-tag, or consist of only an empty-element tag. The string of characters between the start- and end-tags, if any, are the element’s content, and may contain markup, including other elements, which are called “child” elements. In the example above, an element would be

```
<País>Brasil</País>.
```

Attributes

Attributes are “name/value” pairs that exist within a start-tag or empty-element tag. In the example above, the element <Location> has a “number” attribute and a corresponding value:

```
<Localidade número="8">
```

The name of the attribute is “number” and its value is “8.” Attributes can only have a single value in quotes, and each attribute cannot appear more than once in each element.

XML declaration

XML documents should begin by declaring some information about themselves, as in the following example:

```
<?xml version="1.0" encoding="UTF-8"?>
```

KML FORMAT

Keyhole Markup Language (KML) is an XML notation for expressing geographic data and visualization within Internet-based, two-dimensional and three-dimensional map browsers. The format was acquired by Google in 2004 and became the standard used in the [Google Earth](#) application. In 2008, the format became an international standard of the [Open Geospatial Consortium](#).

The KML format has a structure similar to XML, but specifies a set of features, such as place marks, images, polygons, 3D models and textual descriptions. Each location always has a longitude and a latitude. The files are distributed in KMZ packages, which are zipped KML files with a .kmz extension. The contents of a compressed package include a single KML document (“doc.kml”) and optional subdirectories containing images and other files referenced in the KML. A typical KML document is presented as follows:

```
<?xml version="1.0" encoding="UTF-8"?>
<kml xmlns="http://www.opengis.net/kml/2.2">
  <Document>
    <Placemark>
      <name>São Paulo</name>
      <description>Cidade de São Paulo</description>
      <Point>
        <coordinates>-23.5476258,-
          46.6360159</coordinates>
      </Point>
    </Placemark>
  </Document>
</kml>
```

JSON FORMAT

The JSON (JavaScript Object Notation) format is an open format used as an alternative to XML to transmit structured data between a web server and a web application. Its organization logic is similar to XML, but has a different notation. The format gained popularity in web services such as email clients and shopping sites because it can transmit a lot of information between the client and server using fewer characters.

JSON files also work with pairs of keys and values, and instead of markup such as XML, it uses delimiters in sets: {}, [], and “. A typical JSON file is structured as follows:

```

{
  "localidade 1": {
    "Continente": "África",
    "País": "Angola",
    "Capital": "Luanda"
  },
  "localidade 2": {
    "Continente": "América do Norte",
    "País": "Estados Unidos",
    "Capital": "Washington DC"
  },
  "localidade 3": {
    "Continente": "América Central",
    "País": "México",
    "Capital": "Cidade do México"
  },
  "localidade 4": {
    "Continente": "América do Sul",
    "País": "Brasil",
    "Capital": "Brasília"
  },
  "localidade 5": {
    "Continente": "Europa",
    "País": "Espanha",
    "Capital": "Madri"
  },
  "localidade 6": {
    "Continente": "Europa",
    "País": "Alemanha",
    "Capital": "Berlim"
  },
  "localidade 7": {
    "Continente": "Oceania",
    "País": "Austrália",
    "Capital": "Camberra"
  },
  "localidade 8": {
    "Continente": "Ásia",
    "País": "Japão",
    "Capital": "Tóquio"
  }
}

```

The { delimiter marks the beginning of a section and } marks its end. The key and value pairs are separated by : and their values when text is in quotation marks (numbers, for example, are not in quotes). In the example below, “location 6” is a key that receives an array of values (Continent, Country and Capital):

```
"localidade 6": {  
  "Continente": "Europa",  
  "País": "Alemanha",  
  "Capital": "Berlim"  
},
```

Note that the value of “location 6” is a new array of key-value pairs. This new array starts with the delimiter { and ends with }. This logic of linking sets of pairs may be repeated numerous times, creating many levels for the desired data structure.

GEOJSON/TOPOJSON

The [geoJSON](#) and [topoJSON](#) formats are derived from JSON for representing collections of simple geographical features along with their non-spatial attributes. Among the possible features that can be stored in the GeoJSON/topoJSON standard are “points,” including addresses and locations; “line strings,” including streets, highways and boundaries; “polygons,” including countries, provinces or tracts of land; and multi-part collections of these types. The differential of topoJSON compared to GeoJSON is that it stores geospatial topology and typically provides smaller file sizes.

