dimensions and characteristics of the Brazilian

Web

a study of the gov.br

2010

cgi.br
Brazilian Internet Steering
Committee

nic.br
Brazilian Network
Information Center

Brazilian Internet Steering Committee – CGI.br

Coordinator
Augusto Cesar Gadelha Vieira

Advisors
Adriano Silva Mota
Alexandre Annenberg Netto
Carlos Alberto Afonso
Demi Getschko
Ernesto Costa de Paula
Flávio Rech Wagner
Francelino José Lamy de Miranda Grando
Gustavo Gindre Monteiro Soares
Henrique Faulhaber
Jaime Barreiro Wagner
Jorge Santana de Oliveira
Lisandro Zambenedetti Granville
Marcelo Bechara de Souza Hobaika
Marcelo Fernandes Costa
Mario Luis Teza
Nelson Simões da Silva
Nivaldo Cleto
Plínio de Aguiar Junior
Renato da Silveira Martini
Rogério Santanna dos Santos

Managing Director
Hartmut Richard Glaser

Network Information Center – NIC.br

Chief Executive Officer
Demi Getschko

Financial and Managing Director
Ricardo Narchi

Services and Technology Director
Frederico Neves

Special Projects and Development Director
Milton Kaoru Kashiwakura

GENERAL MANAGEMENT

NIC.br / CEPTRO - Center of Study and Research in Network Technology and Operations
Antonio Marcos Moreiras
Heitor de Souza Ganzeli
Pedro Hadek

NIC.br / CETIC – Center of studies on Information and Communication Technologies
Alexandre Barbosa
Juliano Cappi
Robson Tavares

W3C - Brazilian Office
Carlinhos Cecconi
Vagner Diniz
Orípide Cilento Filho

Communications Department
Caroline D'Avo

PARTNERS

InWeb – National Institute for Web Research

Adriano C. Machado Pereira
Cristina Duarte Murta
CEFET-MG - Federal Center of Technological Education of Minas Gerais, IT Department

Altigran da Silva
UFAM - Federal Universiy of Amazonas, Science and IT Department

Dimitri Fazito de Almeida Rezende
Eduardo Luiz Gonçalves Rios-Neto
UFMG - Federal University of Minas Gerais, Demography Department

Dorgival Olavo Guedes Neto
Renato Ferreira
Wagner Meira Jr
UFMG - Federal University of Minas Gerais, Computer Science Department

Ministry of Planning, Budget and Management

Cláudio Muniz Machado Cavalcanti
João Batista Ferri de Oliveira
SLTI - Logistics and Information Technology Bureau

ABEP

Dayse Vianna
PRODERJ - Center of Information and Communication of the State of Rio de Janeiro

Kátia Bruno
CEPROMAT - Center of Data Processing of the State of Mato Grosso

SPECIAL ACKNOWLEDGMENT TO THE FOLLOWING COLLABORATORS:

Gustavo da Gama Torres
José Maria Leocádio
SERPRO - Federal Data Processing Service

Isabele dos Passos Omena
José Nilo Martins Sampaio
ATI Information Technology Agency, Government of the State of Pernanmbuco

Nicolau Reinhard
FEA - School of Economy and Administration, University of São Paulo

Orion Borba
CIASC Center of Information and Automation of the State of Santa Catarina

Paulo Maia
Federal Savings and Loans Bank

Roberto Agune
Public Management Bureau of the State of São Paulo

Tatyana Souza
PRODEB - Data Processing Company of the State of Bahia

index

Preface

Preface

The first principle of the Web, proposed by the W3C Brazil, states that "the most valuable contribution of the Web is social. More important than its technological dimension, is that it is an environment for human communication, commercial transactions, opportunities and sharing knowledge; and it can only be universal, if available to all, regardless of the devices and software used, and, more importantly, regardless of culture, geographic location, intellectual or physical skills, socioeconomic conditions or level of education". Making access to the Web universal and exploring it to its full potential, is only possible if a democratic and pluralist governance model, focused on granting access to all and on its own technological development, is set forth.

Following resolution of the Brazilian Internet Steering Committee – CGI.br, from 2007, the Brazilian Network Information Center – NIC.br set up the W3C office in Brazil – the first in South America. The W3C is an international consortium aimed at leading the Web to the realization of its full potential, by creating standards and guidelines to enable its continuous development.

Assessing and monitoring the evolution of the Brazilian Web is one of the CGI.br's most recent roles; it is managed by the Brazilian office of the W3C and by the Center of Study and Research in Network Technology and Operations (CEPTRO.br), to produce information and indicators aimed at contributing to the understanding of the features and behavior of the Web, regarding accessibility and universality, and to monitor its own development.

We are thrilled to reveal that the results of the surveys published by the CGI. br are used by public managers, in the planning of governmental strategies and design of public policies to meet the needs of the Brazilian population;

by researchers, in the design of academic research; and by private companies, to monitor the Brazilian technological scenario.

The CGI.br presents the first edition of the Survey "Dimensions and Characteristics of the Brazilian Web: a study of the gov.br", an unprecedented study in scale and design, both in Brazil and internationally. Initially, the survey aimed only at the .gov.br domain, its approach scrutinizes the governmental Web. Later on, results for other Web domains shall be published. This survey will be conducted annually, in order to produce a historic series and to enable monitoring the development of the Brazilian Web.

This survey's results reveal characteristics of Brazilian domains, webpages and servers, which show how organizations design their webpages, regarding aspects of accessibility, universality, and the types of technology and documents used. The survey also presents features of Web servers regarding geolocation, synchronization of timestamps and preparation for the IPv6 protocol.

Hence, the CGI.br is proud to present the results and analysis of this survey, and we hope that these data play a key role in the understanding and development of the Brazilian Web.

Hartmut Richard Glaser
Managing Director – CGI.br

Introduction

Introduction

The Internet is probably the most sophisticated information and communication technology currently available to society, due to its structure and effect on the technological, social, economical and political instances of society. It is also the infrastructure required by one of its biggest and most widely known applications: the Web; which is accountable for popularizing the Internet to the extent that it is now mistaken by it. Internet and Web are, therefore, different concepts. The Web may be defined, generally, as the part of the Internet that can be assessed through browsers.

The impact of the Internet and the Web on society, individuals and on organizations has become the object of research beyond the specialized field of applied computing; it is also targeted by organizational and sociological studies. Essentially for being dynamic and unlimited, both physically and virtually, it must be widely known both to enable its free transformation and to ensure its availability, reliability and accessibility to all.

Hence, the Brazilian Internet Steering Committee – CGI.br and the Network Information Center – NIC.br, through the W3C Brazil and the Center of Study and Research in Network Technology and Operations – CEPTRO.br, have designed a project to increase knowledge and understanding of the Brazilian Internet: the **Web.br Census Project**. Conducted in partnership with the Logistics and Information Technology Bureau of the Ministry of Planning, Budget and Management (SLTI / MPOG), the Brazilian Association of State Bodies for Information and Communication Technologies (ABEP) and the National Institute of Web Science and Technology (InWeb), as well as, the methodological support of the Center of Studies on Information and Communication Technologies – CETIC.br, this project aims to produce indicators

that will contribute to the study and development of the Brazilian Web; the scope of which shall be discussed further in this publication.

Since mid 90s, the Brazilian Web has been growing profusely, both in the number of users and in the range of services and applications provided through the network. The increased use by the Brazilian population is striking: from 37 million users in 2005 to 65 million users in 2009. No less impressive is the change in citizens' behavior, who are using increasingly more transactional services through virtual environments, as shown by the CGI.br's surveys.

Aimed at shedding light on the development of the Brazilian Web, understanding its growth and potential, as well as, monitoring its transformation, the project and report hereby presented are the result of a team's relentless six months of hard work; overcoming the challenges of an innovative undertaking, unprecedented in the world in scope and objectives. Its results are available for appreciation and use as of now.

The methodological approach chosen, discussed further in this publication, involved working initially with the ".gov.br" domain. This report will show the characteristics and dimensions of the "governmental Web".

This report is divided into four parts. The first part, written by Antônio Marcos Moreiras, manager of the CEPTRO.br, consists in a description of the project. In this description, he discusses the technological challenges faced by the technical team to create a pioneering survey. Although there have been similar initiatives to the Web Census project, which have been useful for the design and development of the technological tool, the development of the Web itself and the questions we sought to answer make it unique.

The next topic presents the Methodological Aspects, it describes the methodology proposed and written by the InWeb, a scientific-technical partner of the project. This topic consists in a brief description of the methodology chosen and a summary of the data collection process.

The chapter before last presents the Analysis of Results, discussing the importance of each indicator and the results obtained compiling the data collected.

Finally, we present all of the indicators of the ".gov.br" domain with their respective tables and results, and a few individual results per region or State.

This study shall contribute to answering several questions, complementing and functioning as starting point for further initiatives. Examples of these questions are: How many websites are there in the Web.br? How big is the

Web.br and how does it grow? What types of technologies are used? Where are websites hosted? Are they hosted in Brazil or abroad? Are websites compliant with Web standards, such as HTML and CSS? Are websites accessible? Do they provide IPv6 support? Which technologies are used by server, pages, images, documents, videos, etc.? Are servers synchronized with the Brazilian legal time?

The proposal for this project foresees conducting it annually, and we hope that its ongoing results contribute to enable public, private and academic institutions to assess and monitor the development of the Brazilian Web; in addition to contributing to the design of governmental public policies for the electronic government.

Vagner Diniz
Manager – W3C Brazil Office

17

CHAPTER 1

THE TECHNICAL CHALLENGES OF STUDYING THE BRAZILIAN WEB
Dimensions and Characteristics of the Brazilian Web: a study of the gov.br

CHAPTER 1
The technical challenges of studying the Brazilian Web

CHAPTER 1

THE TECHNICAL CHALLENGES OF STUDYING THE BRAZILIAN WEB
Dimensions and Characteristics of the Brazilian Web: a study of the gov.br

The technical challenges of studying the Brazilian Web

At CEPTRO.br, our curiosity about the Web was awakened when we conducted a few simple studies on the geographic location of servers hosting ".br" domains, presented in the last two meeting of the TEP[1] Forum, an event tailored to Brazilian Autonomous Systems – networks comprising the Internet. These data revealed that over a third of the Brazilian Web servers were hosted abroad; which is far from ideal, as it results in higher latencies and costs for national access providers, despite the lower cost of hosting a website. We had to learn more. What kind of websites were these? Were they more or less important? Were they big or small? Were they tailored to the national or to a foreign market?

We had also conducted a project to promote the IPv6 in the country and monitoring its use in the Web would be a good indicator of the effectiveness of our initiatives. Similarly, we'd like to know if Web servers were synchronized to the appropriate time stamp, which is recommended by the CGI.br and enabled by a service provided by the NTP.br and the National Observatory. When the W3C office discussed with us the government's willingness and need to learn more and comply with website accessibility standards, we realized there was a real need to learn more

---

[1]   TEP – Traffic Exchange Point

CHAPTER 1

THE TECHNICAL CHALLENGES OF STUDYING THE BRAZILIAN WEB
Dimensions and Characteristics of the Brazilian Web: a study of the gov.br

about the Brazilian Web and decided to dedicate ourselves to the project. Nonetheless, little did we know about the dimension of the project we were signing up for, especially regarding the technical challenges in store.

The way the Web is structured, in itself, creates several challenges; such as, the fact that there isn't really a "Brazilian Web". The World Wide Web, as its name suggests, is a worldwide network. Limiting the scope of the study was the first challenge we faced. What would we do? Should we base our classification on the original language of the pages? If so, there was the challenge of determining the language and the risk of including websites from other Portuguese-speaking countries. Should we then base our classification on the geolocation of the servers used to host the Web? We knew beforehand that a large percentage of the ".br" domains were hosted abroad. We then decided to only take into account the Web comprised of ".br" domains, even though we know that Brazilian websites hosted under order domains would be excluded from the study. In order to minimize this problem, we included websites under other domains, found through redirections from URLs leading to a ".br" site.

The Web is a network where contents are interconnected through hypertext documents. It can be studied through successive data collection and analysis of pages from a preset list of websites. This search is made automatically by software known as crawlers, collectors or beaters. Nonetheless, the Web is not fully interconnected; although most of it is in fact connected, there are "islands" of different sizes which are not connected to the rest of the network. This means that the initial set of websites from which the search is conducted affects the final results, and finding the right initial set, as complete as possible, is a key step of the research process. When collecting data on the ".gov.br", for instance, ideally we would like to know the domains registered directly under the ".gov.br", as well as, the domains registered under State domains, such as ".sp.gov.br". The first are managed by the federal government, and represent the initial set; the latter are managed by state governments, and we relied on ABEP's (Brazilian Association of State ICT Bodies) support to identify them. Still, only 8 of the federative units (states) had sent us data when we started data collection, which required us to gradually complement data, as websites were found through Web search engines.

There are also traps in the data collection process: websites with infinite number of pages generated dynamically. Simple elements, such as calendars, generated automatically in a website, may create problems like this.

Size and depth restrictions must be set, which then create the risk of preventing data collection from bigger websites than these.

Another relevant aspect is what is referred to as the "Deep Web": the share of the network which requires user authentication for browsing, such as most relationship or community websites. This part of the Web is inaccessible to our methodology, and is therefore not included in the study. There is still the possibility of searching temporarily unavailable servers, or websites that do not have the robots.txt file, which determines if they can be searched or not by Web crawlers, or, websites blocked by this file for data collection.

We must also consider time, processing, connectivity and disk resources available to collect, store and process data. Even now that we've finished the first part of the study, it is still hard to estimate the resources needed for a census study of the whole ".br" Web. Data estimates, for instance, vary between 30 and 300 Tbytes, only taking HTML pages into account.

When studying the possibility of conducting the study, one of the first steps we took was to search previous similar surveys and tools. We found a few academic studies, even about the Brazilian Web, which became useful to us. We also found a few tools which could be of use in data collection processes. We studied three specific software designed for this purpose: Nutch, a data collector used to create search engines; Heritrix, used by the Web Archive, a project which keeps record of the history of the relevant part of the Web; and WIRE, which was used in one of the academic studies we found, and is designed specifically for Web studies. WIRE contains embedded analytical tools which we deemed relevant for our project: page size analysis, type of document, languages, ranking assessment, etc. We concluded that initiating the study with any of those would be beneficial to the design of a completely new tool. We chose WIRE, mainly because it already contains analytical functions embedded in the software.

