dimensões e características da

Web

brasileira: um estudo do .gov.br

2010

cgi.br
Brazilian Internet Steering Committee

nic.br
Brazilian Network Information Center

Comitê Gestor da Internet no Brasil – CGI.br

Coordenador
Augusto Cesar Gadelha Vieira

Conselheiros
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

Diretor Executivo
Hartmut Richard Glaser

Núcleo de Informação e Coordenação do Ponto BR – NIC.br

Diretor Presidente
Demi Getschko

Diretor Administrativo e Financeiro
Ricardo Narchi

Diretor de Serviços e Tecnologia
Frederico Neves

Diretor de Projetos Especiais e de Desenvolvimento
Milton Kaoru Kashiwakura

COORDENAÇÃO GERAL

NIC.br / CEPTRO - Centro de Estudos e Pesquisas em Tecnologias de Redes e Operações
Antonio Marcos Moreiras
Heitor de Souza Ganzeli
Pedro Hadek

NIC.br / CETIC - Centro de Estudos sobre as Tecnologias da Informação e da Comunicação
Alexandre Barbosa
Juliano Cappi
Robson Tavares

W3C - Escritório Brasil
Carlinhos Cecconi
Vagner Diniz
Orípide Cilento Filho

Assessoria de Comunicação
Caroline D'Avo

PARCEIROS

InWeb - Instituto Nacional de Ciência e Tecnologia para a *Web*

Adriano C. Machado Pereira
Cristina Duarte Murta
CEFET-MG - Centro Federal de Educação Tecnológica de Minas Gerais, Departamento de Computação

Altigran da Silva
UFAM - Universidade Federal do Amazonas, Departamento de Ciência da Computação

Dimitri Fazito de Almeida Rezende
Eduardo Luiz Gonçalves Rios-Neto
UFMG - Universidade Federal de Minas Gerais, Departamento de Demografia

Dorgival Olavo Guedes Neto
Renato Ferreira
Wagner Meira Jr
UFMG - Universidade Federal de Minas Gerais, Departamento de Ciência da Computação

Ministério do Planejamento, Orçamento e Gestão
Cláudio Muniz Machado Cavalcanti
João Batista Ferri de Oliveira
SLTI - Secretaria de Logística e Tecnologia da Informação

ABEP
Dayse Vianna
PRODERJ - Centro de Informação e Comunicação do Estado do Rio de Janeiro

Kátia Bruno
CEPROMAT - Centro de Processamento de Dados do Estado de Mato Grosso

AGRADECIMENTO ESPECIAL AOS COLABORADORES:

Gustavo da Gama Torres
José Maria Leocádio
SERPRO - Serviço Federal de Processamento de Dados

Isabele dos Passos Omena
José Nilo Martins Sampaio
ATI Agência de Tecnologia da Informação, Governo do Estado de Pernambuco

Nicolau Reinhard
FEA - Faculdade de Economia e Administração, Universidade de São Paulo

Orion Borba
CIASC Centro de Informática e Automação do Estado de Santa Catarina

Paulo Maia
Caixa Econômica Federal

Roberto Agune
Secretaria de Gestão Pública do Estado de São Paulo

Tatyana Souza
PRODEB - Companhia de Processamento de Dados do Estado da Bahia

Índice

Prefácio

Prefácio

O primeiro princípio da *Web*, proposto pelo W3C Brasil, afirma que "o principal valor da *Web* é o social. Mais do que tecnológico, este é um ambiente de comunicação humana, de transações comerciais, de oportunidades para compartilhar conhecimentos e, para ser um ambiente universal, deve estar disponível para todas as pessoas, independentemente dos equipamentos e *softwares* que utilizem, principalmente da cultura em que inserem, da localização geográfica, das habilidades físicas ou mentais, das condições socioeconômicas ou de instrução". A universalidade da *Web* só pode ser garantida e aprofundada com um modelo de governança democrático e pluralista que tenha foco no acesso por todos e na sua própria evolução tecnológica.

Acompanhando deliberação do Comitê Gestor da Internet no Brasil – CGI.br, em 2007, o Núcleo de Informação e Coordenação do Ponto BR – NIC.br instalou o escritório do W3C no Brasil – o primeiro na América do Sul. O W3C é um consórcio internacional com a missão de conduzir a *Web* ao seu potencial máximo, criando padrões e diretrizes que garantam a sua evolução permanente.

Medir e acompanhar a evolução da *Web* brasileira é uma das mais recentes atividades do CGI.br conduzida pelo escritório do W3C no Brasil e pelo Centro de Estudos e Pesquisas em Tecnologia de Redes e Operações (CEPTRO.br), a fim de se produzirem informações e indicadores que contribuam para o entendimento das características da *Web* e do seu comportamento nas áreas de acessibilidade e universalidade, além de acompanhar a sua própria evolução.

É com satisfação que comprovamos a utilização dos resultados das pesquisas divulgadas pelo CGI.br por gestores públicos na construção de estratégias governamentais e na elaboração de políticas públicas que atendam às necessidades da população brasileira, por pesquisadores na elaboração de pes-

quisas acadêmicas e por empresas privadas no acompanhamento do cenário tecnológico brasileiro.

O CGI.br apresenta a primeira edição da Pesquisa "Dimensões e características da *Web* brasileira: um estudo do .gov.br", pesquisa inédita na sua forma e extensão no Brasil e também fora do País. Inicialmente, a pesquisa dedicou-se apenas ao domínio .gov.br, cujo olhar faz um raio-x da *Web* governamental. Posteriormente, serão divulgados também os resultados sobre os demais domínios da *Web*. Essa pesquisa será realizada anualmente, com objetivo de gerar uma série histórica e de poder acompanhar a evolução da *Web* brasileira.

Os resultados dessa pesquisa revelam características dos domínios, páginas *Web* e servidores *Web* brasileiros, que mostram como as organizações desenvolvem as suas páginas *Web*, considerando aspectos de acessibilidade, universalidade, tipos de tecnologias e tipos de documentos. A pesquisa também mostrará características dos servidores *Web*, considerando aspectos de geolocalização, sincronização de *timestamp* e preparação para protocolo IPv6.

Portanto, é com satisfação que o CGI.br divulga o resultado dessa pesquisa e a análise sobre o seu significado, com a expectativa de que esses dados sejam importantes ferramentas de compreensão e evolução da *Web* brasileira.

Hartmut Richard Glaser
Diretor Executivo - CGI.br

Introdução

Introdução

A Internet é provavelmente a mais sofisticada tecnologia de informação e comunicação atualmente disponível para a sociedade, em função da sua forma de organização e de seus impactos nas esferas tecnológicas, social, econômica e política. Ela é também a infraestrutura necessária para uma de suas maiores e mais conhecida aplicação: a *Web,* grande responsável pela popularização da Internet, a ponto de hoje ser confundida com esta. Internet e *Web* são, portanto, conceitos distintos. A *Web* pode ser definida, grosso modo, como a parte da Internet acessada por meio de navegadores, ou *browsers*.