SQL FORMAT (DUMP)

SQL (Structured Query Language) is a programming language specifically designed for managing data held in [relational database](#) systems. Possible SQL commands include data insert, query, update and delete, database schema creation and modification, and data control. A database “dump” typically results in a list of SQL statements and allows anyone to restore the database by using its database schema and the values contained in it. A “dump” file typically is provided as follows:


```

-- Base de dados
CREATE DATABASE `ex_localidades`;
USE `Exemplos de Localidades`;

-- Estrutura da tabela para a tabela `localidades`
CREATE TABLE `localidades` (
  `id` INT(8) UNSIGNED NOT NULL AUTO_INCREMENT,
  `nome de usuário` VARCHAR(16) NOT NULL,
  `senha` VARCHAR(16) NOT NULL,
  PRIMARY KEY (`id`)
);

-- Dados da tabela `localidades`
INSERT INTO `localidades` VALUES ('Continente', 'País',
'Capital'), ('África', 'Angola', 'Luanda'), ('América
do Norte', 'Estados Unidos', 'Washington DC'), ('América
Central', 'México', 'Cidade do México'), ('América do Sul',
'Brasil', 'Brasília'), ('Europa', 'Espanha', 'Madri'),
('Europa', 'Alemanha', 'Berlim'), ('Oceania', 'Austrália',
'Camberra'), ('Ásia', 'Japão', 'Tóquio');

```

SQL databases are usually created and managed using tools intended for IT professionals. While CSV, XML, and JSON formats can be easily created in common text editors, an SQL database requires further refinement and technical knowledge.

SHAPEFILE FORMAT

Shapefile is a geospatial vector data format for geographic information systems (GIS). It was developed and is regulated by the [Esri](#) Company. It is considered an open format, though it is proprietary. Since it is open, the format is supported by many free and open-source map processing applications. Shapefile can spatially describe vector features (points, lines and polygons of rivers, lakes and wells, for example), and each item usually has attributes that describe it, such as name and temperature.

Despite having a singular name, the shapefile format consists of a collection of files with a common filename and different extensions stored in the same directory. The following three files are mandatory for a shapefile to work properly: .shp, .shx and .dbf. The actual shapefile is the .shp file,

but if distributed alone, it will not be able to display the stored data. The distribution must be done together with the other two files.

Required files are the following:

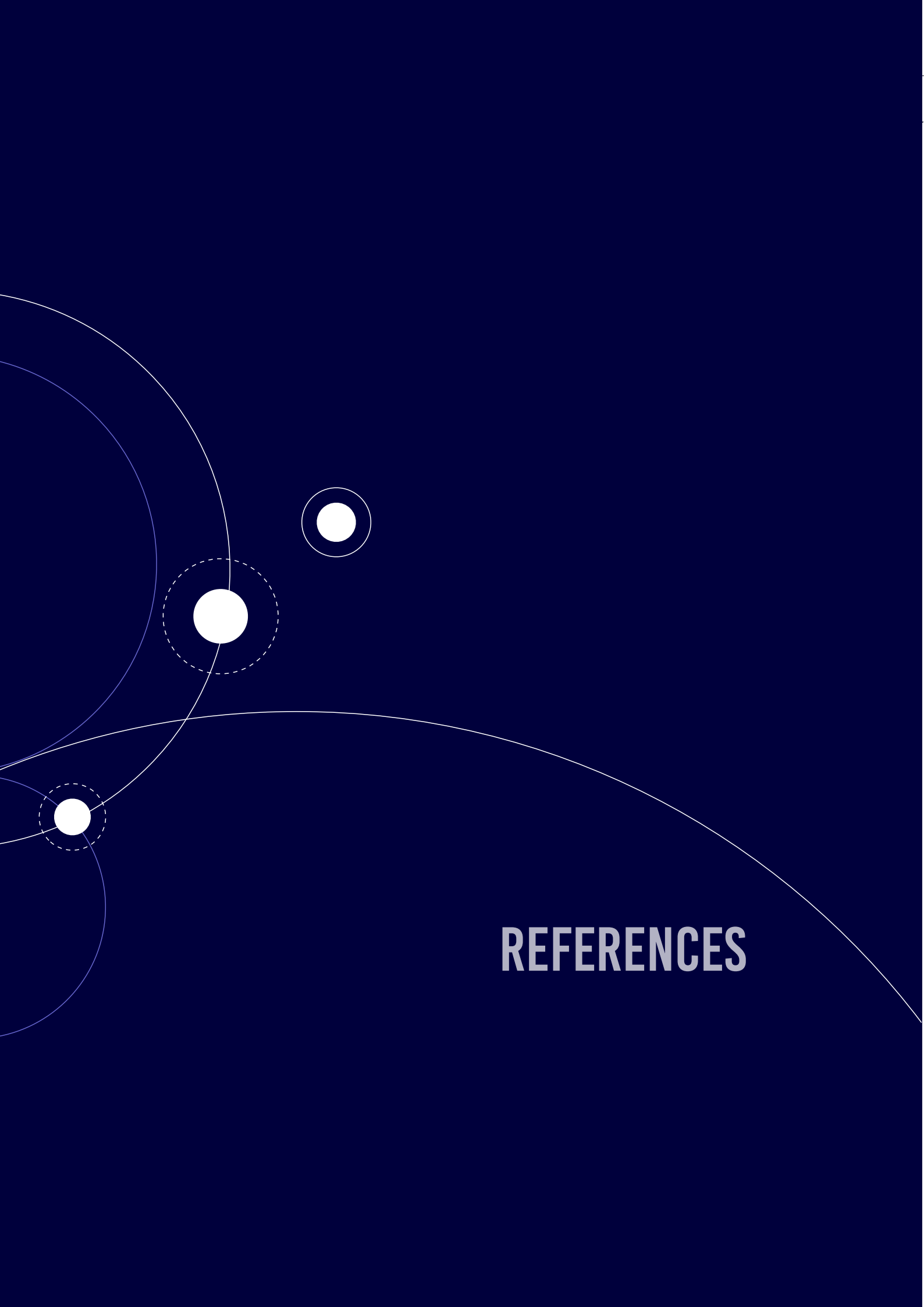
- .shp - shape format; the feature geometry itself