We knew a few changes would have to be made to the original WIRE. For instance, we'd have to store Web pages in full to comply with standards; thus, HTML files collected, which would formerly be filtered to eliminate certain HTML tags and then stored in a large proprietary file format, would be stored in full in the filling system, in folders and sub-folders, in a similar format to the original websites. This change also made WIRE more scalable. Another modification was made to adjust the software's behavior in relation to redirecting, in order to adjust it to the definition of the Brazilian Web, discussed above.

CHAPTER 1

THE TECHNICAL CHALLENGES OF STUDYING THE BRAZILIAN WEB
Dimensions and Characteristics of the Brazilian Web: a study of the gov.br

CHAPTER 1

THE TECHNICAL CHALLENGES OF STUDYING THE BRAZILIAN WEB
Dimensions and Characteristics of the Brazilian Web: a study of the gov.br

Although WIRE had been used in a number of academic studies, several new implementations and behavioral adjustments had to be made, before we could deem it ready to be used in our study. For example, we've significantly improved the language detection tool to enhance its performance. In this context, it is also worth noting that we've normalized pages according to the RFC3986, the HTTP 1.1 treatment, with progressive data transfer, enhancement of page coding treatment, randomization of the order in which documents are downloaded and changes to the treatment of the list of websites to be searched, in addition to correcting several bugs. WIRE is a tool which is difficult to test. Achieving these results took months of programming, and a lot of data collected from significant parts of the Brazilian Web.

We aimed, through our Web study, to answer several questions which could not have been answered by the results of the analysis made by WIRE; such as the geolocation of servers, use of the IPv6 and the NTP, and compliance with HTML and accessibility (EMAG / WCAG) standards. These analyses could be incorporated into WIRE or implemented through a separate tool. We chose the second alternative, to avoid accidentally inserting new bugs into WIRE's code, with which we weren't completely familiar at the time. The tool created to analyze the aspects mentioned was temporarily named AnáliseInternet; additionally, it also stores WIRE data, as well as its own analyses, in a single database. When possible, we've re-used ready-made tools; such as when verifying compliance with the HTML standard, we used the validator created by the W3C, running locally and being consulted by AnáliseInternet. For accessibility tests we incorporated a popular software designed by the Brazilian government, the ASES.

At the end of the data analysis for this first partial study of the ".gov.br" Web, we concluded that we have a set of reliable tools which shall be very reliable in our future studies. We know, however, that there are limitations which must still be overcome; reason why we continue to implement changes and extensively test them. Among the modifications currently in progress, the following are noteworthy: analysis of right time using the NTP protocol, instead of the time provided by the HTTP, when possible; measuring the size of non-HTML objects found in webpages, such as images and videos, without downloading them, through HTTP HEAD searches; improving how we deal we the previously mentioned traps and speeding up data collection and analysis. Also, we need to automate part of the statistical analysis required to generate this report, possibly using tools such as the Data Warehouse and Data Mining.

We are still preparing to disclose the codes used, with free licenses, in order to ensure full transparency regarding our methodology and, hopefully, to attract other programmers and users of the software to cooperate to overcoming the remaining challenges.

Antonio M. Moreiras
Manager – CEPTRO.br

23

CHAPTER 1

THE TECHNICAL CHALLENGES OF STUDYING THE BRAZILIAN WEB
Dimensions and Characteristics of the Brazilian Web: a study of the gov.br

CHAPTER 2
Methodology

Methodology

Concepts and guidelines
for the Web.br Census

The word census comes from Latin and it means the "full count or listing" of a population of individuals or set of objects. Hence, a census is the final result of a specific count, which yields a set of statistical data regarding numerous variables in the population studied.

In order to conduct a census, it is paramount to clearly define the concept of the empirical units which will be analyzed, as well as, the technical-methodological procedures used to portray the population (determining profiles and limitations within the target population), data collection (individual characteristics to be determined) and result compilation (determined in accordance with the requisites of a tabular plane).

In this context, conducting a census depends on clearly knowing and determining, prior to the survey, the "limits within the population" within which individual objects censed must remain. In other words, in order to study the composition and size of the Brazilian Web, its domains, and consequently its limitations, must be previously determined.

Hence, the initial steps of consolidating a Census of the Brazilian Web, involved defining the concept of the units to be researched, as websites referred to by a name under the .BR domain. Thus, we've assumed that a specific content is part of the Brazilian Web, if the domain in the name of the website matches one of the following definitions:

1. It is part of a .BR domain;

2. It is not part of a .BR domain, but is re-directs from a .BR domain. For example, a multinational company registers a .BR domain under its brand, but this address re-directs to the company's main website under a .com domain.

Other aspects are also taken into account, such as links to documents on pages from .BR domains, even if they are hosted outside the scope of the domain hierarchy.

Nonetheless, one of the main obstacles faced in this census is the topology of the Web environment itself, which limits the technical ability to assess the real size and composition of what would be a "population of domains and technical objects". Besides issues regarding the complexity of determining the boundaries of the "Deep Web", the well known space of the Web.br, for instance, due to its inherent dynamics, seems unmanageable in terms of search and data collection methods, which makes the job of counting and determining domain and object profiles a lot harder, specially regarding determining their "popularity" among the population.

In this context of uncertainty regarding the dynamics, size and composition of the Web, it'd initially seem very daring to propose a robust methodology that could be broadly used to objectively measure the Brazilian Web. Therefore, it is important to note that the development and consolidation of this methodology is part of a broader, integrated process involving systematic planning of its different stages, setting a model for "identification', "collection", "validation" and "analysis" of all of the information available to determine a population of .br domains.

In other words, the importance of putting the concept of "census" to perspective and determining a "population of domains" is clear, and it must consolidate a methodology which will enable a strict count, in a near future. This shall be enabled by the development of this methodology and successive counts conducted as of now. This methodology is currently being developed and standardized for future use.

In order to argue the case for a "Web.br Census", we may compare it to the methodology consolidated for studies with the general population, in the field of Demography. Hence, a key point to be determined in the Web Census is the need to conduct periodic counts of the whole population of .br domains. Previous experiences with demographic experiments, determines that firstly we must define the concept of "population"; which also involves determining the set of techniques required to identify and count this population [1].

For instance, in order to count human populations, the first step is to determine the households where every member of the target population undoubtedly lives. Hence, counting may be made by city council's fiduciary real estate records. In this case, the census could be done by simply collecting data from fiduciary records from every city council in the country and effectively adding up all of the members of every household listed. In this population count (of individuals), it is assumed (common in demography) that every person is part of a household, that is, lives in a single household (there are exception and techniques to accommodate them).

Hence, when counting the Brazilian population, we have a set universe of individuals, based on the households and individuals associated with a living unit, and population count techniques depend of the quality of the study design and non-trivial organization of field work; or, effectively, the quality of the work of census officers in every existing (and dully registered) household when recording the number of residents who live there.

We may extrapolate this brief example, using the same logic, to a census of the Brazilian Web. The key point would be to determine the boundaries of the target population, even if they are only estimated, and never empirically verified. In this case, determining the reference "metrics" to analyze the objects collected at different points in time is paramount. Thus, we begin taking the information collected from official .br domain registries as reference of the target population; the limits of reference for the population are set by the first level .br domain. According to the logic presented for demographic studies, from the determination of a "digital mesh" of registered ".br" domains, we are able to determine the relationship between each individual observable object and its reference domain. Consequently, we have a population universe defined basically by the size of the set of first level domains and their composition of attributed objects.

However, this procedure in itself does not solve the whole problem of counting, because it doesn't provide a clear picture of the size of the Web; besides, it is widely known that effectively tracing an entire population (both the human population and a population of objects on the Web), that is, the so called "census coverage", perfect in every count, depends on a series of often uncontrollable aspects, which disable full count. For example, counting the number of individuals in a household may be impaired when an individual refuses to answer the interviewer's questions. Hence, even in demography, there are limitations to the ability of conducting a "perfect census" and, recurrently, the best demographic census sets an error margin of 2 to 8% of individuals/households in relation to the overall population.

Although there are limitations imposed by data collection procedures (quality of the crawlers/ officers), distribution of the population (isolated or unreachable objects) and the dynamic nature of the Web, there are also demographic methods specifically designed to correct census coverage errors, which could be adapted and used in the Web.br Census. In this case, the main issue is to define the "level of coverage" in relation to the estimated real population and, based on this parameter, to adjust to the effective size of the real population.

This report describes an initial effort towards establishing a methodology capable of estimating the "level of coverage", in order to consequently correct estimates of the size of the Web.br.

Thus, we reach our next challenge: establishing a robust methodology to estimate the level of coverage and the probable size of the target population.

At this stage, some of the challenges that have been studied, regarding adjusting the methodological approach to the scope of computing, and estimating the size of the Web.br, arise. In principle, there are two basic methods for calculating the coverage and estimating the effective size of a population: 1) estimating the coverage in a census, comparing demographics with a previous census; 2) using specific statistical methods to count rare populations;

1. Regarding estimating coverage for the results of two census, there were two immediate limitations to applying this to the Web.br Census. Firstly, a census (or at least identical attempt to count all of the .br domains) would have to have occurred in time T1, and a second one in a time T2. In the traditional demographic analysis of human populations, two censuses are used as parameters for the whole period of exposure of the target population, which through direct and indirect demographic analyses of individuals "joining" and "leaving" the general population, enable safely estimating the real size and composition of a population. Thus, in fact, this methodology aims to target a population at any given time and, from taking direct records of its numbers (that is, counting at two different times), using specific structural variables to estimate the direct and indirect effects of the transformations occurred to the population between T1 and T2. By comparing the results between the model of size and composition expected for the population and the population observed in the second count, it is possible to estimate the approximate size of the population at time T2. This requires specific knowledge of the structure of the population, that is, the main variables of the population (in human demographics these variables are age and

gender, as they represent the direct effects of joining and leaving the population - birth - deaths in the general population), which determine its structure and dynamics. Reproducing this strict methodology is not a possibility for the Web.br Census, as there are no structural variables of the population of domains and objects.

2. There's a second methodology which seems to be more appropriate and totally feasible for the Web.br Census. It involves using statistical techniques designed to estimate the size of rare populations or populations that are difficult to count. One of the most used population estimation techniques in biological sciences, used to estimate ecologic populations (as well as in demography to control the level of census coverage), is called "capture-recapture" [7,1]. Reproducing this technique basically involves determining the universe of .br domains and identifying (marking) them one by one. In fact, all it required in this experiment was an exclusive marker for every domain that came up on the first marking sample. Following a period of time enough to allow the population to change (for example, for new domains to be created), a second marking is conducted in accordance with the parameters set for the first procedure. Hence, we obtain two samples of the population of domains in which the individuals identified (.br domains and their linked objects) in one sample will not necessarily appear in the following sample, and vice-versa. A simple mathematical model is then used to estimate the probable overall size of the population, based on the likelihood of recapturing and not capturing domains in different samples of the same population (.br domains and their link objects). Consequently, from "capture-recapture" procedures, a thorough and statistically safe methodology is established to estimate the overall size of a population of known structure.

As previously stated, probably the biggest challenge for conducting a census is to determine census coverage parameters; as "coverage" refers to the degree of precision of the count in relation to the initial population. This count assumes that limitations (space and time) can be previously determined, so that the final result of the count and listing realistically portrays the overall number of "objects" comprising the initial population.

Generally when the population investigated is previously known (especially when its size at T1 is know), the census level of coverage can be estimated (for example, the level of precision of the census) from direct and indirect demographic techniques, comparing the structure of the popula-

tion at T1 and T2; however, this is the specific case of human populations, as clearly discussed in case 1.

The reduction in the population structure at T1 and T2 is due to two factors: effective changes to characteristics of the population over time, and the coverage error in counting/ listing objects and characteristics during the censuses at T1 and T2.

Regarding traditional demographic censuses, the coverage error is a direct result of undue omission or inclusion of specific households and the people who live in them, as well as, the people who live in specific occupied households regarded as the same in both censuses compared (T1 and T2). Regarding the Web.br, coverage errors are a direct consequence of undue omission of .br sites (and their linked objects) in a data collection at T1 and another one at T2.

Calculating the coverage error is paramount, as it reveals the degree of precision of assessments of the size of the Brazilian Web and, if needed, indicates the parameters needed to amend estimates. Hence, coverage error assessment is calculated from website omission indicators (equivalent to household) and objects (equivalent to people).

Accordingly, the method chosen to estimate these indicators is the so called Dual System Estimation [4,3,1], a methodology based in the "capture-recapture" technique mentioned in case 2. It requires similar data collection events taking place at two (or more) points in time, and statistic independence between events/ samples is paramount. Regarding the design of a methodology capable of measuring the size of the Web.br, ensuring that the seeds and the crawler (count tool) are launched independently from the same list of domains is paramount. At this stage, a new methodology is developed to validate the data collected, to estimate census coverage, taking into account information related to the number of registered domains (in this specific case, those registered under the ".gov.br" domain), errors found in data collection procedures and indicators related to the Web universe, such as increased volume of registered domains, changes to the size of information objects, among others. These results are presented in the forthcoming sections under results and outcomes.

As a novel methodology, our goal is to further develop it to ensure it is used over time, through new data collected and new techniques proposed to conduct a census of Web objects.

Below there's a brief description of the method used to estimate census coverage and its degree of precision. The method used to estimate "omis-

sion" indicators (coverage error) was the Dual System Estimation, based on "capture-recapture" techniques. Its use requires both data collection events to take place independently (data collection at T1 and T2) and requires information to be compared (shown in Table 2.1) as follows:

a is the number of units found at T1 and T2;

b is the number of units found only at T1;

c is the number of units found only at T2;

d is the number of unknown units, which were not included at T1 or at T2 (unknown); and

t is the overall number of units in the population.