O impacto do uso da Internet e da *Web* na sociedade, nos indivíduos e nas organizações tornou-se objeto de pesquisa, extrapolando o campo especializado da computação aplicada, e atingindo áreas de estudos organizacionais e sociológicos. Por ser essencialmente dinâmica e sem fronteiras, tanto do ponto de vista físico como virtual, é importante que seja conhecida em detalhes, tanto para assegurar sua livre transformação quanto para permitir sua disponibilidade, confiabilidade e acessibilidade por todos.

Assim, o Comitê Gestor de Internet do Brasil – CGI.br e o Núcleo de Informação e Coordenação do Ponto BR – NIC.br, por meio do W3C Brasil e do Centro de Estudos e Pesquisas em Tecnologias de Redes e Operações – CEPTRO.br, criou mais uma iniciativa para um melhor conhecimento e entendimento da Internet brasileira: o **Projeto Censo da *Web* .br**. Realizado em parceria com a Secretaria de Logística e Tecnologia da Informação do Ministério do Planejamento, Orçamento e Gestão (SLTI / MPOG), a Associação Brasileira de Entidades Estaduais de Tecnologia da Informação e Comunicação (ABEP) e o Instituto Nacional de Ciência e Tecnologia para a *Web* (InWe*b*), ainda

com o apoio metodológico do Centro de Estudos sobre as Tecnologias de Informação e Comunicação – CETIC.br, esse projeto tem como objetivo criar indicadores para contribuir para o estudo e evolução da *Web* brasileira, cujo escopo é definido mais adiante.

Desde meados dos anos 90, a *Web* brasileira tem mostrado acentuado crescimento, tanto no número de usuários como no leque de serviços e aplicações oferecidos por meio da rede. É flagrante o avanço de seu uso pela população brasileira: de 37 milhões de usuários, em 2005, passou a aproximadamente 65 milhões, em 2009. Igualmente impressionante é a mudança de comportamento do cidadão, que utiliza cada vez mais serviços transacionais em ambientes virtuais, conforme mostram as pesquisas do CGI.br.

Para compreender o fenômeno do desenvolvimento da *Web* brasileira, entender o seu crescimento e potencial, bem como acompanhar a sua transformação, esse projeto e seu relatório agora apresentados são um esforço de seis meses de contínuo trabalho e de superação de uma equipe diante de uma empreitada inovadora, única no mundo em seu escopo e objetivos, cujos primeiros resultados poderão ser apreciados e utilizados a partir de agora.

Como opção metodológica apresentada adiante, trabalhamos inicialmente o domínio ".gov.br". O que veremos nesse relatório são as características e as dimensões da "*Web* governamental".

Esse relatório divide-se em quatro partes. A primeira, escrita por Antônio Marcos Moreiras, gerente do CEPTRO.br, será a descrição do projeto, pelo qual ele apresenta os desafios tecnológicos enfrentados pela equipe técnica diante de um levantamento pioneiro. Embora já tenha havido iniciativas parecidas com esse projeto Censo da *Web*, as quais foram úteis para a concepção do projeto e desenvolvimento da ferramenta tecnológica, a própria evolução da *Web* e as respostas buscadas às perguntas propostas tornaram-no único.

O tópico seguinte será Aspectos Metodológicos, uma descrição da Metodologia proposta e escrita pela InWeb, parceira técnico-científica do projeto. Esse tópico constará de uma breve descrição da metodologia escolhida e um sumário do processo de coleta de dados.

O penúltimo capítulo apresentará uma Análise dos Resultados, buscando explicar a importância de cada indicador definido e compreender os resultados obtidos com a compilação das informações coletadas.

Finalmente, apresentamos todos os indicadores do domínio ".gov.br" com suas respectivas tabelas de resultados, com alguns recortes por Estado ou por região.

Esse estudo ajudará a responder várias questões, complementando e servindo de subsídio para outras ações. Por exemplo: Quantos sítios há na *Web*.br? Qual o tamanho da *Web*.br, e como se dá seu crescimento? Que tipo de tecnologias são utilizadas? Onde os sítios *Web* estão hospedados? No Brasil ou no exterior? Os sítios são aderentes aos padrões *Web*, como HTML e CSS? Os sítios são acessíveis? Há suporte a IPv6? Quais tecnologias são usadas para os servidores, páginas, imagens, documentos, vídeos etc? Os servidores mantém seus relógios sincronizados com a Hora Legal Brasileira?.

A proposta deste projeto é que ele seja realizado anualmente e esperamos que os seus resultados contínuos possam servir para que instituições públicas, privadas e acadêmicas possam medir e acompanhar a evolução da *Web* brasileira e das políticas públicas governamentais na área de governo eletrônico.

Vagner Diniz
Gerente - W3C Escritório Brasil

CAPÍTULO 1

Os desafios técnicos para o estudo da *Web* brasileira

Os desafios técnicos para o estudo da *Web* brasileira

No CEPTRO.br, nossa curiosidade pela *Web* foi aguçada ao realizarmos alguns estudos simples sobre a geolocalização dos servidores que hospedavam os domínios ".br", apresentados nas duas últimas reuniões do PTT Fórum[1], evento destinado aos Sistemas Autônomos brasileiros — redes que compõem a Internet. Esses dados nos informavam que mais de um terço dos servidores *Web* estavam hospedados fora do Brasil, cenário muito aquém do ideal, já que implica em latências mais altas e custos mais elevados para os provedores de acesso nacionais, embora o valor de hospedagem para o sítio possa ser menor. Precisávamos saber mais. Que tipos de sítios eram esses? Eram os mais ou menos importantes? Grandes ou pequenos? Voltados ao mercado nacional ou ao exterior?

Conduzimos também um projeto para a disseminação do IPv6 no país, e acompanhar a sua adoção na *Web* brasileira seria um ótimo indicador da efetividade de nossas ações. De forma semelhante, gostaríamos de saber se os servidores *Web* estavam sincronizados com a hora correta, o que é recomendado pelo CGI.br e possibilitado através do serviço NTP.br oferecido em conjunto com o Observatório Nacional. Quando o escritório do W3C apresentou-nos o desejo e a necessidade do governo de conhecer melhor a aderência aos padrões de acessibilidade dos sítios, percebemos que realmente era uma necessidade conhecer melhor a *Web* brasileira e decidimos

[1] PTT - Ponto de Troca de Tráfego

nos dedicar ao projeto. Não tínhamos ideia, contudo, do tamanho do desafio ao qual nos proporíamos, principalmente em seus aspectos técnicos.