- .shx - shape index format; an index with the geometry's features to allow faster searches

- .dbf - attribute format; columnar attributes for each "shape"

In each of the required files, the shapes in each file correspond to each other in sequence (the first record in the .shp file corresponds to the first record in the .shx and .dbf files, and so on).

It is not feasible to generate shapefiles manually, as would be possible to create CSV, XML and JSON files, because they are binary. Typically, these files are created from manipulating points and features on maps in its own GIS software.



REFERENCES

In addition to giving references in the main text of the guide using external links, below is a list of further reading that we used as a basis for building the concepts presented in this guide. The references also contribute to expanding and extrapolating the building processes for open data, its challenges, development, and technical details.

CSV:

https://en.wikipedia.org/wiki/Comma-separated_values

Código de Conduta para uso de dados abertos do governo (Reino Unido):

<http://data.gov.uk/code-conduct>

Exemplos de Licenças Abertas (Governo dos Estados Unidos):

<https://project-open-data.cio.gov/license-examples/>

geoJSON:

<http://geojson.org/>

Guia de Implantação de um Portal de Transparência (CGU):

http://www.cgu.gov.br/Publicacoes/transparencia-publica/brasil-transparente/arquivos/guia_portaltransparencia.pdf

Guia sobre Informações Classificadas (CGU):

http://www.acessoainformacao.gov.br/lai-para-sic/sic-apoio-orientacoes/guias-e-orientacoes/guia_informacoesclassificadas.pdf/@download/file/Guia_InformacoesClassificadas.pdf

Hampshire County Open Licence (Reino Unido):

<http://www3.hants.gov.uk/opendata/licence.htm>

JSON:

<http://www.json.org/>

Kit de Dados Abertos (Infraestrutura Nacional de Dados Abertos):

<http://kit.dados.gov.br/>

KML:

<https://developers.google.com/kml/documentation/?hl=pt-br>

Manual de Lei de Acesso à Informação para Estados e Municípios (CGU):

http://www.cgu.gov.br/Publicacoes/transparencia-publica/brasil-transparente/arquivos/manual_lai_estadosmunicipios.pdf

Open Definition (Open Knowledge):

<http://opendefinition.org/od/index.html>

Open Government Data (book):

<https://opengovdata.io/>

Open Government Guide (Open Government Data - Sunlight Foundation & Open Knowledge):

<http://www.opengovguide.com/topics/open-government-data/>

Open Government Licence for Public Sector Information (Reino Unido):

<http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/>

Open Government Toolkit (Banco Mundial):

<http://data.worldbank.org/open-government-data-toolkit>

Publishing Open Data: do you really need an API?:

<https://www.peterkrantz.com/2012/publishing-open-data-api-design/>

Shapefile:

<http://doc.arcgis.com/pt-br/arcgis-online/reference/shapefiles.htm>

SQL (dump):

https://en.wikipedia.org/wiki/Database_dump

Texto Lei de Acesso à Informação - Lei nº 12.527, de 18 de novembro de 2011. (Governo Federal):

http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm

Tool Kit para publicação de dados em formato aberto:

<http://platform.od4d.org/article?locale=pt&uri=http%3A%2F%2Fplatform.od4d.org%2Fposts%2F58>

topoJSON:

<https://en.wikipedia.org/wiki/GeoJSON#TopoJSON>

Uso e reuso de dados governamentais:

<http://br.okfn.org/2013/08/28/dados-meio-abertos-sobre-o-uso-e-reuso-dos-dados-governamentais-brasileiros/>

XML:

<http://www.w3.org/XML>

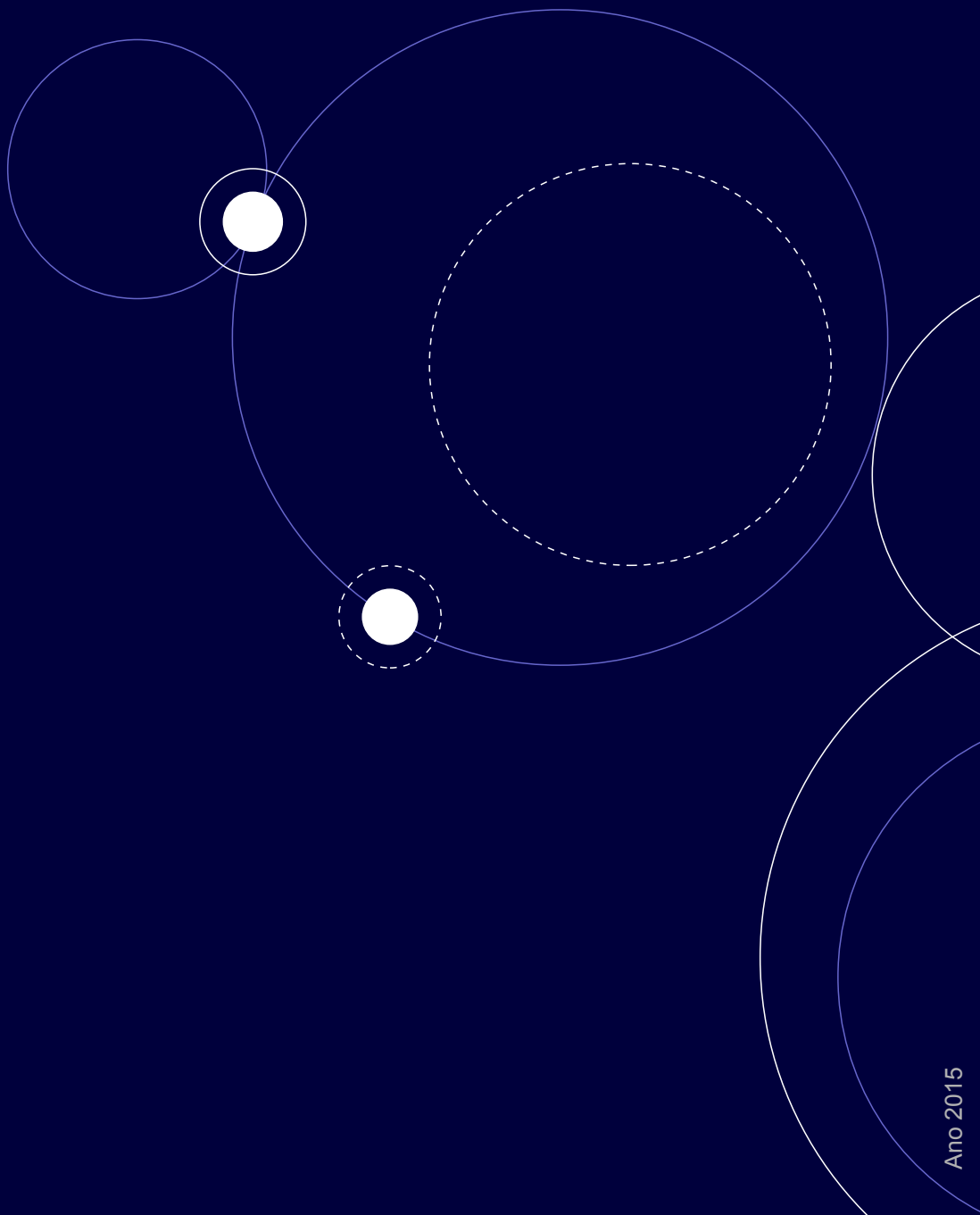


SPUK



Improving business environment through transparency in São Paulo State

Melhoria do ambiente de negócios por meio da transparência no Estado de São Paulo



Ano 2015

ceweb.br nic.br cgi.br

SEADE
Fundação Sistema Estadual
de Análise de Dados

Fundap



Embaixada Britânica
Brasília



This content is licensed under Creative Commons.
Attribution-NonCommercial-NoDerivs
CC BY-NC-ND