DATA COLLECTION AT T1	DATA COLLECTION AT T2		
	TOTAL	UNITS INCLUDED	UNITS NOT INCLUDED
TOTAL	t	a + c	b + d
UNITS INCLUDED	a + b	a	b
UNITS NOT INCLUDED	c + d	c	d

Tabela 2.1: "Capture-Recapture" Information Table

Only **d** is, in fact, unknown data, as it represents the number of websites not captured at T1 and at T2 [4]. When the first "capture" at T1 is compared to the results of the "recapture" at T2, the coverage error (**d**) can be corrected from the probabilities known for **a**, **b** and **c**, that is, P(T1) = a+b and P(T2) = a+c, as a represents websites included at both data collection events; **b** is comprised of the websites collected at T1, but not recaptured; and c are the websites not captured at T1, but captured at T2.

Finally, from successive data collection events (which may be scaled to a large series in time), we are able to:

• Calculate the size of the coverage error (**d**); its output shall be refereed to as "omission rate";

• Estimate the size of the Web.br and number of pages related to it at a specific date;

• Estimate sizes at different points in time to assess the development of the Brazilian Web (such as, the growth and dynamics of the composition of the Web);

- Determine a variety of indicators, according to the different characteristics and composition of Web.br sites and pages.

Uses

In summary, until now we've been discussing the real possibilities of reproducing a demographic census in the population of .br domains. As previously mentioned, effectively conducting a population count/ listing which enables estimating the size and composition of the Brazilian Web, requires a non-trivial methodology still under development for consolidation.

In this initial effort, we've developed the applications needed to determine the population universe to be studied (.br domains and linked objects), the appropriate information collection and validation techniques, as well as, the methodology used to analyze and assess the size of the Web.br.

In order to achieve our initial goals, we had to redefine the study procedures and guidelines experimentally used to collect data exclusively from .gov.br domains. From this initial experiment, we tested a few procedures to estimate the population size of the Brazilian Web under .gov.br domains.

At this initial stage, our goal was restricted to trying, assessing and validating the pre-set methodological procedures. The "step-by-step" description of the procedure is presented below:

Determining the study question and the data required to answer it

Firstly, our goal was to estimate the size of the share of the Web.br under the .gov.br domain. In order to do that, we used information on the number of websites (gov.br) collected at two different times, as well as, the number of pages related to the set of websites collected.

Hence, there were two basic variables: 1) number of gov.br sites, and 2) number of pages related to the websites collected.

Secondly, as the data collection for the .gov.br took place at two different times (T1 and T2), we aimed to determine the number of websites and pages to which they referred. As a means of enabling use of the "capture-recapture"

method to estimate the size of the gov.br Web, both data collection events at T1 and t2 were compliant with the required parameters (events took place individually, and ensured seeds were launched randomly).

Brief description of the methodology

The method known as Dual System Estimation (DSE) – here referred to as the "capture-recapture" method – is commonly used by national of population statistics institutes, usually to check the quality of censuses [7, 1].

Estimates regarding population size derive from basic mathematical and statistical relationships, as long as a few elementary procedures are complied with: independent events, random distribution of objects in the population and the same random chance of an object being collected in all data collection events. Clearly, not all of these parameters can be complied with empirically in the Web universe. Particularly the random distribution of objects and their connections in the online universe (the topology of the online network is structured in a free scale, within the restrictions of a power law, and consequently, the distribution of vortexes and arches doesn't comply with normality patterns) [2,5,6].

Nonetheless, these are the first steps into using a methodology which requires adjustment to the Web universe, as previously done by Jianguo Lu and Dingding Li to estimate the size of the deep Web [6]. Hence, a plausible correspondence is revealed between estimates and the data collected.

In order to understand the "capture-recapture" method, an unknown population was studied (the size of the gov.br Web), and its objects (individuals) initially producing a set of n1 objects; and later, at a different time, a set of n2 objects. It is noteworthy that the listing involved an extensive collection of all objects in the target-population. Comparisons between the two sets of data collected (n1 and n2), produced an m set of duplicated objects, that is, objects captured at both times.

We've assumed that both data collection events were independent and the objects collected are equally likely to be collected at both data collection events. As shown by Alho and Spencer [1], the set of m duplicated objects follows a hipergometric probability distribution when the overall number of objects (captured and non-captured) in the population N are known. Indirectly, from the hipergometric probability distribution equation, it is possible

to estimate the overall size of the population N from a maximum similarity estimator, which makes the m set of objects captured as likely as possible.

Hence, the EN estimator is the N value which maximizes the probability of the observed set of duplicated objects m being true for the whole population. Here the likelihood estimator is:

$$EN = \frac{n_1 * n_2}{m}$$

In which **n1** and **n2** represent the set of objects collected at T1 and T2, and **m** represents the set of objects collected at both times.

The equation 3.1 shows the classic estimator of the "capture-recapture" method, established since Francis Bacon (1560) and reinvented numerous times, until consolidation by Laplace (1802) and its applied use in Biology by Petersen (1896), known as the Petersen estimator [7,1]. Additionally, other estimators were created to suit the empirical reality of data. In order to enable comparisons, we're also presenting an estimator derived from Petersen's and used by Lu and Li [6], the widely known Shumacher estimator, indicated for populations with uniform distribution; as it is the aim of this work group to acquire more in-depth knowledge on the methodology and to develop estimators suited to the empirical scenario of the .br Web.

Results

VALUES	WEBSITES	OK SITES	OK HTML PAGES
N1	18.911	12.891	6.334.054
N2	19.300	12.279	6.575.751
N1 -N2 = M	18.053	11.309	3.459.590
N1 + N2 = T	20.158	13.861	9.450.215

Tabela 2.2: Table of Results

Using the techniques presented in the methodology to estimate the number of websites (hosts), results obtained were:

- Consistency Ratio (correction factor) $R = \dfrac{EN}{t} = 1{,}0029$

- Classic Schumacher Estimator $EN = \dfrac{n_1 * n_2}{m} = 20.217$

Using the techniques presented in the methodology to estimate websites with valid pages, OK websites, results obtained were:

- Consistency Ratio (correction factor) $R = \dfrac{EN}{t} = 1{,}0097$

- Classic Schumacher Estimator $EN = \dfrac{n_1 * n_2}{m} = 13.996$

Applying the same techniques to estimate the number of webpages (valid HTML pages), results obtained were:

- Consistency Ratio (correction factor) $R = \dfrac{EN}{t} = 1{,}2740$

- Classic Schumacher Estimator $EN = \dfrac{n_1 * n_2}{m} = 12.039.334$

Analyses

It's noteworthy that both estimators used (Petersen and Shumacher) resulted in the same figure for the probable size of the population of sites and pages of the .gov.br Web. The consistency ratio (that is, the estimator of coverage of the data collection at different times) reveals the limitations of individual collection events. In other words, the sum of all of the objects collected at two different points in time adds up to 20,158 sites, 13,861 OK sites and 12,039,334 valid (OK) pages, under the .gov.br domain. Nonetheless, the consistency ratio, the correction factor for data collection coverage, indicates that the size of the population of .gov.br sites has been underestimated by 0.3% and by 0.97% for OK sites. Regarding valid HTML pages, results were underestimated by a much larger margin, 27%, due to variability of pages between both collection events. This is explained by the dynamic characteristic of the Web and also by the nature of its pages, which often vary technologically at every event, as a result of the concept of dynamic pages.

Final Comments

The concepts adopted as part of this methodology are in line with the objectives of the project and its current outcomes. Regarding the indicators produced and their analysis, it is worth noting that they are valid and relevant to the question we sought to answer, given existing limitations regarding data collection for analysis.

For future estimates and forecasts regarding the number of Web domains, the techniques used have not yet proven to be effective. This project's scenario is very dynamic and challenging, which requires further research to design new methods to progress from the results presented and produce forecasts for changes in the Brazilian Web. This reinstates the relevance of the strategy chosen to count domains until now in the analysis of the universe of the .gov.br Web, which shall be scaled to other domains in the following steps of our project.

Nonetheless, there is still interest in researching and developing new techniques which will enable, in addition to the counting method (Census), estimating and assessing future trends in the Brazilian Web; thus, enabling comparisons between analyses and ensuring higher quality regarding the study and assessment of the Web's qualitative and quantitative features.

Technical Team

InWeb – National Institute for Web Research

Bibliography

_____ [1] JUHA M. ALHO AND BRUCE D. SPENCER. Statistical Demography and Forecasting (Springer Series in Statistics). Springer, August 2005.

_____ [2] RICARDO BAEZA-YATES, CARLOS CASTILLO, and Efthimis N. Efthimiadis. Characterization of national Web domains. ACM Trans. Internet Technol., 7(2):9, 2007.

_____ [3] BEVERLEY CAUSEY. Dual system estimation based on iterative proportional fitting. Technical Report, Statistical Research Report - Bureau of the Census, Washington, USA,1984.

_____ [4] Brazilian Institute of Geography and Statistics. Methodologia do Censo Demografico 2000. Serie Relatorios Metodologicos, 25, 2003.

_____ [5] DANIEL GOMES E JOAO MIRANDA. Arquivo e Medicao da Web Portuguesa. In Pedro Isaias, editor, Proceedings of Ibero-Americana I ADIS WWW/ Internet 2008, Lisbon Portugal, December 2008.

_____ [6] JIANGUO LU AND DINGDING LI. Estimating deep Web data source size by capture-recapture method. Inf. Retr., 13 (1):70-95, 2010.

_____ [7] TRENT L. MCDONALD STEVEN C. AMSTRUP. Handbook of Capture-Recapture Analysis. Princeton University Press, USA, 2005.

CHAPTER 3
Analysis of Results

Analysis of Results

Defining the concept of Web

The World Wide Web, also known as the Web, or simply WWW, is an enormous compilation of pages, documents, data, applications and services interconnected through the worldwide web of computers; it is available to people from anywhere in the globe, at any time, and through several devices, from computers to mobile devices, such as mobile phones. This vast compilation is able to gather several digital contents, from pages to hypertexts, and even image files, figures, sounds, videos, programming codes, among others. Every file available through the Web is identified by a single, exclusive address, called URL, which stands for Uniform Resource Locator. An URL indicates the place where a digital file is located on the Web. Tim Berners-Lee's great invention was that by creating a Web localization system, he enabled documents to be accessible from anywhere on the Globe.

Each of these pools is identified by a domain name, usually known as website or site. When we browse the Web, we type in names to access the websites we want, such as: http://www.cgi.br, http://www.google.com.br, http://www.receita.fazenda.gov.br. It is noteworthy that although the Web is a powerful and broadly used application, it is only part of the network, an application.

Domain names have their special organization, not exclusive to the Web, and are globally structured in hierarchical levels. First level domains are called TLDs, acronym for Top Level Domains. There are several types of

TLDs, such as ".net", ".com", ".org", among others. Domains which identify their countries of origin, such as .br, are known as Country Code. Hence, the .br domain is a Country Code Top Level Domain – ccTLD; Brazil's first level domain. In Brazil, under this first level domain there are other levels, such as the ".gov.br", the ".com,br", and the ".org.br". Under these domains only, and following the hierarchical structure, domains are registered and created, such as the "governoeletronico.gov.br" domain. Thorough understanding of this structure of domains is paramount to the understanding of the analysis presented in this documented.

Dimensions and characteristics of the Brazilian Web

In order to determine the scope of this project, the Brazilian Web is defined as the network of public access websites identified by a .br domain, plus the websites to which there's automatic re-directing, via server, from a .br website, from which only the initial page was analyzed.

According to data from the Registro.br survey, the body in charge of domain name registry in Brazil, at the end of May 2010 there were 2.1 million domain names registered under this respective ccTLD, containing websites from several private, governmental and educational institutions, third sector organizations, independent professionals, individuals, etc. This requires studies on the specific context of domain names, such as ".com.br", ".org.br", ".nrt.br", ".gov.br" and other smaller groups, in order to measure their characteristics in the Web.

As the starting point for a broader study of the dimensions and characteristics of the .br, we have chosen to collect extensive data on the Brazilian governmental Web, from websites and pages under the .gov.br domain.

The main results and few of the conclusions from this compilation are presented in this report. This initial study is largely useful and shall become the foundation for a broader data gathering effort to assess the dimensions of the whole Brazilian Web under the ccTLD .br.

Dimensions and characteristics of the .gov.br

Data collection about governmental domains was conducted in October 2009 and 18,796 websites were found under the *.gov.br*, from the URLs searched. Determination of the overall number of websites began with the analysis of data from the following sources:

a. Domains identified as *.gov.br* (domains exclusive to the federal government). The list of these domains was provided by the Brazilian domain name registry authority, Registro.br, instructed by the Ministry of Planning, which is in charge of the use of *.gov.br* domains.

a. Domains identified under the *acronym -fu..gov.br*, registered by state data processing companies, and related to state governments.

a. Results of research and information searches, using search engines to complement previous information.

These different sources were compiled and served as seeds for a collection system. This compilation aimed to produce a registry which could include the largest possible number of governmental websites, in order to make it as similar as possible to a census of the Brazilian governmental Web. However, not all data processing companies from federal units, in charge of *acronym-.fu.gov.br*, were able to reply in time. This created an additional difficulty to conducting a census of the governmental Web, along with the difficulties inherent to the Web itself.

Regional participation in the composition of the Brazilian governmental Web

Data collection results revealed the participation of each of the five Brazilian regions, as well as of the federal government, in the composition of the Web; from analysis of the sub-domains corresponding to states, such as the sub-domain .sp.gov.br which was regarded as part of the Southeast region, and the gov.br as the federal government's. Two aspects of the participation were assessed:

a. the overall number of websites corresponding to each of the five regions of the country and the federal government;

a. The overall volume of bytes per region, as well as the federal government, taking only HTML documents into account (that is, images, videos and

other types of files were not compiled in the table below. See figure17 for an analysis of other types of files).

Table 3.1 presents a summary of the number of websites and the volume of bytes collected for each of the Brazilian regions. Although data collection found 18,796 websites, this analysis took into account those that had at least one HTML document. Overall, 11,856 websites were found to match these conditions.