A forma como é constituída a *Web*, por si só, traz diversas dificuldades. Por exemplo, não há realmente uma "*Web* brasileira"; a *World Wide Web*, como o próprio nome diz, é uma rede de alcance mundial. Limitar o escopo do estudo foi o primeiro dos desafios. O que faríamos? Consideraríamos a linguagem das páginas? Se o fizéssemos, haveria a dificuldade em si, de identificar o idioma, e o risco de considerarmos sítios dos demais países lusófonos. Consideraríamos, então, a geolocalização dos servidores utilizados para hospedar a *Web*? Sabíamos de antemão que uma grande percentagem dos domínios ".br" estavam hospedados fora do país. Decidimos considerar apenas a *Web* formada pelos domínios ".br", conscientes de que há sítios nacionais hospedados em outros domínios que ficariam fora do estudo. Para minimizar o problema, consideramos no estudo os sítios sob outros domínios encontrados por meio de um redirecionamento a partir de uma URL que apontasse para um ".br".

A *Web* é uma rede cujos conteúdos estão interligados através de documentos de hipertexto. Seu estudo é possível por um processo de análise e coleta sucessiva das páginas, a partir de um conjunto de sítios previamente conhecidos. Essa busca é feita de forma automática por um programa de computador normalmente chamado de *crawler*, coletor, ou batedor. Nem toda a *Web* está interligada, contudo, embora a maior parte dela esteja: há "ilhas" de tamanhos variados sem ligação com o restante da rede. Isso significa que o conjunto inicial de sítios a partir dos quais a pesquisa é feita influencia o resultado, e encontrar o conjunto adequado, geralmente o mais completo possível, é um passo importante. Na coleta do ".gov.br", por exemplo, a situação ideal seria conhecermos os domínios registrados diretamente sob o ".gov.br", mais os domínios registrados sob as siglas das unidades federativas, como ".sp.gov.br". Os primeiros estão sob responsabilidade do Governo Federal, e obtivemos a base; os demais são responsabilidade dos Governos Estaduais e contamos com o apoio da ABEP (Associação Brasileira de Entidades Estaduais de TICs) em sua obtenção. Ainda assim, apenas 8 unidades federativas haviam nos enviado os dados na época da coleta, obrigando-nos a, paliativamente, complementar os dados usando sítios encontrados em buscadores *Web*.

Há também armadilhas para o processo de coleta: sítios com um número infinito de páginas, geradas dinamicamente. Elementos simples, como um calendário gerado automaticamente no sítio, podem criar situações desse gênero. Limites de tamanho e profundidade têm de ser estabelecidos, com o risco de impedirem a coleta de partes de sítios maiores que estes.

Outro ponto a ser considerado é o que apelidamos de "*Web* profunda": a parte da rede em que é requerida a autenticação do usuário para a navegação, por exemplo a maior parte dos sítios de relacionamento ou comunidades. Essa parte da *Web* é inacessível através do método utilizado, tendo ficado fora do estudo. Existe ainda a possibilidade de serem consultados servidores temporariamente indisponíveis, ou de serem encontrados sítios sem o arquivo *robots.txt*, que especifica se eles podem ou não ser visitados por batedores automatizados, ou sítios em que esse arquivo negue a possibilidade da coleta.

Consideramos, ainda, os recursos de tempo, processamento, conectividade e disco, para coletar, armazenar e processar os dados: mesmo agora, com a primeira parte do estudo concluída, temos ainda dificuldade em estimar o que seria necessário para um estudo no formato censitário de toda a *Web* ".br". As estimativas de quantidade de dados, por exemplo, variam entre 30 e 300Tbytes, considerando-se apenas as páginas em formato HTML.

Ao aventarmos a possibilidade de fazer o estudo, um dos primeiros passos foi procurar por pesquisas similares realizadas anteriormente, e por ferramentas. Encontramos algumas pesquisas de cunho acadêmico, inclusive realizadas sobre a *Web* brasileira, que nos auxiliaram no processo. Encontramos também algumas ferramentas que poderiam ser aproveitadas para a coleta dos dados. Em particular, estudamos três programas de computador para essa finalidade: o Nutch, um coletor utilizado para a criação de buscadores; o Heritrix, usado no *Web Archive*, um projeto que mantém um arquivo histórico de parte relevante da *Web*; e o WIRE, utilizado em um dos estudos acadêmicos que encontramos, escrito justamente com a finalidade de realizar estudos sobre a *Web*, tendo já embutidas algumas ferramentas de análise que consideramos de interesse: análise do tamanho das páginas, tipos de documentos, idiomas, cálculo de *rankings*, etc. A conclusão foi: começar o estudo utilizando qualquer uma delas traria vantagens em relação ao desenvolvimento de uma ferramenta inteiramente nova. Optamos pelo WIRE, principalmente pela existência das funcionalidades de análise, já incorporadas ao programa.

Sabíamos que algumas modificações teriam de ser feitas no WIRE original. Por exemplo, seria necessário que armazenássemos as páginas *Web* integralmente, para possibilitar a aderência aos padrões, então os arquivos HTML coletados, que antes passavam por um filtro para eliminar algumas *tags* HTML, e eram armazenados em um grande arquivo de dados de formato proprietário, passaram a ser armazenados integralmente no sistema de arquivos, em pastas e subpastas, num formato similar ao original dos próprios

sítios. Essa modificação ajudou também a tornar o WIRE mais escalável. Outra modificação foi feita para acertar o comportamento do *software* em relação aos *redirects*, de forma que se adequasse à definição de *Web* brasileira explicada anteriormente.

Embora o WIRE tivesse sido usado em vários estudos acadêmicos, foram necessárias ainda diversas novas implementações e correções de comportamento para que o considerássemos pronto para ser usado no estudo. Fizemos uma melhora significativa na ferramenta de identificação de idiomas, com objetivo de melhorar seu desempenho. Pode-se citar ainda, nesse contexto: a normalização das páginas segundo a RFC3986, o tratamento do HTTP 1.1, com a transferência progressiva dos dados, a melhora do tratamento da codificação das páginas, a aleatorização da ordem em que os documentos são baixados e mudanças no tratamento das listas de sítios a serem percorridos, além de diversas correções de *bugs*. O WIRE é uma ferramenta difícil de ser testada. Para alcançar esse resultado, foram necessários meses de desenvolvimento, e muitas coletas de partes significativas da *Web* brasileira.

Gostaríamos, com o estudo da *Web*, de responder a várias questões que não estavam contempladas nos resultados das análises feitas pelo WIRE. Por exemplo: a geolocalização dos servidores, a aderência ao IPv6 e ao NTP, e a aderência aos padrões HTML e de acessibilidade (eMAG / WCAG). Essas análises poderiam ser incorporadas ao WIRE ou implementadas numa ferramenta separada. Optamos pela segunda alternativa, de forma a evitar a inserção acidental de novos *bugs* no código do WIRE, com o qual ainda não estávamos completamente familiarizados. Foi criada a ferramenta cujo nome provisório é AnáliseInternet, que realiza os testes citados, e tem a função adicional de armazenar tanto os dados do WIRE, quanto os de suas próprias análises, num banco de dados único. Reutilizamos, quando possível, ferramentas já prontas. Por exemplo, para verificar a aderência ao padrão HTML usamos o validador criado pelo W3C, rodando localmente, o qual é consultado pelo AnáliseInternet. Para os testes de acessibilidade, incorporamos ao programa rotinas do ASES, programa desenvolvido pelo Governo Brasileiro.

Ao terminar a análise dos dados desse primeiro estudo parcial, da *Web* ".gov.br", concluímos que temos um conjunto de ferramentas confiáveis que nos servirão bem nos estudos adicionais que faremos. Sabemos, no entanto, de limitações que ainda precisam ser vencidas, por isso modificações continuam a ser feitas, seguidas de testes extensivos. Dentre as modificações em curso, podemos destacar: a análise do tempo correto através do protocolo NTP, no lugar de usar apenas a hora fornecida pelo próprio HTTP, quando possível; a contagem do tamanho dos objetos não HTML presentes nas pá-

ginas, como imagens e vídeos, sem baixá-los, através de consultas HTTP HEAD; a melhora no tratamento às "armadilhas" citadas anteriormente e a melhora na velocidade das coletas e análises. Além disso, há a necessidade de automatizarmos parte das análises estatísticas necessárias para a geração deste relatório, com a possibilidade de utilização de ferramentas do tipo *Data Warehouse* e *Data Mining*.

Estamos, ainda, nos preparando para em breve tornar públicos os códigos utilizados, com licenças livres, de forma a garantir a transparência total sobre a metodologia e, quiçá, conseguir a colaboração de outros desenvolvedores e utilizadores dos programas para vencer os muito desafios que ainda nos restam.

Antonio M. Moreiras
Gerente - CEPTRO.br

CAPÍTULO 2

Metodologia

Metodologia

Conceitos e orientações para o Censo *Web* .br

A palavra censo origina-se no latim *census* e significa hoje a "contagem ou enumeração completa" de uma população de indivíduos ou objetos determinados. Portanto, censo é o resultado final de uma contagem específica que define o conjunto de dados estatísticos sobre as diversas variáveis de uma população investigada.

Para a realização de um censo, é fundamental definir rigorosamente o conceito das unidades empíricas que serão objetos de análise, além dos procedimentos técnico-metodológicos para elaboração do quadro populacional (definição dos perfis e dos limites da população objeto de investigação), coleta dos dados (características individuais a serem identificadas) e tabulação dos resultados (definida segundo os requisitos de um plano tabular).

Neste sentido, a possibilidade de realização de um censo está diretamente condicionada ao conhecimento e à definição prévia dos "limites populacionais" aos quais devem-se ater os objetos individuais a serem recenseados. Em outras palavras, para o estudo do tamanho e composição da *Web* brasileira, é necessário a definição de seus domínios e consequentes limites.

Então, para uma primeira consolidação de um Censo da *Web* Brasileira, definiram-se conceitualmente as unidades a serem pesquisadas como aqueles sítios da *Web* referenciados por um nome sob o domínio .BR. Assim sendo, assume-se que um conteúdo pertence à *Web* brasileira se o domínio de topo do nome do seu sítio *Web* respeita uma das seguintes condições:

1. Está sob a hierarquia .BR;
2. Não está sob a hierarquia .BR, mas existe um redirecionamento a partir de um domínio sob o .BR. para ele. Por exemplo, uma empresa multinacional que registra o domínio .BR com a sua marca, porém o aponta (redireciona) para o sítio *Web* principal da empresa que está sob a hierarquia .com.

Consideraram-se, ainda, em algumas das análises, os *links* para documentos presentes nas páginas de sítios .BR, mesmo que estejam hospedados fora desta hierarquia de domínios.

Contudo, um dos maiores problemas encontrados até agora para a consecução deste censo esbarra exatamente na topologia do universo virtual da *Web*, que limita a capacidade técnica de mensuração do tamanho e composição real do que seria uma "população de domínios e objetos virtuais". Para além das questões que cercam a complexidade de identificação dos limites da "*Web* profunda", o próprio espaço conhecido da *Web* .br, por exemplo, devido à sua dinâmica inerente, parece intratável quanto as técnicas de rastreamento e coleta de informações, dificultando em muito o trabalho de contagem e de identificação dos perfis de domínios e de objetos e, principalmente, sobre o conhecimento da "popularidade" desses objetos na população.

Diante do quadro de incertezas sobre a dinâmica, tamanho e composição da *Web*, em princípio pareceria extrema ousadia a proposição de uma metodologia rigorosa de ampla aplicação para mensuração objetiva da *Web* brasileira. Portanto, deixa-se claro que o avanço e consolidação dessa metodologia refere-se a um processo maior e integrado de planejamento sistemático sobre diferentes etapas que devem definir um modelo para "identificação", "coleta", "validação" e "análise" de todas as informações disponíveis para determinação de uma população de domínios .br.

Em outras palavras, há a consciência de que a aplicação do conceito de "censo" e a determinação de uma "população de domínios" deve ocorrer em perspectiva e consolidar uma metodologia apropriada para a realização rigorosa de uma contagem definitiva em um futuro próximo, que se realizará a partir do aperfeiçoamento dessa metodologia e das contagens sucessivas que se pretende conduzir desde agora. Nesse momento, desenvolvem-se essa metodologia e sua padronização para realizações futuras.

Para a defesa da ideia de um "Censo da *Web* .br", poder-se-ia se traçar um paralelo com a metodologia consolidada nos estudos de população em geral, nomeadamente a área da Demografia. Assim, um ponto fundamental a ser definido no Censo da *Web* .br é a realização eventual de uma contagem/enumeração completa da população de domínios .br. Partindo da experiên-

cia desenvolvida na Demografia, para proceder à enumeração propriamente dita, é necessário definir conceitualmente o que é "população"; faz parte desse entendimento definir também o conjunto de técnicas necessárias para a identificação e registro dessa população [1].

Por exemplo, para a contagem da população humana, definem-se os domicílios de referência onde reside inequivocamente cada indivíduo membro da população-alvo. Assim, a contagem pode ser feita por meio do registro fiduciário de imóveis em prefeituras municipais. Nesse caso, o censo poderia se resumir simplesmente à coleta de informações em cada prefeitura do país sobre o registro fiduciário de cada domicílio e soma efetiva de todos os membros associados a cada domicílio enumerado. No caso dessa contagem populacional (de indivíduos), parte-se do pressuposto (forte em demografia) de que cada pessoa faz parte de um domicílio, ou seja, reside em um e apenas um domicílio (existem exceções e também técnicas para ajustar tais exceções).