REGION	VOLUME IN GIGABYTES	OVERALL NUMBER OF WEBSITES	REGIONAL PARTICIPATION IN THE OVERALL VOLUME IN BYTES	REGIONAL PARTICIPATION IN THE OVERALL NUMBER OF .GOV.BR WEBSITES
SOUTH	26	3.416	18%	29%
SOUTHEAST	32	3.358	22%	28%
NORTH	7	816	5%	7%
NORTHEAST	27	1.786	18%	15%
GOV.BR	38	1.668	26%	14%
CENTER-WEST	17	812	11%	7%
TOTAL	**148**	**11.856**	**100%**	**100%**

Table 3.1 – Number of websites and volume in gigabytes per geographic region

The percentage distribution of the data presented in Table 3.1 may be analyzed in the graph presented in Figure 3.1.

PARTICIPATION OF THE REGION IN THE TOTAL VOLUME OF GIGABYTES
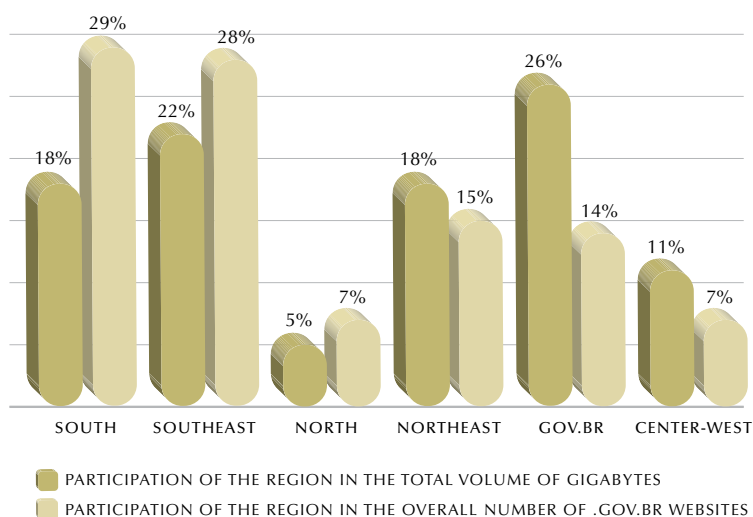PARTICIPATION OF THE REGION IN THE OVERALL NUMBER OF .GOV.BR WEBSITES

Figure 3.1 – Participation of regions and the federal government
in the composition of the governmental Web.

The largest byte volume index of Brazilian governmental websites was found in the Federal Government, which recorded 26% of the overall byte volume. Regarding the overall number of websites, the South region recorded 33% of the 18.7 thousand websites captured.

Regarding participation of units of the federation in the composition of the governmental Web, the domain **pr.gov.br**, from the State of Paraná, recorded the highest participation in the overall number of Brazilian governmental websites captured, approximately 17%; as shown in Figure 3.2. The Federal Government, comprising websites under the .gov.br domain, was next along with the State of São Paulo (**sp.gpv.br**); both of which recorded each 14% of the websites of the Brazilian governmental Web.

PARTICIPATION OF FEDERATIVE UNITS IN THE COMPOSITION OF THE BRAZILIAN WEB - % IN NUMBER OF WEBSITES

PARTICIPATION OF FEDERATIVE UNITS IN THE COMPOSITION OF THE BRAZILIAN WEB - % IN BYTE VOLUME
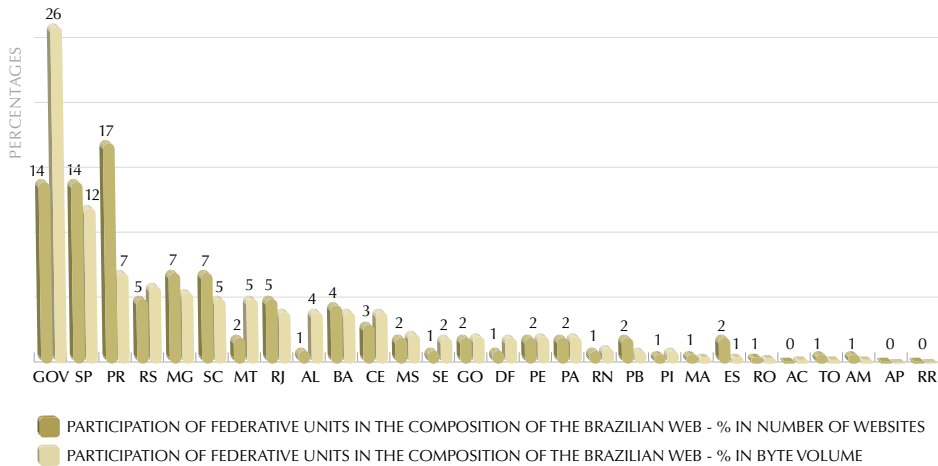
Figure 3.2 – Percentage distribution of the number of websites per FU,
including the Federal Government (gov).

The largest index regarding content in bytes of Brazilian governmental websites was recorded by the Federal Government, which recorded 26% of the overall byte volume; followed by the state of São Paulo, with 12%. Remaining states presented index lower than 10% of the overall byte volume.

The relationship between the number of websites, or occasionally the number of pages, and the volume in bytes may also be considered as a potential approach in the future. Figure 3.2 shows that states, in general, share both the same percentage share of the overall byte volume and the same number of websites of the governmental Web in Brazil, which indicates that the average byte size of websites is equivalent. Few states show different patterns. On one hand, Federal Government domains, present the highest byte volume in relation to websites; on the other, the State of Paraná (PR)

records the highest overall number of websites and overall participation in the Brazilian governmental Web. A closer look into these ratios may reveal that there are many governmental domains on the Web with no relevant content, or even governmental domains with excessive content, which indicates difficulties in accessing relevant public information.

Other languages in the governmental Web

Among the 3,182,202 pages that could have had its language identified by an automated analysis, 97% were in Portuguese. The software used to identify the language compares the text available on the pages with dictionaries with key words in Portuguese, English, Spanish and French, listing the words that appear in a given document. If the number of key-words for one of the languages tested is higher than a minimum reference number, and there's no ambiguity (more than one language with sufficient number of key-words), it is regarded as identified. Results for foreign languages are shown in Figure 3.3.
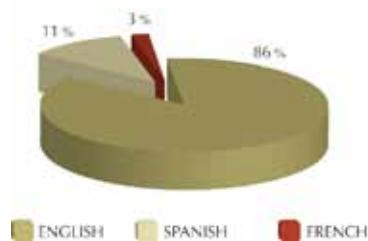

Figure 3.3 – Other languages found in the governmental Web

It is also noteworthy that in 47% of the 6.3 million HTML pages captured the language could not be identified by the procedure used. The sample is significant and may reveal that, essentially, the content of the Brazilian governmental Web is aimed at the internal public, and little information is available in other languages. On the other hand, without a semantic approach it is impossible to determine if the information in other languages is relevant, that is, content aimed at international relations, or tourism. Determining the type of information which could and should be available in other languages was not part of the scope of this survey.

Compliance with W3C's HTML standards

Analysis of compliance of HTML pages with W3C standards was made through validating software, designed by the consortium itself. As stated by the W3C and good practices of Web development, compliance with Web standards is a key indicator of the universality of access by any device connected to the Web, as well as, by any operating environment. The more compliant with the standards, the better the access offered by the page to any user, regardless of the device used and the operating environment. On the other hand, non- complaint pages may restrict access from certain devices and operating systems, which renders them non-universal. Especially for content from the governmental Web, compliance with standards and ensuring universality of access are mandatory principles and must be constantly taken into account and enforced.

This analysis involved counting the number of non-compliances found by the validating software. From the 6.3 million websites with HTML pages collected, approximately 91% presented more than one non-compliance; only 5% were fully compliant with the standard, and 4% could not be assessed; as shown in Figure 3.4.
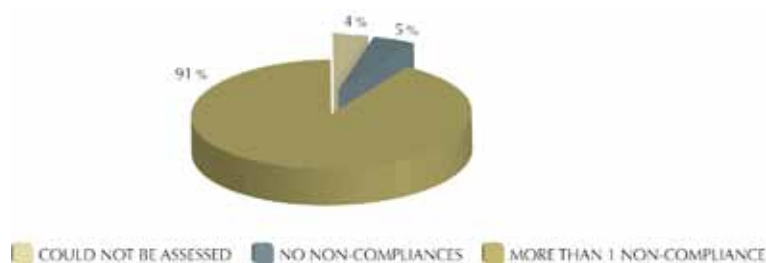

COULD NOT BE ASSESSED   NO NON-COMPLIANCES   MORE THAN 1 NON-COMPLIANCE

Figure 3.4 – Compliance with the W3C's HTML standards

Compliance with the ASES[1] accessibility standards

Compliance with accessibility standards aims to ensure universal access to Websites, even by disabled users. The accessibility criteria are divided into three lev-

---

[1]   T.N. ASES is the acronym in Portuguese for Site Accessibility Assessor and Simulator – tool designed by the Department of Electronic Government of the Ministry of Planning, Budget and Management and OSIP to enable assessment, simulation and correction of the accessibility of Web pages, sites and portals.

els of accessibility or compliance, according to the WCAG standard. The A level of compliance is regarded as mandatory in order for a website to be deemed accessible. The AA level of compliance comprises practices which should be adopted, in addition to the basic ones; and the AAA level of compliance comprises optional practices, which could greatly improve website accessibility.

The Brazilian government has created the e-MAG – Electronic Government Accessibility Model, in compliance with international standards; it consists in a set of recommendations which should be taken into account to ensure that granting the accessibility of Brazilian governmental sites and portals is a standardized and easily implemented process. It also created ASES, software that supports Web development by enabling the design of accessible websites, compliant with the e-MAG.

Analysis of compliance of HTML pages collected with accessibility standards followed the same protocol used in the ASES. The assessment process consists in measuring the level of compliance of pages.

From the 6.3 million HTML pages collected, 98% are not compliant in any way with the accessibility standards, as shown in Figure 3.5.

98%

2%

■ PRESENTED NON-COMPLIANCES
□ FULLY COMPLIANT

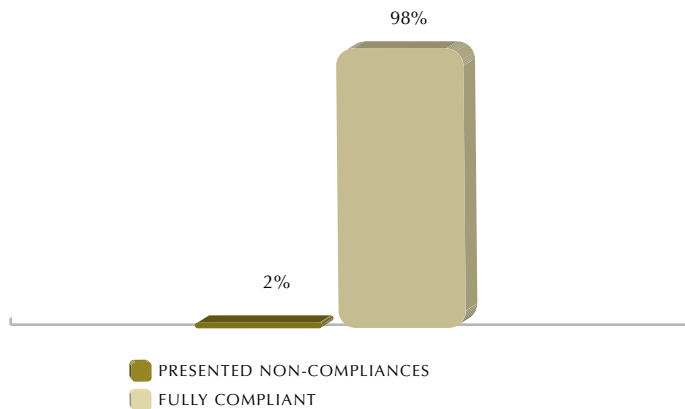Figure 3.5 – Proportion of pages compliant with the ASES accessibility standards

Technologies used to feed files to the governmental Web

Mapping out the feeding and information storage technologies used are among the noteworthy results obtained in the .gov.br data collection. Below, there's a chart showing the split between the main technologies used to feed files into the governmental Web (Figure 3.6).
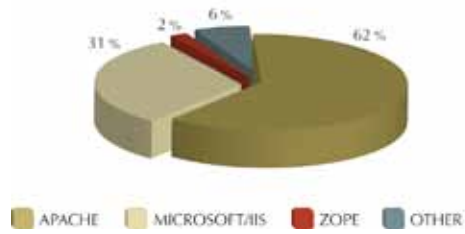
Figure 3.6 – Technologies used to feed files into the Web

Technologies based on open code software were found in over 60% of the pages captured. Proprietary platforms accounted for just over 30% of the document feed systems in the governmental Web.

Technologies used in file servers in the five Brazilian region

As shown in the following chart, pages captured under federative units' sub-domains for the South region account for the highest use of Web servers based on open code systems, as well as, the smallest percentage of use of proprietary systems; even when pages from the federal government are taken into account.
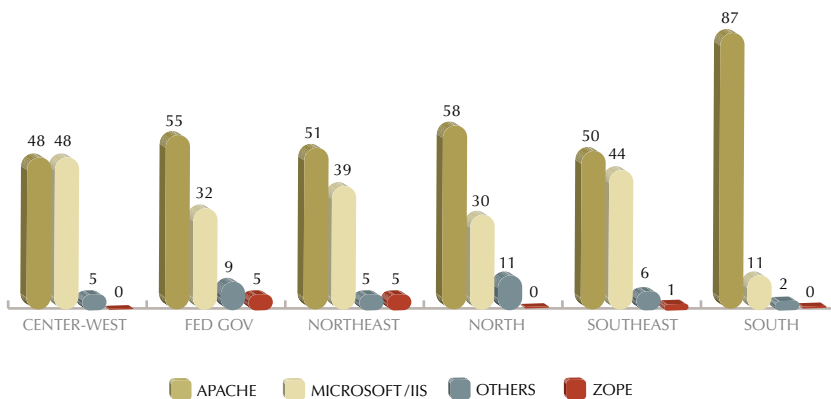
Figure 3.7 – Technologies used to feed files per region

Technologies used in file servers in the Brazilian States

The type of server platform used was also analyzed per federative unit. The chart in Figure 3.8 presents usage per type of Web server platform, in relation to the overall number of websites in the FU. Amapá and Paraná ranked first in relative use of open code systems to feed Web content. Regarding the used of proprietary software, the FD presents the highest use of this type of system to feed content.

Figure 3.8 – Use of the main Web servers by Brazilian FUs

We've still verified that the majority of Federal Government websites is hosted in Apache servers, which is an open technology.