Assim, quando se enumera a população brasileira, aponta-se um quadro populacional definido, baseado nos domicílios e nos indivíduos referidos à unidade de residência, e as técnicas de contagem da população resumem-se à qualificação do desenho de pesquisa e organização não trivial do trabalho de campo, ou efetivamente à qualidade do trabalho dos recenseadores em cada domicilio existente (e devidamente registrado) para catalogar o número de residentes em cada habitação.

A partir desse pequeno exemplo, imagina-se a aplicação de uma lógica semelhante de pesquisa para enumeração da *Web* brasileira. O ponto principal seria definir um limite referencial para o universo da população alvo, mesmo que este seja apenas estimado e nunca verificado empiricamente, pois, nesse caso, o que importa é estabelecer uma "métrica" como referência para análise dos objetos coletados em diferentes momentos no tempo. Assim, parte-se das informações coletadas sobre os registros oficiais dos domínios .br como uma referência sobre a população alvo; os limites referenciais para a população são dados pela definição do domínio de primeiro nível .br . Seguindo a lógica demográfica indicada, a partir da definição de uma "malha digital" dos domínios registrados ".br", estabelecem-se os vínculos de cada objeto individual observável do universo virtual com seu domínio de referência. Consequentemente, obtém-se um quadro populacional definido basicamente pelo tamanho do conjunto de domínios de primeiro nível e sua composição por objetos atribuídos.

Contudo, esse procedimento em si mesmo não resolve todo o problema da contagem, porque não indica uma ideia real do tamanho da *Web*; além dis-

so, sabe-se que o rastreamento efetivo de toda a população (tanto a da população humana quanto a de objetos na *Web*), ou seja, a chamada "cobertura censitária", perfeita em qualquer contagem, depende de uma série de fatores muitas vezes não controlados, que inviabilizam um fechamento completo da enumeração. Por exemplo, a contagem de indivíduos em um domicílio pode ser prejudicada pela recusa do residente em receber um recenseador. Assim, até mesmo em Demografia, existem limitações para a realização de "censos perfeitos" e, recorrentemente, os melhores censos demográficos assumem um erro de cobertura aceitável entre 2 a 8% dos indivíduos/domicílios em relação à população total.

Em que pesem as limitações impostas pelo próprio processo de coleta (qualidade dos batedores/recenseadores), distribuição populacional (objetos isolados ou inatingíveis) e natureza dinâmica da *Web*, existem também métodos demográficos específicos para correção dos erros de cobertura censitária, que poderiam ser estendidos e aplicados no caso do Censo da *Web* .br. Nesse caso, a questão seria definir o "grau de cobertura" em relação à provável população real e, a partir desse parâmetro, promover a correção do tamanho efetivo da população alvo.

Esse relatório enseja um primeiro esforço a fim de estabelecer a metodologia capaz de estimar o chamado "grau de cobertura" para a consequente correção das estimativas do tamanho da *Web* .br.

Chega-se, assim, ao desafio seguinte, um segundo ponto: a definição de um procedimento metodológico rigoroso para estimar o grau de cobertura e o tamanho mais provável da população-alvo.

Aqui surgem alguns desafios que têm sido estudados para se adequarem à aplicação metodológica no âmbito da computação e da estimativa do tamanho da *Web* .br. Em princípio, existem duas formas básicas de cálculo da cobertura e estimativa do tamanho real de uma população: 1) estimar a cobertura em um censo, a partir da comparação demográfica com um censo anterior; 2) utilizar técnicas estatísticas específicas para se definirem populações difíceis de serem contadas (raras);

1. No caso da estimativa de cobertura a partir de dois censos, existiriam duas limitações imediatas para aplicação no Censo da *Web* .br. Primeiro, seria necessário haver um censo (ou pelo menos um esforço idêntico de contagem de todos os domínios .br) num tempo T1, e outro num tempo T2. Na análise demográfica tradicional de populações humanas, utilizam-se dois censos como parâmetro para se balizar todo o período de exposição da população-alvo que, mediante análises demográficas

diretas e indiretas sobre as "entradas" e "saídas" de indivíduos da população geral, possibilitarão a estimativa segura de um tamanho e de uma composição populacional. Assim, na realidade, essa metodologia propõe tomar uma população exposta num período qualquer e, a partir do seu registro direto (isto é, da contagem em dois momentos distintos), utiliza variáveis estruturais específicas para estimar efeitos diretos e indiretos de transformação da população do tempo original T1 para T2. Ao comparar as resultantes entre o modelo do tamanho e da composição da população esperada com a população observada no segundo momento, obtém-se uma definição aproximada do tamanho populacional no tempo T2. Isso Demanda um conhecimento específico sobre a estrutura populacional, ou seja, que se conheçam as variáveis populacionais principais (no caso da demografia humana, são as variáveis de idade e sexo, pois expressam diretamente o efeito de entrada e saída – nascimento - óbitos na população geral) que definem a estrutura da população e sua dinâmica. A replicação dessa metodologia estrita, no caso do Censo da *Web* .br, não se mostra factível, dada a inexistência de variáveis estruturais da população de domínios e objetos.

2. Há uma segunda metodologia que se apresenta mais adequada e plenamente realizável para a consecução do Censo *Web* .br., e diz respeito às técnicas estatísticas desenvolvidas para estimativas de tamanhos de populações raras ou difíceis de contar. Uma das técnicas de estimativas de tamanho populacional mais utilizadas nas ciências biológicas (e também na demografia para controle do grau de cobertura censitária) para contar populações ecológicas é a chamada "captura-recaptura" [7, 1]. A replicação dessa técnica consiste basicamente em enumerar o universo dos domínios .br e identificá-los (marcá-los) um a um. Na realidade, basta um identificador exclusivo para cada domínio que surgiu na amostra dessa primeira enumeração. Depois de um intervalo de tempo suficiente para haver transformações nessa população (por exemplo, surgimento de novos domínios), proceder-se-ia a uma segunda enumeração, seguindo os mesmos parâmetros executados na coleta anterior. Tem-se assim duas amostras da população de domínios, em que os indivíduos expostos (domínios .br e seus objetos vinculados) em uma amostra não necessariamente aparecerão na amostra seguinte, e vice-versa. Utiliza-se, então, um modelo matemático simples para estimar o tamanho provável da população total a partir da probabilidade de haver defasagens e repetições da presença dos domínios em diferentes amostras da mesma população (domínios .br e seus objetos vincula-

dos). Consequentemente, estabelecer-se-ia, a partir dos procedimentos de "captura-recaptura", uma metodologia rigorosa e estatisticamente segura para estimativa do tamanho real de uma população com estrutura desconhecida.

Como se afirmou, talvez o maior problema para a consecução de um censo seja o estabelecimento dos parâmetros de cobertura censitária, visto que a "cobertura" reflete o grau de acuidade da contagem frente à população inicial, cuja contagem pressupõe que seus limites (espaciais e temporais) sejam definíveis a *priori*, de maneira que o resultado final da contagem e listagem reflita realisticamente o total de "objetos" que devem fazer parte da população inicial.

Em geral, quando se conhece de antemão a população a ser investigada (especialmente quando se conhece seu tamanho no tempo inicial T1), pode-se definir a estimativa do grau de cobertura (por exemplo, o grau de acuidade do censo) a partir de técnicas demográficas diretas e indiretas, comparando-se a composição populacional nos tempos T1 e T2; entretanto, esse é o caso específico de populações humanas, como ficou claro no item 1.

A defasagem na composição populacional de T1 e T2 deve-se a dois fatores: mudanças efetivas nas características populacionais ao longo do tempo, e erro de cobertura da contagem/listagem de objetos e características nos censos em T1 e T2.

No caso dos censos demográficos tradicionais, o erro de cobertura é uma consequência direta da omissão ou inclusão indevida de domicílios particulares e das pessoas neles residentes, assim como das pessoas residentes em domicílios particulares ocupados e considerados os mesmos nos dois censos comparados (T1 e T2). No caso do censo da *Web* .br, os erros de cobertura serão consequencia direta da omissão indevida de sítios .br (e de seus objetos vinculados) numa contagem em T1 e outra, em T2.

A medição do erro de cobertura é essencial, pois pode informar o grau de precisão (acuidade) das medições do tamanho da *Web* brasileira e, caso necessário, orientar os parâmetros para correção das estimativas. Então, a medição do erro de cobertura é feita a partir da construção de indicadores de omissão de sítios (equivalentes aos domicílios) e objetos (equivalentes às pessoas).

Não por acaso o método escolhido para estimação desses indicadores é o chamado *Dual System Estimation* [4, 3, 1], uma metodologia baseada na técnica de "captura-recaptura", referida no item 2. cujo pressuposto é a amostragem e as coletas semelhantes em dois (ou mais) momentos no tempo, tendo a independência estatística entre as amostras/coletas como requisito. No caso do desenvolvimento dessa metodologia para a medição da *Web* .br,

deve-se garantir a independência em relação ao lançamento das sementes e do batedor (ferramenta de contagem), a partir de uma mesma lista de domínios. Nesse momento, desenvolve-se uma nova metodologia para validar os dados da coleta, visando uma estimativa da cobertura censitária, considerando informações relacionadas às quantidades de domínios registrados (nesse caso, especificamente, aqueles registrados com domínios do "gov.br"), erros retornados no procedimento de coleta de dados e indicadores relacionados ao contexto da *Web*, como crescimento do volume de domínios registrados, modificação do tamanho de objetos informacionais, dentre outros. Esses resultados poderão ser acompanhados mais adiante, na seção de apresentação de resultados e desdobramentos.

Como se trata de algo novo, a proposta é a evolução da metodologia a ser adotada com o tempo, a partir de novas coletas realizadas e novas técnicas propostas para tratar um censo de objetos da *Web*.

A seguir, descrever-se-á brevemente o método de estimação da cobertura censitária e do seu grau de acuidade. O método utilizado para cálculo dos indicadores de "omissão" (erro de cobertura) será o *Dual System Estimation*, que se baseia nas técnicas de "captura-recaptura". Sua utilização requer independência na coleta das duas pesquisas (coletas em T1 e T2) e pressupõe o confronto das informações da seguinte maneira (ilustrada na Tabela 2.1), onde:

- **a** é o número de unidades incluídas em T1 e T2;
- **b** é o número de unidades incluídas apenas em T1;
- **c** é o número de unidades incluídas apenas em T2;
- **d** é o número de unidades desconhecidas que não foram incluídas nem em T1 nem em T2 (desconhecido) e;
- **t** é o total de unidades da população.

COLETA T1	COLETA T2		
	TOTAL	UNIDADES INCLUÍDAS	UNIDADES NÃO INCLUÍDAS
TOTAL	t	a + c	b + d
UNIDADES INCLUÍDAS	a + b	a	b
UNIDADES NÃO INCLUÍDAS	c + d	c	d

Tabela 2.1: Tabela de Informações "Captura-Recaptura"

Apenas **d** é, de fato, um dado desconhecido, pois é o provável número de sítios não coletado nas amostras em T1 e T2 [4]. Quando se compara a primeira "captura" em T1 com o resultado da "recaptura" em T2, observa-se que o erro de cobertura (**d**) pode ser corrigido a partir das probabilidades conhecidas para **a**, **b** e **c**, ou seja, P(T1)=a+b e P(T2)=a+c, visto que **a** são os sítios incluídos em ambas coletas; **b** é composto pelos sítios coletados em T1, mas que não foram recapturados; e **c** são os sítios não capturados em T1, mas capturados em T2.

Finalmente, a partir das coletas sucessivas (que podem ser ampliadas para uma série temporal maior), poderemos:

- Calcular o tamanho do erro de cobertura (**d**); como *output* serão definidas "taxas de omissão";
- Estimar o tamanho da *Web* .br e do número de páginas vinculadas, em uma data específica;
- Estimar tamanhos em diferentes pontos no tempo para avaliação da evolução da *Web* brasileira (por exemplo, crescimento e dinâmica da estrutura e composição da *Web*);
- Estabelecer indicadores variados, segundo as diversas características de composição dos sítios e páginas da *Web* .br.

Aplicação

Em resumo, até esse momento discutiram-se as possibilidades reais para replicação de um censo demográfico sobre a população de domínios .br. Como já se ressaltou, a realização efetiva de uma contagem/enumeração populacional que permita estimar o tamanho e composição da *Web* brasileira implica uma metodologia não trivial, e que ainda está em desenvolvimento para consolidação.

Nesse primeiro esforço, desenvolvem-se as aplicações necessárias para determinação do quadro populacional a ser trabalhado (domínios .br e seus objetos vinculados), as técnicas apropriadas de coleta e validação dos procedimentos e informações coletadas, bem como a metodologia adequada para análise e aferição do tamanho da *Web* .br.

Para se atingirem os objetivos traçados inicialmente, foi preciso redefinir os procedimentos e orientações do estudo, experimentalmente aplicados

à coleta restrita dos domínios .gov.br. A partir dessa primeira experiência, testaram-se alguns procedimentos para estimativa do tamanho populacional da *Web* brasileira sob os domínios .gov.br.

Nesse primeiro momento, o objetivo restringe-se à tentativa de aplicação, avaliação e validação dos procedimentos metodológicos pré-definidos. A seguir, descreve-se seu "passo-a-passo":

Definição da pergunta e os dados para respondê-la

Primeiro, o objetivo é definir uma estimativa para o tamanho da parte da *Web* .br sob o domínio .gov.br. Para tal, utilizaram-se as informações sobre o número de sítios (.gov.br) coletados em dois momentos distintos, bem como o número de páginas referidas ao conjunto de sítios coletados.

Portanto, há duas variáveis básicas: 1) número de sítios .gov.br, e 2) número de páginas vinculadas aos sítios coletados.