Domains as sites structured in pages

Only domains with some form of verifiable content can be regarded as structured websites; thus, we assessed how many domains pointed to at least one HTML document, commonly known as HTML pages; that is, a file which can be interpreted by a Web browser.
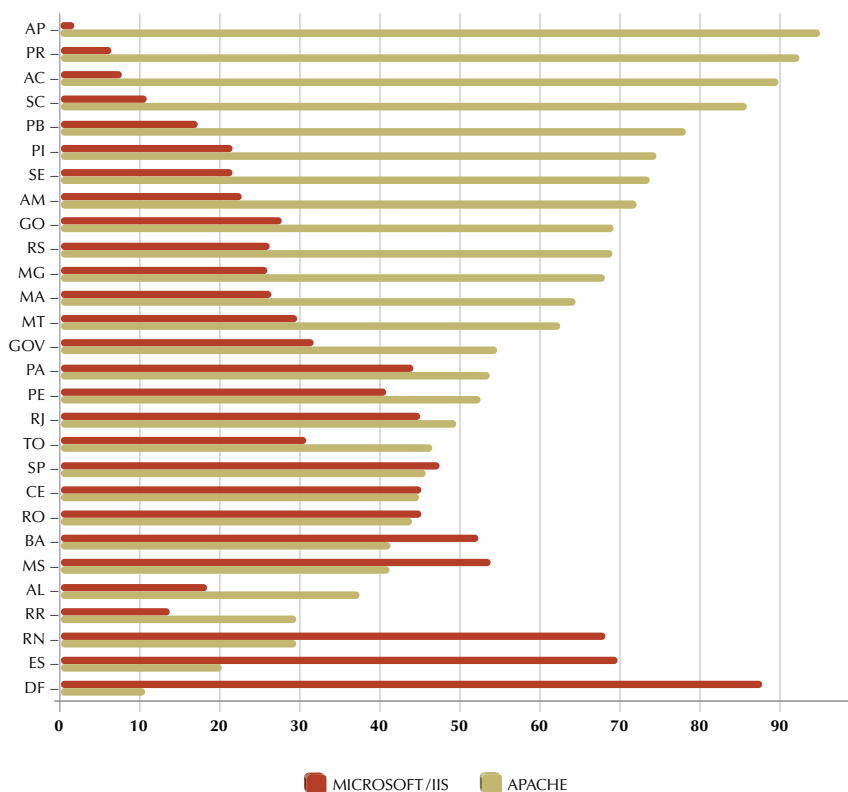
The term site, equivalent to Website or electronic site, refers to a set of HTML pages under the same name (here regarded as a full domain name) on the Internet. For example, http://www.prefeitura.sp.gov.br (everything following http:// and before the first "/" was regarded as an URL). For the purpose of the survey, URLs such as http://sítio.prefeitura.sp.gov.br/pagina1.html and http://sítio.prefeitura.sp.gov.br/calendario/evento.html are part of the same site, whereas, http://www.prefeitura.sp.gov.br is regarded as part of a different website.

From the initial figure of 18,796 websites, only 11,586 presented these characteristics. The overall number of digital files in each website was also determined.

Data collection revealed 7,947,607 digital files; from these, 6,331,256 were HTML documents, created or not by automated content generation systems. The remaining 1,616,351 non-HTML digital files were files in other formats, such as TXT, SWF, EXE, ZIP, RAR.

Also according to the data collected, the average number of HTML documents per site is 534 documents. These numbers together comprise a brief account of characteristics of governmental websites in the Brazilian Web.

Most used objects in pages of the governmental Web

The survey indicated that among all of the 192.2 million links found in .gov. br webpages, 89% correspond to some type of chart, 8.3% to some type of hypertext file and 2.5% a type of text file, such as .DOC, .PDF, .XML, .ODT, as presented in Figure 3.9.
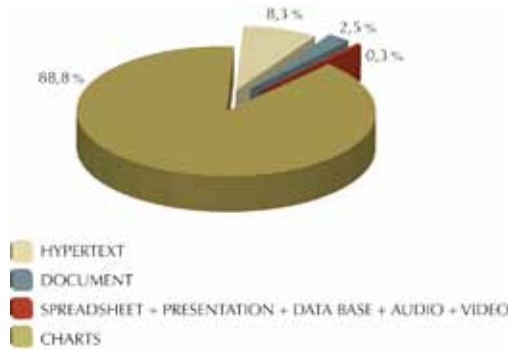
Figure 3.9 – Most frequent objects in the governmental Web
Base used: 192,247,032 links analyzed

Technologies used to make data and content available through the governmental Web

The technologies used to automatically distribute information through Brazilian governmental websites can be basically grouped into two types of technology: PHP and ASP. Technologies based on open code system, such as PHP, prevailed among all governmental pages collected; 70% of HTML pages collected had .PHP extensions.
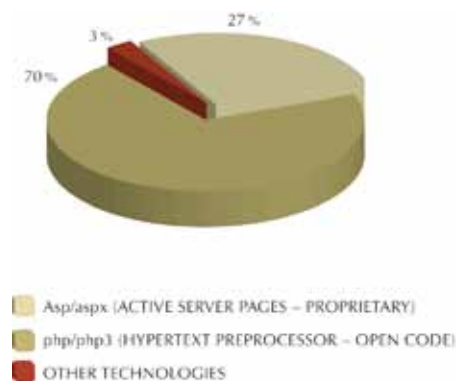
Figure 3.10 – Types of technology used to produce documents

From the links to graphic objects identified, approximately 99% of the images were in .GIF, .JPG, .PNG or .BMP format. Files in .PDF format represented 80% of all types of documents collected; whereas, files in .DOC format accounted for 13%. These results are illustrated in the charts in Figures 3.11 and 3.12.
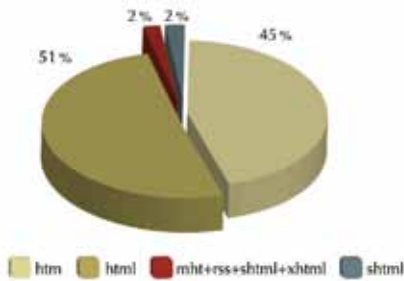
Figure 3.11 – Most used types of hypertext
Base used: 15,957,331 objects collected
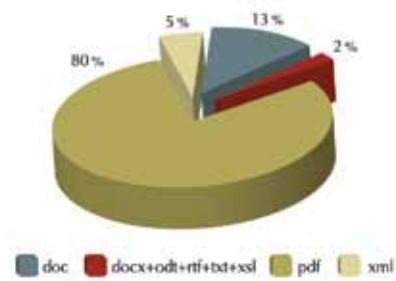
Figure 3.12 – Most used types of documents
Base used: 4,821,244 objects collected

It is, therefore, evident that .PDF is the preferred format for document publication.

Among editable content, the DOC format is the most common. The publication of editable format may not be a good practice, except in cases of publication of models used by users to send other information. It is also evident that the .XML format, used for data interoperation, is not widespread.

Brazilian web servers time synchronization

Synchronizing servers' clocks, work stations and other devices connected to the Internet is important to enable several applications to function properly, as well as, when logs need to be analyzed by applications to handle security incidents and related events. The NIC.br offers a public free service to provide the Brazilian Legal Time over the Internet, in partnership with the National Observatory, body in charge of setting the legal time. It also provides a website with information and guidelines on how to us this service: the http://ntp.br. The CGI.br formally recommends synchronization of all devices connected to a network, in its

resolution CGI.br/RES/2008/009/P (http://www.cgi.br/regulamentacao/resolucao2008-009.htm).

In order to assess the synchronization of servers hosting .gov.br websites, the time of its clocks was obtained via http protocol, and compared with the correct time. Results show that just over half of the servers are adequately synchronized, and the remaining present differences in relation to the Brazilian Legal Time of between 1 second and up to two hours, evidencing the need to revise settings.
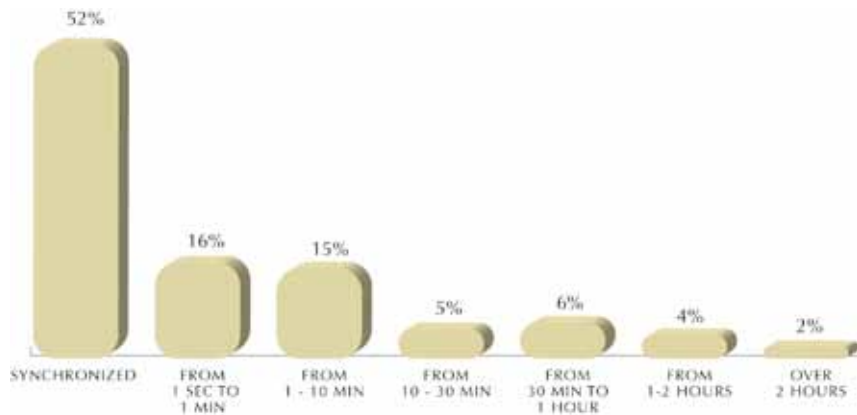
Figure 3.13 – Server time synchronization

Geolocation of IP addresses

The geolocation of servers was estimated using the GeoLite database, from MaxMind[2], consulted from server IP addresses.

Web servers hosted in areas far from users presented lower access speeds, due to the time it takes packages to travel. Hosting servers abroad, additionally, contributes to increasing the cost of accessing the Internet in Brazil, as it requires telecommunication operators to use more international loops, at high costs.

About 6% of gov.br websites are hosted abroad.

_____

[2] "This product includes GeoLite data created by MaxMind, available from http://maxmind.com/"
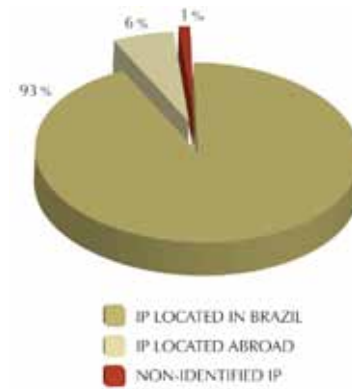
Figure 3.14 – Geolocalization of IP addresses
Base: 11,856 site with at least one HTML document

Average response time of
Brazilian servers

Server average response time in this study is regarded as the time servers took to respond to a simple (HEAD) http survey; including the time it took data packages to go back and forth plus server processing time. The test is, therefore, affected by the location of the network assessor at NIC.br's, in São Paulo.
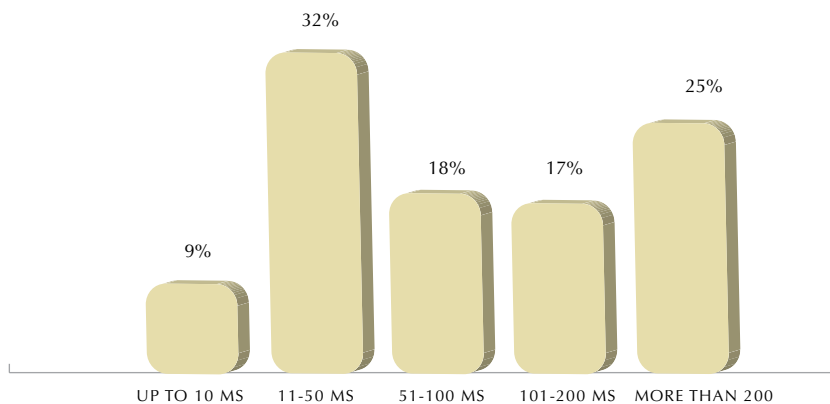
Figure 3.15 – Distribution of the average response time for websites hosted in Brazil

Briefly, the indicator may be regarded as website performance assessment from the perspective of users in São Paulo. It's noteworthy that websites hosted outside Brazil (Figure 3.16) presented worse results than the ones hosted

in the country. Among websites hosted in the country, approximately 59% recorded times of up to 100ms, which is regarded as a good result. Nonetheless, this study reveals room for improvement in the infrastructure of other servers and even in the infrastructure of the whole Brazilian Internet.
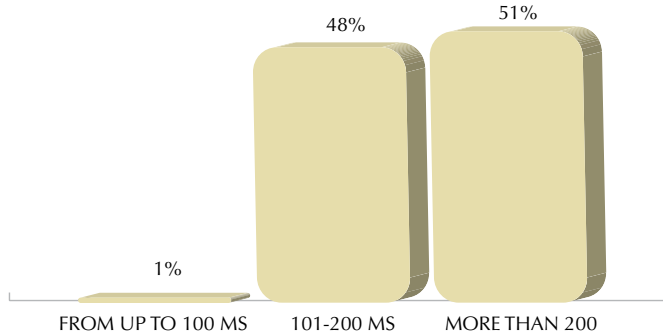
Figure 3.16 - Distribution of average response time for websites hosted abroad.

Brazilian governmental websites' response to IPV6 requests

The IP protocol is the technological foundation supporting the Internet; it is what enables intelligent use of the telecommunications infrastructure, which renders the Internet a much more ubiquitous, versatile and accessible network in relation to traditional telecommunications services. All Internet applications, including the Web, are supported, in a more basic level, by this protocol. The information on the Internet is divided in packages which travel independently across the network, and the IP enables them to find their way, identifying every device on the Internet with a number; the "IP address".

Currently, the most used version of the protocol is the IPv4, which has approximately 4 billion potential addresses; however, about 94% of these are already in use. With the IPv4, the Internet shall only be able to sustain its current growth rate for another year or two. Hence, a new protocol, the IPv6, shall be introduced. This protocol will function along with the IPv4 for a transition period, and replace it in the long term, enabling the Internet to continue to expand. We should all diligently implement it, as when the IPv4 runs out, there'll increasingly be more services and uses that'll only be able to communicate using only the IPv6.

The federal government seems to be aware of the seriousness of the issue, as it has produced a reference document for the e-PING: "Bodies of the federal public administration shall interconnect using the IPv4 and plan its future migration to the IPv6. New hires and network updates must foresee support to the coexistence of IPv4 and IPv6 protocols and products supporting both products.[3]"

No website was available via IPv6 protocol in the Web census for the *.gov.br.*

Technical team of the CETIC.br

Center of Studies on Information and Communication Technologies

---

[3]   Available at http://www.governoeletronico.gov.br/anexos/e-ping-versao-3.0. Accessed on the 23rd of July 2010.

CHAPTER 4
Indicators and Data Universe

Indicators and Data Universe

This section presents the indicators assessed only for the universe of .gov. br domains, presenting their definition and objectives, the methodology used to capture them and results. The indicators assessed at this stage of the project were:

A1: Overall size of the Brazilian Web: number of Websites and pages.

A2: Overall size of the Brazilian Web: size in Gigabytes.

B1: Proportion of Websites using the IPv6 protocol.

B2: Proportion of Websites using alternative IPv6 domain (ipv6.domain).

B3: Proportion of Websites which respond to Ping IPv6

B4: Proportion of Websites which respond to the GET command at the IPv6 address.

C1: Distribution of the use of languages in the Brazilian Web: proportion of languages.

E1: Proportion of Web pages compliant with the W3C's HTML standards.

F1: Proportion of Web pages compliant with Ases accessibility standards.

G1: Proportion of types of objects used in Web pages: percentage per type of object.

G2: Proportion of types of technologies used in Web pages.

H1: Average age (last update) of Brazilian Web pages.

H2: Proportion of dynamic pages in the Brazilian Web.

I1: Information on time synchronization of Brazilian Web's servers

I2: Information on average response time of servers in the Brazilian Web

In this project, variables related to the unit assessed (domain) are:

• URL (Uniform Resource Locator): URL is a compact string, defined by W3C standards, which indicates the location of a resource available on the Internet. All files available on the Web and files produced dynamically are captured via URLs. Web documents are identified by the substring http:// at the start of the URL.

• Web Page (HTML Document): a webpage is a resource written in HTML language. A page may have links to other types of resources, such as pdf files, image files and resources that trigger programs to run with specific parameters and generate dynamic pages.

• Website: is the set of one or more URLs associated with a domain.

This project comprises the following data universe:

• Universe of websites: all websites found during the data collection, regardless of whether they had one or more valid HTML pages or not.

• Universe of websites with valid HTML pages: there was at least one valid web page, that is, captured successfully (http-status code 2XX).

• Universe of valid HTML pages: universe of HTML pages requested and successfully obtained (http-status code 2XX).

There are indicators related to only one universe; others, to more than one. Regarding general indicators for Web size, they are general, and not related to another universe. Indicators are presented divided according to the universe to which they belong.

The following subsections present each of the indicators for the capture of .gov.br domains in detail.

A1: Overall size of the Brazilian Web – number of Web sites and pages

Definition of the indicator

Overall number of websites collected under the .gov.br domain and their respective pages, including pages outside the .gov.br domain, redirected from a .gov.br domain.

Objective

To determine the overall number of Web sites and pages in the Brazilian Web for the different data universes; that is, accounting for the overall number of websites and pages found, and the number of websites which issued a valid response to the HTTP request made (that is, issued response code 2XX).

Methodology

A crawler was run through the pages that met the previous definition, from an initial set of seeds supplied manually.

Presentation of Results

Tables 4.1, 4.2 and 4.3 present the results obtained for this indicator, using incrementally restrictive cuts. While Table 4.1 takes into account any objects found by the collector during the process, Table 4.2 limits results to the set of HTML pages and the servers in which they are contained. Next, Table 4.3 restricts this set to websites comprised in the .gov.br sub-domain.

NUMBER OF WEB OBJECTS AND SITES		
NUMBER OF WEBSITES	OVERALL NUMBER OF WEB OBJECTS	AVERAGE NUMBER OF OBJECTS PER SITE
18.796	7.947.607	422,84

Table 4.1: Indicator A1 - Number of Web sites and objects,
including all of the objects found by the collector.

NUMBER OF HTML PAGES AND WEBSITES		
NUMBER OF WEBSITES	OVERALL NUMBER OF HTML WEB PAGES	AVERAGE NUMBER OF HTML PAGES PER SITE
12.891	6.334.054	491,35

Table 4.2: Indicator A1 – Number of HTML pages and websites found by the collector
.

NUMBER OF HTML PAGES AND WEBSITES - .GOV.BR		
NUMBER OF WEBSITES	OVERALL NUMBER OF HTML WEB PAGES	AVERAGE NUMBER OF HTML PAGES PER SITE
11.856	6.331.256	534,01

Table 4.3: Indicator A1 – Number of HTML pages and websites
found by the collector, containing the .gov.br suffix.

SUBDOMAIN	SITES	SUBDOMAIN	PAGES	PAGES/SITE
ac.gov.br	39 (0,33%)	rr.gov.br	6.366 (0,10%)	163,23
rr.gov.br	51 (0,43%)	ap.gov.br	12.323 (0,19%)	241,62
ap.gov.br	58 (0,49%)	am.gov.br	28.091 (0,44%)	484,32
ro.gov.br	91 (0,77%)	ro.gov.br	41.342 (0,65%)	454,30
ma.gov.br	110 (0,93%)	ma.gov.br	48.330 (0,76%)	439,26
to.gov.br	117 (0,99%)	ac.gov.br	48.875 (0,77%)	417,73
pi.gov.br	121 (1,02%)	rn.gov.br	52.277 (0,83%)	432,04
se.gov.br	125 (1,05%)	to.gov.br	59.427 (0,94%)	475,41
am.gov.br	132 (1,11%)	es.gov.br	64.612 (1,02%)	489,48
al.gov.br	150 (1,27%)	pi.gov.br	68.905 (1,09%)	459,36
df.gov.br	160 (1,35%)	pb.gov.br	95.276 (1,50%)	595,47
rn.gov.br	170 (1,43%)	pa.gov.br	110.814 (1,75%)	651,84
mt.gov.br	189 (1,59%)	go.gov.br	121.225 (1,91%)	641,40
pb.gov.br	192 (1,62%)	ms.gov.br	129.391 (2,04%)	673,91
pe.gov.br	208 (1,75%)	df.gov.br	131.323 (2,07%)	631,36
pa.gov.br	218 (1,84%)	pe.gov.br	133.296 (2,11%)	611,44
go.gov.br	221 (1,86%)	se.gov.br	147.673 (2,33%)	668,20
ms.gov.br	242 (2,04%)	ce.gov.br	183.836 (2,90%)	759,65
es.gov.br	290 (2,45%)	ba.gov.br	185.756 (2,93%)	640,53
ce.gov.br	353 (2,98%)	al.gov.br	204.487 (3,23%)	579,28
ba.gov.br	467 (3,94%)	rj.gov.br	215.681 (3,41%)	461,84
rj.gov.br	572 (4,82%)	mt.gov.br	287.227 (4,54%)	502,14
rs.gov.br	605 (5,10%)	sc.gov.br	368.328 (5,82%)	608,80
sc.gov.br	791 (6,67%)	mg.gov.br	404.006 (6,38%)	510,75
mg.gov.br	832 (7,02%)	pr.gov.br	416.006 (6,57%)	500,00
sp.gov.br	1.664 (14,04%)	rs.gov.br	430.720 (6,80%)	258,84
pr.gov.br	2.020 (17,04%)	sp.gov.br	828.095 (13.08%)	409,94
gov.br	1.668 (14,07%)	gov.br	1.507.568 (23,81%)	903,81
TOTAL	11.856 (100,00%)	TOTAL	6.331.256 (100,00%)	534,01
(a)		(b)		

Table 4.4: Distribution of .gov.br Sites (a) and Pages (b) per Unit of the Federation

SITES	PAGES		SITES	PAGES
1 (0,01%)	>12000		70 (0,60%)	900 - 1000
29 (0,22%)	10000 - 12000		93 (0,80%)	800 - 900
181 (1,40%)	9000 - 10000		123 (1,05%)	700 - 800
61 (0,47%)	8000 - 9000		191 (1,64%)	600 - 700
38 (0,29%)	7000 - 8000		299 (2,56%)	500 - 600
55 (0,43%)	6000 - 7000		187 (1,60%)	400 - 500
51 (0,40%)	5000 - 6000		269 (2,30%)	300 - 400
71 (0,55%)	4000 - 5000		411 (3,52%)	200 - 300
108 (0,84%)	3000 - 4000		855 (7,32%)	100 - 200
176 (1,37%)	2000 - 3000		9.179 (78,61%)	<100
443 (3,44%)	1000 - 2000			
12.891 (100,00%)	TOTAL		11.677 (100,00%)	TOTAL
(a)			**(b)**	

Table 4.5: Distribution of pages per site;
results for all websites (a) and websites with less than 1000 pages (b).

A2: Overall size of the Brazilian Web – size in Gigabytes

Definition of the indicator

Sum of the size of pages under the .gov.br domains, according to the premises set for this project.

Objective

To calculate the volume occupied by Brazilian websites and webpages.

Methodology

A crawler was run through the pages that met the previous definition, from an initial set of seeds supplied manually.

Presentation of Results

Tables 4.6, 4.7 and 4.8 present the results obtained for this indicator, using incrementally restrictive cuts. While Table 4.6 takes into account any objects found by the collector during the process, Table 4.7 limits results to the set of HTML pages and the servers in which they are contained. Next, Table 4.8 restricts this set to websites comprised in the .gov.br sub-domain.

SIZE OF THE .GOV.BR WEB		
SIZE OF THE WEB (VOLUME IN GB)	AVERAGE SIZE OF WEBSITES (IN MB)	AVERAGE SIZE OF WEB PAGES (IN KB)
169,7	9,43	24,17

Table 4.6: Indicator A2 – Overall Size the Web under the .gov.br domain
(Size in GigaBytes)

VOLUME OF PAGES COLLECTED IN BYTES		
SIZE OF THE WEB (VOLUME IN GB)	AVERAGE SIZE OF WEBSITES (IN MB)	AVERAGE SIZE OF WEB PAGES (IN KB)
148,37	11,79	24,56

Table 4.7: Volume of HTML pages collected in bytes
and number of websites found by the collector

VOLUME OF PAGES COLLECTED IN BYTES - .GOV.BR		
SIZE OF THE WEB (VOLUME IN GB)	AVERAGE SIZE OF WEBSITES (IN MB)	AVERAGE SIZE OF WEB PAGES (IN KB)
148,33	12,81	24,57

Table 4.8: Volume of HTML pages collected in bytes and number of websites
where pages were found, taking only websites with the .gov.br suffix into account.

SUBDOMAIN	VOLUME (GB)	PAGES	VOLUME/PAGES (KB))
rr.gov.br	0,10 (0,07%)	6.366	15,73
ap.gov.br	0,18 (0,12%)	12.323	14,68
am.gov.br	0,48 (0,32%)	28.091	17,83
to.gov.br	0,81 (0,54%)	59.427	13,63
ac.gov.br	1,04 (0,70%)	48.875	22,02
ro.gov.br	1,06 (0,71%)	41.342	26,21
es.gov.br	1,09 (0,73%)	64.612	16,78
ma.gov.br	1,24 (0,84%)	48.330	26,21
pi.gov.br	1,91 (1,29%)	68.905	28,31
pb.gov.br	2,19 (1,48%)	95.276	23,07
rn.gov.br	2,22 (1,49%)	52.277	44,04
pa.gov.br	2,37 (1,60%)	110.814	22,02
pe.gov.br	2,75 (1,85%)	133.296	20,97
df.gov.br	2,86 (1,93%)	131.323	22,02
go.gov.br	2,89 (1,95%)	121.225	24,12
se.gov.br	2,90 (1,96%)	147.673	19,92
ms.gov.br	3,23 (2,18%)	129.391	25,17
ce.gov.br	5,05 (3,40%)	183.836	28,31
ba.gov.br	5,18 (3,49%)	185.756	28,31
al.gov.br	5,24 (3,53%)	204.487	26,21
rj.gov.br	5,50 (3,71%)	215.681	26,21
mt.gov.br	7,79 (5,25%)	287.227	28,31
sc.gov.br	7,85 (5,29%)	368.328	22,02
mg.gov.br	8,22 (5,54%)	404.006	20,97
rs.gov.br	8,73 (5,88%)	430.720	20,97
pr.gov.br	9,88 (6,66%)	416.006	24,11
sp.gov.br	17,62 (11,88%)	829.095	22,02
gov.br	37,96 (25,59%)	1.507.568	26,21
TOTAL	148,33 (100,00%)	6.331.256	24,12

Table 4.9: Distribution of the volume in Gigabytes under the .gov.br domain,
per unit of the federation

SITES	SIZE		SITES	SIZE
1 (0,01%)	>2.100 M		22 (0,18%)	48 -50 M
1 (0,01%)	1.300 -2.100 M		26 (0,21%)	46 -48 M
2 (0,02%)	1.000 -1300 M		22 (0,18%)	44 -46 M
1 (0,01%)	800 -1.000 M		20 (0,16%)	42 -44 M
2 (0,02%)	750 -800 M		28 (0,23%)	40 -42 M
1 (0,01%)	700 -750 M		31 (0,25%)	38 -40 M
2 (0,02%)	650 -700 M		23 (0,19%)	36 -38 M
1 (0,01%)	600 -650 M		27 (0,22%)	34 -36 M
8 (0,06%)	550 -600 M		35 (0,29%)	32 -34 M
4 (0,03%)	500 -550 M		34 (0,28%)	30 -32 M
11 (0,09%)	450 -500 M		51 (0,42%)	28 -30 M
11 (0,09%)	400 -450 M		44 (0,36%)	26 -28 M
9 (0,07%)	350 -400 M		63 (0,51%)	24 -26 M
22 (0,17%)	300 -350 M		36 (0,29%)	22 -24 M
54 (0,42%)	250 -300 M		69 (0,56%)	20 -22 M
60 (0,47%)	200 -250 M		71 (0,58%)	18 -20 M
99 (0,77%)	150 -200 M		74 (0,60%)	16 -18 M
138 (1,07%)	100 -150 M		105 (0,86%)	14 -16 M
227 (1,76%)	50 -100 M		120 (0,98%)	12 -14 M
12.237 (94,93%)	<50 M		146 (1,19%)	10 -12 M
			195 (1,59%)	8 -10 M
			433 (3,54%)	6 -8 M
			402 (3,29%)	4 -6 M
			732 (5,98%)	2 -4 M
			9.428 (77,05%)	<2 M
12.891 (100,00%)	TOTAL		12.237 (100,00%)	TOTAL
(a)			**(b)**	

Table 4.10: Distribution of volume in bytes per site;
for all websites (a) and for websites with less than 50 MB (b).

C1: Distribution of the use of languages in the Brazilian Web - proportion of languages

Definition of the indicator

Percentage of pages from the .gov.br domain divided according to a pre-set list of languages.

Objective

To obtain a percentage distribution of pages from the .gov.br based on their language.

Methodology

A crawler was run through the pages that met the previous definition, from an initial set of seeds supplied manually. The crawler used determines the frequency of words in a given language, based on pre-set dictionaries for a pre-set list of languages..