Em segundo lugar, como a coleta do .gov.br foi feita em dois momentos distintos (T1 e T2), a que o número de sítios e páginas diz respeito. Como forma de se garantir a aplicação do método de "captura-recaptura" para estimar o tamanho da *Web* .gov.br, as duas coletas feitas em T1 e T2 satisfazem os requisitos necessários (independência das coletas, e garantia do lançamento aleatório das sementes).

Breve descrição da metodologia

O método conhecido como *Dual System Estimation* (DSE) – aqui tratado como método de "captura-recaptura" – é comumente utilizado pelos institutos nacionais de estatísticas de população, especialmente para conferência (checagem) da qualidade censitária [7, 1].

As estimativas sobre o tamanho da população derivam de relações matemáticas e de estatísticas elementares, desde que alguns pressupostos fundamentais sejam observados: independência das coletas, distribuição aleatória dos objetos na população e a mesma chance aleatória de o objeto ser coletado em todas as coletas. Claramente, alguns desses pressupostos não são ob-

servados empiricamente no universo da *Web*. Em especial, a distribuição aleatória de objetos e de suas conexões no universo *online* (sabe-se que a topologia da rede *online* possui uma distribuição em escala-livre, observando os requisitos de uma *power law* e, consequentemente, a distribuição de vértices e arcos não segue um padrão) [2, 5, 6].

De qualquer forma, inicia-se a aplicação de uma metodologia que deve ser ajustada ao universo da *Web*, como fizeram Jianguo Lu e Dingding Li para estimar o tamanho da *Web* profunda [6]. Observa-se, portanto, que existe uma correspondência plausível entre as estimativas e as coletas feitas.

Para a compreensão do método "captura-recaptura", considerou-se uma população desconhecida (o tamanho da *Web* .gov.br), cujos objetos (indivíduos) foram listados em um primeiro momento, gerando um conjunto de objetos **n1**, e posteriormente, em um segundo momento, um conjunto de objetos **n2**. É importante frisar que a listagem representou a coleta exaustiva de todos os objetos da população-alvo. Ao se comparar os dois conjuntos coletados (**n1** e **n2**), notou-se que existe um conjunto **m** de objetos duplicados, isto é, objetos presentes nas duas coletas.

Assume-se que as duas coletas são independentes e que os objetos coletados têm a mesma probabilidade de serem coletados em ambas as coletas. Como mostram Alho e Spencer [1], o conjunto de objetos duplicados **m** segue uma distribuição de probabilidade hipergeométrica quando conhecemos o tamanho da população total de objetos **N** (observados e não observados). Pode-se, indiretamente, a partir da equação da distribuição de probabilidade hipergeométrica, estimar o tamanho total da população **N** a partir de um estimador de máxima verossimilhança que torne o conjunto de objetos **m** observados o mais provável possível.

Portanto, o estimador EN será o valor de **N** que maximiza a probabilidade de o conjunto observado de objetos duplicados m ser verdadeiro para toda a população. Aqui o estimador de máxima verossimilhança é:

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

em que **n1** e **n2** representam o conjunto de objetos coletados em cada momento T1 e T2, e **m** representa o conjunto de objetos coletados em ambos momentos.

A equação 3.1 mostra o estimador clássico do método de "captura-recaptura", definido desde Francis Bacon (1560) e reinventado diversas vezes, até a consolidação com Laplace (1802) e a sua especificação no campo da

biologia com Petersen (1896), conhecido como estimador de Petersen [7, 1]. Além disso, outros estimadores foram desenvolvidos para se adequarem à realidade empírica dos dados. Apenas para efeitos comparativos, indica-se aqui um estimador derivado de Petersen, utilizado por Lu e Li [6], o conhecido estimador de Shumacher, indicado para populações com distribuição uniforme, visto ser objetivo do grupo de trabalho aprofundar o conhecimento sobre a metodologia e desenvolver estimadores adequados à realidade empírica da *Web* .br.

Resultados

VALORES	SÍTIOS	SÍTIOS OK	PÁGINAS HTML OK
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: Tabela de Resultados

Usando as técnicas apresentadas na metodologia para avaliar a estimativa para sítios (*Hosts*), os valores obtidos foram:

- Razão de Consistência (fator de correção) $R = \frac{EN}{t} = 1{,}0029$
- Estimador Clássico de Shumacher $EN = \frac{n_1 * n_2}{\mathrm{m}} = 20.217$

Usando as técnicas apresentadas na metodologia para avaliar a estimativa para sítios com páginas válidas, Sítios OK, os valores obtidos foram:

- Razão de Consistência (fator de correção) $R = \frac{EN}{t} = 1{,}0097$
- Estimador Clássico de Shumacher $EN = \frac{n_1 * n_2}{\mathrm{m}} = 13.996$

Aplicando essas mesmas técnicas para avaliar a estimativa para Páginas da *Web* (Páginas HTML válidas), os valores obtidos foram:

- Razão de Consistência (fator de correção) $R = \frac{EN}{t} = 1{,}2740$
- Estimador Clássico de Shumacher $EN = \frac{n_1 * n_2}{\mathrm{m}} = 12.039.334$

Análises

Brevemente, aponta-se que os dois estimadores utilizados (Petersen e Shumacher) apresentam o mesmo valor para o tamanho da provável população de sítios e páginas da *Web* .gov.br. Por meio da razão de consistência (isto é, do estimador de cobertura das coletas feitas), percebe-se claramente os limites de coletas isoladas. Em outras palavras, quando se obtém o somatório de todos os objetos coletados em dois momentos distintos no tempo, têm-se um total de 20.158 sítios, 13.861 sítios OK e 12.039.334 páginas válidas (OK), sob o domínio .gov.br. Contudo, a razão de consistência, fator de correção para a cobertura das coletas, indica que houve uma subestimativa na ordem de 0,3% para o tamanho da população de sítios .gov.br e 0,97% para sítios OK. No caso de páginas HTML válidas, a subestimativa foi bem mais acentuada, de aproximadamente 27%, devido à grande variabilidade de páginas entre as 2 coletas, justificada pela característica dinâmica da *Web* e também pela natureza de suas páginas, que muitas vezes variam tecnologicamente a cada execução, o que diz respeito ao conceito de páginas dinâmicas.

Conclusão

Os conceitos adotados como parte metodológica estão adequados aos objetivos do projeto e seus desdobramentos até o presente. No que diz respeito aos indicadores gerados e suas análises, cabe ressaltar que estas são válidas e pertinentes às questões que se buscavam responder, respeitadas às limitações existentes em termos de coleta de dados realizada para a análise.