Results

Table 4.11 presents the distribution of pre-set languages on pages of the .gov.br domain: Portuguese, English, Spanish and French. It is noteworthy that the language could not be identified using this method for a number of HTML pages, which totaled 2,912,597 pages (47.8% of the overall number of pages).

	PERCENTAGE OF WEB PAGES FOR EACH LANGUAGE FROM A PRE-SET LIST OF LANGUAGES		
DISTRIBUTION OF LANGUAGES USED BY PAGES FROM THE DOMAIN	Portuguese	3.088.680	97,05 %
	English	80.726	2,54 %
	Spanish	10.623	0,33 %
	French	2.623	0,08 %

Table 4.11: Distribution of the language of pages from the .gov.br domain

E1: Proportion of Web pages compliant with the W3C's HTML standards

Definition of the indicator

Percentage of HTML pages in the Brazilian web ".br" compliant with the W3C standards, according to the type of document.

Objective

To assess the quality of HTML pages in the Brazilian Web ".br", regarding compliance with the W3C's HTML standard.

Methodology

A W3C document validator identifies the type of document and reveals the number of errors obtained according to type. The document validator issues the overall number of errors obtained from the analysis of compliance with the W3C regulations.

Presentation of Results

From the validation of Web pages made by the W3C validator, resulting figures were consolidated indicating the number of non-compliances found in the page.

Table 4.12 presents general webpage validation results, yielded by the W3C validation tool.

VALUE ISSUED BY THE TOOL	NUMBER	PERCENTAGE (%)
IMPOSSIBLE TO VALIDATE	267.137	4,24
PAGES VALIDATED	316.501	5,02
PRESENTED NON-COMPLIANCES > 0	5.717.315	90,74

Table 4.12: Number and percentage of pages in the governmental Web compliant with the W3C standards

NON-COMPLIANCES	WEB PAGES	
	NUMBER	PERCENTAGE (%)
≤ 10	1.212.156	21,20
≥ 10 e < 20	738.550	12,92
≥ 20 e < 30	673.568	11,78
≥ 30 e < 40	394.189	6,89
≥ 40 e < 50	332.285	5,81
≥ 50 e < 60	302.258	5,29
≥ 60 e < 70	241.251	4,22
≥ 70 e < 80	245.156	4,23
≥ 80 e < 90	183.045	3,20
≥ 90 e < 100	158.907	2,78
≥ 100	1.235.950	21,6

Table 4.13: Compliance of the governmental web to W3C standards
– Distribution of non-compliances per range

FU	W3C VALIDATION (HTML PAGES ON THE WEB)		
	NON COMPLIANCE	COMPLIANCE	% COMPLIANCE
Acre - AC	64.227	333	0,51
Alagoas - AL	212.728	4.724	2,17
Amapá - AP	21.055	969	4,39
Amazonas - AM	29.759	68	0,22
Bahia - BA	173.239	9.181	5,03
Ceará - CE	158.334	19.346	10,88
Distrito Federal - DF	119.812	3.553	2,88
Espírito Santo - ES	69.865	9.921	12,43
Goiás - GO	118.375	2.097	1,74
Maranhão - MA	51.023	277	0,53
Mato Grosso - MT	274.311	12.990	4,52
Mato Grosso do Sul - MS	135.955	1.219	0,88
Minas Gerais - MG	364.647	37.625	9,35
Pará - PA	135.466	2.230	1,61
Paraíba - PB	95.327	1.930	1,98
Paraná - PR	380.268	30.607	7,44
Pernambuco -PE	125.689	8.528	6,35
Piauí -PI	82.204	588	0,71
Rio de Janeiro -RJ	198.123	17.442	8,09
Rio Grande do Norte -RN	53.568	668	1,23
Rio Grande do Sul -RS	417.061	6.486	1,53
Rondônia -RO	72.109	10.251	12,44
Roraima -RR	6.538	32	0,48
Santa Catarina -SC	365.692	9.036	2,41
São Paulo -SP	799.181	50.790	5,97
Sergipe -SE	154.299	50	0,03
Tocantins -TO	83.248	1.361	1,6
Total	4.762.103	242.302	4,84
Other domains	955.212	74.199	7,2

Table 4.14: Compliance of the governmental web to W3C standards
– Per Federative Unit

CHAPTER 4 | gov.br
INDICATORS AND DATA UNIVERSE
Dimensions and Characteristics of the Brazilian Web: a study of the gov.br

F1: Proportion of Web pages compliant with Ases accessibility standards

Definition of the indicator

Percentage of valid HTML pages compliant with the accessibility standards set, regarding compliance standards A, AA, and AAA.

Objective

To assess the number of HTML pages compliant with accessibility standards WCAG 1.0 (W3C) and eMAG (Brazilian Government).

Methodology

In order to assess accessibility, the following were conducted:

- page data collection, as per the data collection procedures foreseen for this project.

- running of the ASES validator, which attributes an accessibility value (A, AA, AAA or non-compliance) for every HTML page collected.

In order to determine the means of assessing accessibility, a reference document was adopted, the WCACG 1.0 – Web Content Accessibility Guidelines 1.0. It was used to reveal compliance with accessibility A, AA and AAA. In order to achieve that, the following priority levels and definitions were adopted.

The work group attributed a priority level based on accessibility for every verification point. These levels are as described below:

• Priority 1: Items which Web content creators must comply with fully. If they don't, one or more groups of users will be prevented from accessing the information in the document. Fully complying with these items is a basic requisite to ensure that certain groups are able to access documents available on the Web.

• Priority 2: Items which Web content creators should comply with. If they don't, one or more groups or users will find it difficult to access the information contained in that document. Complying with these items removes significant barriers preventing access to documents available on the Web.

• Priority 3: Items which Web content creators may comply with. If they don't, one or more groups may find it difficult to access the information contained in the document. Complying with this type of item will improve access to the document stored on the Web.

Certain verification points specify a level of priority which could change under certain conditions (disclosed). Hence, accessibility compliance for Web pages was defined as follows:

• Compliance level "A": compliance with all priority 1 verification points;

• Compliance level "AA": compliance with all priority 1 and 2 verification points;

• Compliance level "AAA": compliance with all priority 1, 2 and 3 verification points;

• Non-compliance: no verification points were complied with in full; the Web page analyzed is non-compliant.

It's noteworthy that Web pages rated level A were not AA or AAA, likewise, AA pages were not AAA.

Presentation of Results

Table 4.15 presents the results obtained for the accessibility indicator, from the analysis of HTML Web pages from the .gov.br universe. Data presented was produced from 6,279,206 HTML pages. A further 54,848 pages (0.86%) were not classified, as the validator did not yield any of the expected results.

Table 4.17 presents the results obtained for the accessibility indicator, from the analysis of HTML pages, dividing the data into Federative Units (FUs) identified in the HTML pages' URLs (http:// ...fu.gov.br).

	COMPLIANCE WITH PRIORITY LEVELS (HTML WEB PAGES) – DIVIDE BY UNIT			
	PRIORITY 3	PRIORITY 2	PRIORITY 1	NON COMPLIANCE
QUANTITATIVE	39.440	14.662	71.628	6.153.476
PERCENTAGE	0,63	0,23	1,14	98,00

Table 4.15: F1 Indicator – Compliance of governmental web pages
with priority levels (HTML webpages)

FU	Priority levels (HTML webpages)				
	3	2	1	Non Compliance	
Acre -AC	6	5	8	65.213	99,97%
Alagoas -AL	0	0	1.773	216.815	99,19%
Amapá -Ap	60	0	2	22.154	99,72%
Amazonas -AM	12	0	23	32.373	99,89%
Bahia -BA	9	1	380	167.227	99,77%
Ceará -CE	1.762	0	171	182.977	98,95%
Distrito Federal -DF	210	0	638	129.955	99,35%
Espírito Santo -ES	157	6	458	86.622	99,29%
Goiás -GO	1.053	0	17	120.113	99,12%
Maranhão -MA	67	16	427	51.092	99,01%
Mato Grosso -MT	88	0	2	287.222	99,97%
Mato Grosso do Sul -MS	7.093	0	607	130.144	94,41%
Minas Gerais -MG	1.111	300	407	406.274	99,55%
Pará -PA	94	159	6.468	131.378	95,13%
Paraíba -PB	78	2	15	97.531	99,90%
Paraná -PR	5.537	49	4.255	407.748	97,64%
Pernambuco -PE	75	8	123	134.526	99,85%
Piauí -PI	32	12	17	82.997	99,93%
Rio de Janeiro -RJ	298	228	893	215.409	99,35%
Rio Grande do Norte -RN	36	12	31	54.442	99,86%
Rio Grande do Sul -RS	4.922	4319	1.913	438.921	97,52%
Rondônia -RO	380	824	169	81.014	98,33%
Roraima -RR	1	0	3	6.575	99,94%
Santa Catarina -SC	4.393	23	506	376.020	98,71%
São Paulo -SP	7.489	655	4.370	848.759	98,55%
Sergipe -SE	29	2	66	156.452	99,94%
Tocantins -TO	1.940	104	496	83.847	97,06%
Total	36.932	6.725	24.238	5.013.800	98,66%
Other domains	2.508	7.937	47.390	1.139.676	95,17%

Table 4.16: Compliance with priority levels
(HTML webpages) – divided by Federative Unit

G1: Proportion of types of objects used in Web pages - percentage per type of object

Definition of the indicator

Percentage of the types of objects used in pages of the Brazilian Web ".br", divided by category (images, scripts, videos, etc.).

Objective

To obtain a distribution of the types of objects used in pages of the Brazilian Web ".br", classified according to pre-set categories (images, scripts, videos, etc.).

Methodology

Pages were collected using a Web crawler, taking both the URL of every page and URLs present in each page into account. All extensions were converted to lower case. The taxonomy of types of documents was taken from e-Ping, Interoperability Standards of the Electronic Government, Reference Document Version 2.0, 11 of December 2009.

Presentation of Results

Results are presented in Table 4.17.

GROUP	NUMBER PER GROUP	%	TYPE	NUMBER PER TYPE	%
HIPERTEXTO	15.957.331	8.30	htm	7.220.067	45.25
			html	8.089.407	50.69
			mht	5.128	0.03
			rss	61.829	0.39
			shtml	318.241	1.99
			xhtml	1.731	0.01
			xml	260.928	1.64
DOCUMENT	4.821.244	2.51	doc	627.197	13.01
			docx	225	0.00
			odt	8.516	0.18
			pdf	3.864.991	80.17
			rtf	24.766	0.51
			txt	32.932	0.68
			xml	260.928	5.41
			xsl	1.689	0.04
SPREADSHEET	156.623	0.08	ods	331	0.21
			xls	156.240	99.76
			xlsx	52	0.03
PRESENTATION	28.533	0.01	odp	158	0.55
			ppt	28.302	99.19
			pptx	73	0.26
DATA BASE	6.531	0.00	csv	6.405	98.07
			myd	63	0.96
			myi	63	0.96
CHARTS	170.538.106	88.71	bmp	118.730	0.07
			gif	660.78.840	38.75
			gif	66.078.840	38.75
			jpeg	51.888	0.03
			jpg	28.281.181	16.58
			odg	24	0.00
			png	9.915.715	5.81
			svg	480	0.00
			tif	12.408	0.01
AUDIO AND VIDEO	472.158	0.25	avi	7.964	1.69
			mid	20	0.00
			mp3	412.649	87.40
			mp4	49.252	10.43
			mpg	1.519	0.32
			ogg	251	0.05
			wav	503	0.11
TOTAL	192.247.032	100.00			

Table 4.17: Number and percentage of objects found in HTML pages, per type of document

G2: Proportion of types of technologies used in Web pages – percentage per type of technology

Definition of the indicator

Percentage distribution of the types of technology used in pages of the Brazilian Web (.br).

Objective

To obtain a distribution of the types of languages used in pages of the Brazilian Web ".br", based on a pre-set list of values (PHP, ASP, ASPX, JSF, JSP, etc.).

Methodology

Determining the technologies used is challenging, because pages collected don't necessarily contain information on the technology used to produce them. One of the alternatives adopted in this study was to check the extension of the file in the URL.

In order to determine the technologies used, the starting point was a dictionary of 406 file extensions. We processed the file in valid pages (OK), verifying in each URL listed in that file the presence of a valid extension. A valid extension must occur before the URL's "?" ad from the last "/" preceding this "?"´. The process then checks, by matching standards, the presence of extensions in the chain within the URL's "/" and "?". For extensions with the same root (e.g. asp and aspx), the longest extension is taken into account.

The starting point of this methodology was to determine potential extensions, as an assessment of the technologies used. In this case, a list of 406 file extensions was found at the website: http://www.file-extensions.org. Based on the extensions, the URL for every page collected was analyzed in order to determine which extensions were present in that URL.

Once extensions had been determined, there were three possible scenarios. The first scenario: no extension was found in the URL, which made it impossible to determine the technology used. The second scenario: there a single extension, in which case the technology is immediately determined. The third scenario: more than one URL is found in association with the URL, which makes it necessary to determine the most relevant extension.

Two criteria were used to determine the most relevant extension. The first criterion is the place where the extension occurs in the URL, based on the assumption that the extension present in file positioned first determines its technology of origin. This criterion was validated in a significant number of cases. The second criterion is for extensions occurring in the same position, and it determines that the longest one is selected, as it is naturally the most descriptive; for instance, when the extensions php and php3 are found in the same position in an URL, as php is part of php3. In this example, the extension selected would be php3, as it is the longest and most distinctive.

The last step of this methodology was to select, among the extensions identified, those which were related to technologies. This process was performed manually, verifying the extension which effectively occurred and their respective descriptions.

Presentation of Results

Results are presented in Table 4.18.

TEC	AMOUNT	%	DESCRIPTION
asp	868.183	24,34	ASP script, Page
aspx	94.017	2,64	ASP.NET script, page
cfm	10.003	0,28	Cold Fusion Markup
cgi	6.186	0,17	Common Gateway Interface
com	73	0,00	Common Object Module
dbc	1	0,00	Database Container
dll	6.515	0,18	Dynamic Link Library file
do	38.690	1,08	Oracle Application Server
exe	4	0,00	Executable file
js	1	0,00	JavaScript file
jsp	53.260	1,49	JAVA Server page
nsf	86	0,00	IBM Notes
php	2.483.013	69,61	PHP script, page
php3	335	0,01	PHP version 3 script file
py	1.424	0,04	Python
sql	115	0,00	Structured Query Language Data SQL
wsp	5.346	0,15	SharePoint Services Solution
TOTAL	3.567.252	100,00	

Table 4.18: Number and proportion of technologies used in the Brazilian Web

Definition of the indicator

Average age of the pages in the Brazilian Web ".br" based on the date of the webpage's last update.

Objective

To obtain the average age of pages in the Brazilian Web, based on their last update date.

Methodology

The software used to collect data always attempts to record the age of the pages collected, showing the difference between the date and time when an URL is collected and the date and time shown by the server, through the last update of that page. This information (last update date of every page) is not provided by all servers, nor for every type of content. As this is not compulsory information, it is often not available through data collection. Additionally, errors in the time set in Web servers may lead to errors in date and time information provided by them. In the data collected, pages for which last update date was not provided remained with unknown record of age.

The age of the pages with age information was determined in days to simplify the analysis.

Presentation of Results

Results are presented in the table below.

AVERAGE AGE OF PAGES	AVERAGE AGE OF PAGES OF THE BRAZILIAN GOVERNMENTAL WEB
	656 days

Note: From the 6,331,256 pages, just over 10% (614,770) presented valid age information.

GROUP	OVERALL NUMBER OF PAGES	OVERALL NUMBER OF PAGES WITH AGE INFO	%	AVERAGE AGE IN DAYS
ac.gov.br	48.875	458	1	225,82
al.gov.br	204.487	111.374	54	491,84
ap.gov.br	12.323	2.803	23	450,68
am.gov.br	28.091	581	2	562,51
ba.gov.br	185.756	6.321	3	455,69
ce.gov.br	183.836	10.955	6	635,69
df.gov.br	131.323	7.806	6	779,38
es.gov.br	64.612	4.557	7	1.242,94
go.gov.br	121.225	19.341	16	538,9
ma.gov.br	48.330	3.320	7	1.545,25
mt.gov.br	287.227	19.946	7	1.150,8
ms.gov.br	129.391	2.765	2	712,44
mg.gov.br	404.006	28.967	7	377,94
pa.gov.br	110.814	4.129	4	868,85
pb.gov.br	95.276	2.715	3	477,63
pr.gov.br	416.006	17.593	4	664,57
pe.gov.br	133.296	4.206	3	1.385,63
pi.gov.br	68.905	6.322	9	176,09
rj.gov.br	215.681	16.132	7	399,39
rn.gov.br	52.277	3.598	7	678,95
rs.gov.br	430.720	24.370	6	685,02
ro.gov.br	41.342	7.389	18	270,44
rr.gov.br	6.366	306	5	601,63
sc.gov.br	368.328	18.909	5	767,16
sp.gov.br	828.095	100.790	12	600,65
se.gov.br	147.673	1.291	1	1.986,37
to.gov.br	59.427	1.053	2	1.594,59
Total for states	4.823.688	427.997	9	607,55
Other .gov.br	1.507.568	186.773	12	768,31
Total for the .gov.br	6.331.256	614.770	10	656,24
Other pages	2.798	1.110	40	440,07
TOTAL	6.334.054	615.880	10	655,85

Table 4.19: Age of the pages in the Brazilian governmental Web per state

H2: Proportion of dynamic pages in the Brazilian Web

Definition of the indicator

Percentage of pages regarded as dynamic in the Brazilian governmental Web (.gov.br). A dynamic page is generally a reference to the use of server-side programming language, such as PHP, ASP, JSP, ColdFusion, among others in the development of intranet and extranet sites and applications. Its is called dynamic because its generated in execution time, producing static content which a user visualizes when he/she makes a request, via HTTP request.

Objective

To determine the percentage of dynamic content produced from pages of the Brazilian governmental Web (.gov.br).

Methodology

The collector used has a set of internal rules to determine if a page is dynamic or static. These rules assess the extension used for the content file (for example, extensions such as .jsp or .php are associated with dynamic files), as well as, the presence of parameters related to the URL.

Presentation of Results

Results are presented in the table below.

PERCENTAGE OF DYNAMIC PAGES	PERCENTAGE OF DYNAMIC PAGES IN THE BRAZILIAN GOVERNMENTAL WEB
	74,8 %

GROUP (XX.GOV.BR)	OVERALL NUMBER OF PAGES FOUND	OVERALL NUMBER OF DYNAMIC PAGES	PERCENTAGE
.ac.gov.br	48.875	48.297	99
.al.gov.br	204.487	65.766	32
.ap.gov.br	12.323	8.832	72
.am.gov.br	28.091	14.660	52
.ba.gov.br	185.756	164.832	89
.ce.gov.br	183.836	89.914	49
.df.gov.br	131.323	101.298	77
.es.gov.br	64.612	52.169	81
.go.gov.br	121.225	97.212	80
.ma.gov.br	48.330	42.655	88
.mt.gov.br	287.227	243.994	85
.ms.gov.br	129.391	106.624	82
.mg.gov.br	404.006	335.168	83
.pa.gov.br	110.814	79.986	72
.pb.gov.br	95.276	87.897	92
.pr.gov.br	416.006	353.119	85
.pe.gov.br	133.296	113.213	85
.pi.gov.br	68.905	65.701	95
.rj.gov.br	215.681	184.889	86
.rn.gov.br	52.277	47.317	91
.rs.gov.br	430.720	370.529	86
.ro.gov.br	41.342	34.350	83
.rr.gov.br	6.366	6.010	94
.sc.gov.br	368.328	315.023	86
.sp.gov.br	828.095	637.510	77
.se.gov.br	147.673	128.428	87
.to.gov.br	59.427	53.480	90
Total for states	4.851.779	3.863.533	80
Other .gov.br	1.479.477	873.269	59
Total for the .gov.br	6.331.256	4.736.802	75
Other pages	2.798	1.499	54
Total	6.334.054	4.738.301	75

Table 4.20: Percentage of dynamic pages in the Brazilian governmental Web

B1: Proportion of Websites using the IPv6 protocol

Definition of the indicator

Percentage of websites in the universe of servers of pages of the Brazilian Web (".br") operationally structured to respond according to protocol IPv6.

Objective

To determine the current number of web servers operationally structured to communicate using the IPv6.

Methodology

Specific requests for the IPv6 protocol to Web servers. A response indicates that the server is structured to respond to IPv6 requests.

Presentation of Results

Just 4 (four) of the 12,891 hosts in the first screening responded to the IPv6 protocol. Hosts are listed in Table 7.1. It's noteworthy that, according to the Web Census Project's definition, all hosts captured as a result of redirections from the Brazilian governmental Web are regarded as part of this subdivision of the Web. Therefore, the hosts listed below, although not comprised in the .gov.br dub-domain, in this Project are regarded as part of the Brazilian governmental Web.

URL
www.google.com
www.lacnic.net
www.itu.int
www.terra.com.br

Table 4.2.1 presents the list of the hosts which responded to an IPv6 protocol request

Therefore, the proportion of hosts which responded to the IPv6 request in the Brazilian governmental Web is 4/12,891 = 0.031%.

B2: Proportion of Websites using alternative IPv6 domain (ipv6.domain)

Definition of the indicator

Percentage of websites in the universe of servers of pages of the Brazilian Web ("br"), which respond to an IPv6 request in an alternative domain (ipv6.domain).

Objective

To determine the current number of web servers tested to communicate using the IPv6.

Methodology

Specific software issues IPv6 requests to Web servers, and their responses are used to determine if servers are operationally structured to respond to the IPv6 request.

Presentation of Results

No website responded to the request to alternative domain. Hence, the proportion is zero.

B3: Proportion of websites which respond to ping IPv6

Definition of the indicator

Percentage of websites in the universe of servers of pages of the Brazilian Web ("br") that respond to a PING command in IPv6 addresses.

Objective

To determine the current number of active Web servers which respond to a PING command in the IPv6 domain.

Methodology

Specific software issues IPv6 requests to Web servers, and their responses are used to determine if servers are operationally structured to respond to the IPv6 request.

Presentation of Results

URL
www.itu.int

Table 4.22: Hosts that respond to the PING via IPv6 protocol

Results revealed that only one website of the Brazilian governmental Web responded to a PING in IPv6 addresses. The proportion is, therefore, 1/12,891 = 0.008%.

B4: Proportion of Websites that respond to the GET command to the IPv6 address

Definition of the indicator

Percentage of websites in the universe of servers of pages of the Brazilian Web ("br"), which respond to a GET command to port 80 of the address with IPv6 protocol.

Objective

To determine the current number of active Web servers that can respond to a GET command to port 80 of the IPv6 address.

Methodology

Specific software issues IPv6 requests to Web servers, and their responses are used to determine if servers are operationally structured to respond to the IPv6 request.

Presentation of Results

URL
www.google.com
www.lacnic.net
www.itu.int

Table 4.23: Hosts that respond to the HTTP GET via IPv6 protocol

I1: Information on time synchronization of Brazilian Web servers

Definition of the indicator

Estimate of the difference of the synchronization of time in servers of the Brazilian Web (.gov.br), in relation to the official time worldwide, known as CUT (Coordinated Universal Time).

Objective

To estimate the degree of synchronization of servers of the Brazilian governmental Web (.gov.br) in relation to the official global time.

Computers are able to synchronize time using a time server and a protocol. Generally, the protocol used is the NTP (network Time Protocol), which converts time into a comprehensible language to the server. This mechanism is key to ensure that transactions, as well as, other communications on the Web are duly recorded.

Methodology

A HTTP request was issued to the server using the HEAD method. The server responded with the date and time under the Date field. The RTT (round-trip time) of the request was measured. The time issued by the server was estimated as follows: half of the time of the RTT was added to the time provided in the Date field. It's noteworthy that the RTT is measure in milliseconds and the Date field is issued in seconds. Estimate of the synchronization was calculated by the calculation of the module of the difference between the estimated time of the server and the time recorded by the

machine conducting the survey, synchronized via NTP. The final result is yielded in seconds.

Presentation of Results

Results revealed 12,836 valid times, from the 12,891 servers surveyed. Table 4.24 present the main statistics regarding the times found.

MIN	MEDIAN	AVERAGE	MAX	CV	Q1	Q2	Q3	P90	P97	P99
0	1	150.766	336.045.799	40	0	1	170	3.435	5.750	11.860

Table 4.24: Difference in numbers between server time and CUT in seconds

FU	MEDIAN	AVERAGE	MAXIMUM	CV
AC	150	1.040	10.534	2,17
AL	31	4.279	79.706	2,36
AM	4	1.670	57.403	3,93
AP	1194	1.212	3.953	0,61
BA	2	3.415	585.462	9,80
CE	1	1.193	203.281	9,35
DF	2	432	3.903	1,72
ES	0	1.193	236.686	11,76
GO	0	929	29.154	2,8
MA	284	2,36e+06	252.563.955	10,29
MG	1	551	37.261	3,38
MS	3507	2.778	68.881	1,94
MT	1	2.364	191.897	6,52
PA	25	2.537	348.840	9,30
PB	195	1.088	22.034	2,21
PE	3	15.583	2.631.634	11,88
PR	0	122.887	247.348.217	44,82
RN	0	437	8.462	3,39
RO	21	1.962	50.529	2,90
RR	0	867	8.355	2,37
RS	0	1.010	86.762	4,60
SC	10	1,03e+06	246.533.888	14,96
SE	1	676	11.389	2,92
SP	6	152.961	215.427.138	34,94
TO	321	3.134	82.862	3,05

Table 4.25: Indicator I1 – Synchronization statistics
per unit of the federation: time in seconds

I2: Information on average response time of servers in the Brazilian Web

Definition of the indicator

Average response time of Web servers, regarding all websites identified in the data collection for the .gov.br.

Objective

This indicator aims to shed light on the average response time of Web servers in the .gov.br universe.

Methodology

In order to collects data on the response time of a specific website, a specific request is made to a specific Web server, and the time it takes the server to respond is recorded (in milliseconds).

Thus, it is a simple method, which only yields the approximate time needed to access the server; however, it provides a general assessment of this quality indicator of the response time to a request.

Presentation of Results

Table 4.26 presents the results obtained for the average response time indicators for websites in the .gov.br universe.

Data presented in the table refer to the 12,871 websites which had at least one HTML page issuing a valid response. A further 20 websites (0.15%) were not included in the analysis, as there was no response to the time request (which happens because the website server doesn't accept this type of request or due to an unavailability error).

RESPONSE TIME (MILLISECONDS)	WEBSITES	
	NUMBER	PERCENTAGE
≤ 10	1.101	8,55
> 10 e ≤ 50	4.111	31,94
> 50 e ≤ 100	2.278	17,70
> 100 e ≤ 200	2.143	16,65
> 200 e ≤ 300	1.184	9,20
> 300 e ≤ 400	534	4,15
> 400 e ≤ 500	311	2,42
> 500 e ≤ 600	274	2,13
> 600 e ≤ 700	176	1,37
> 700 e ≤ 800	152	1,18
> 800 e ≤ 900	100	0,78
> 900 e ≤ 1000	77	0,60
> 1000	430	3,34

Table 4.26: Average response time of websites

AVERAGE	MIN	MAX	MEDIAN	STANDARD DEVIATION	CV	Q1	Q2	Q3	P90	P97	P99
190,20	1	8313	71	368,78	1,94	27	71	201	475	1049	1595

Table 4.27: Response Time – Statistical Analysis

D2: Proportion of countries hosting Brazilian websites

Definition of the indicator

Percentage of websites in the Brazilian Web, per host country of the website.

Objective

To obtain a percentage distribution of websites from the Brazilian Web, according to host country.

Methodology

The geolocation of servers was determined using the GeoLite database, from MaxMind, in which IP addresses were searched.

Presentation of Results

Table 4.28 shows this indicator's results, obtained from the universe of 11,856 websites with .gov.br domains and at least one valid HTML document.

LOCATION	BRAZIL	ABROAD	UNIDENTIFIED
PROPORTION OF WEBSITES HOSTED	93%	6%	1%

Table 4.28: Proportion of servers hosted in Brazil and abroad