Em termos de estimativas futuras e previsões acerca do universo de domínios da *Web*, as técnicas aplicadas até aqui ainda não se mostraram eficientes, dado o cenário deste projeto ser muito dinâmico e desafiador, o que demanda novos estudos científicos, que poderão gerar novos métodos que permitam extrapolar os resultados apresentados e fazer previsões futuras de mudanças da *Web* brasileira. Isso reforça a boa escolha da estratégia de contagem adotada até aqui para análise do universo da *Web* .gov.br, que deverá ser ampliado para outros domínios nas etapas seguintes do trabalho.

Mesmo assim, existe interesse em pesquisa e desenvolvimento de novas técnicas que permitam, de forma complementar ao método de contagem (Censo), fazer estimativas e avaliar tendências futuras para a *Web* brasileira, a fim

de se possilibitar o confronto de análises e a garantia de melhor qualidade acerca do estudo e da avaliação de características quantitativas e qualitativas sobre a *Web*.

Equipe técnica

InWeb – Instituto Nacional de Ciência e Tecnologia para a *Web*

Bibliografia

_____ [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] Instituto Brasileiro de Geografia e Estatística. Metodologia do censo demográfico 2000. Série Relatórios Metodológicos, 25, 2003.

_____ [5] DANIEL GOMES E JOÃO MIRANDA. Arquivo e Medição da *Web* Portuguesa. In Pedro Isaias, editor, Proceedings of Ibero-Americana IADIS 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.

CAPÍTULO 3

Análise dos Resultados

Análise dos resultados

Definindo o conceito de *Web*

A *World Wide Web*, também conhecida como *Web*, ou simplesmente WWW, é um gigantesco acervo universal de páginas, documentos, dados, aplicações e serviços interligados por meio da rede mundial de computadores, disponibilizado às pessoas de qualquer lugar do globo, a qualquer momento e por diversos dispositivos, desde computadores até aparelhos móveis, como telefones celulares. Esse imenso acervo pode reunir diversos tipos de conteúdos digitais, desde páginas de hipertextos, até arquivos no formato de imagens, figuras, som, vídeos, e códigos de programação, dentre outros. Todo arquivo disponível na *Web* é identificado por um endereço único e exclusivo, chamado URL, que significa *Uniform Resource Locator*, em português Localizador Padrão de Recursos. Uma URL indica o local onde se localiza o arquivo digital na *Web*. Essa foi a grande invenção de Tim Berners-Lee, que, ao criar todo um sistema de localização na *Web*, possibilitou que os documentos pudessem ser acessíveis em qualquer lugar do globo.

Cada um destes acervos é identificado por um nome ou domínio, comumente conhecido por *Website*, sítio, ou sítio *Web*. Toda vez que navegamos na *Web*, digitamos esses nomes para acessarmos os sítios que desejamos, por exemplo: http://www.cgi.br, http://www.google.com.br, http://www.receita.fazenda.gov.br. É importante notar que a *Web*, embora seja uma aplicação poderosa e de ampla utilização, é apenas uma parte da rede, uma aplicação.

Os nomes de domínios também têm sua organização própria, não sendo de uso exclusivo da *Web*, e estão estruturados globalmente em níveis hierár-

quicos. Os domínios de primeiro nível são chamados de TLDs, acrônimo de *Top Level Domains*; existem diversos tipos, por exemplo o ".net", o ".com", ".org", entre outros. Os domínios que identificam um determinado país de origem, como o .br, são chamados de código de país ou *Country Code*. Assim, o .br é um *Country Code Top Level Domain* – ccTLD, o domínio de primeiro nível do Brasil. Para o nosso caso brasileiro, abaixo desse domínio de primeiro nível existem outros níveis, como o ".gov.br", o ".com.br", o ".org.br". Somente abaixo desses domínios, e seguindo esse esquema hierárquico, os diversos domínios são registrados e criados, por exemplo o domínio "governoeletronico.gov.br". O correto entendimento dessa estrutura de domínios é importante para também compreender as análises expostas neste documento.

As dimensões e as características da *Web* brasileira

Para fins de determinação do escopo, a *Web* brasileira é definida no contexto do desse projeto como a rede formada pelos sítios de aceso público identificados por um domínios .br, mais os sítios para os quais há redirecionamentos diretos, via servidor, a partir de um sítio.br, dos quais analisa-se apenas a página principal.

De acordo com dados do *Registro.br*, autoridade de registro para nomes de domínio no Brasil, o mês de maio de 2010 foi encerrado com cerca de 2,1 milhões de nomes de domínios registrados sob esse respectivo ccTLD, os quais contêm sítios das mais variadas instituições privadas, governamentais, instituições de ensino, organizações do terceiro setor, profissionais liberais, pessoas físicas, etc. Isso exige a realização de estudos sobre universos específicos de nomes de domínios, por exemplo ".com.br", ".org.br", ".net.br", ".gov.br" e outros grupos menores, com o objetivo de medir suas características na *Web*.

Como ponto de partida para um levantamento mais amplo das dimensões e características do .br, optou-se por uma coleta exaustiva da *Web* governamental brasileira, aquela constante nos sítios e páginas sob o domínio *.gov.br*.

Os principais resultados e algumas conclusões desse levantamento são apresentados neste relatório. Esse primeiro estudo será de grande utilidade e subsidiará o planejamento de uma coleta mais ampla e detalhada das dimensões de toda a *Web* brasileira sob o ccTLD .br.

As dimensões e as características do .gov.br

A coleta de dados sobre os domínios do governo foi realizada em outubro de 2009 e identificou um total de 18.796 sítios sob o *.gov.br*, a partir de URLs percorridas. A identificação do total de sítios partiu de dados fornecidos das seguintes fontes:

a. Domínios identificados como *.gov.br* (domínios reservados ao Governo Federal), cuja lista foi fornecida pela autoridade de registro para nomes de domínio no Brasil, o Registro.br, com autorização do Ministério do Planejamento, responsável pelo uso dos domínios sob o *.gov.br*.
b. Domínios identificados como *sigla-uf..gov.br*, registrados pelas empresas estaduais de processamento de dados, vinculadas aos governos estaduais;
c. Resultados de consultas e buscas de informações, utilizando ferramentas de busca, com o objetivo de complementar as informações anteriores.

Essas diferentes fontes foram unificadas e serviram como semente para um sistema coletor. Objetivou-se com esse levantamento produzir um cadastro que pudesse contemplar o maior número possível de sítios governamentais, de tal modo que fosse o mais próximo de um censo da *Web* governamental brasileira. Porém, nem todas as empresas de processamento de dados das unidades das federações e responsáveis pelos registros dos domínios *sigla-uf.gov.br* puderam responder em tempo, fato que introduziu mais uma dificuldade para a realização de um censo da *Web* governamental, além daquelas inerentes a própria *Web*.

Participação das regiões na composição da *Web* governamental

A partir dos resultados da coleta, investigou-se a participação de cada uma das cinco regiões brasileiras e também a do Governo Federal na composição da *Web* a partir da análise dos subdomínios correspondentes aos estados, por exemplo o subdomínio *.sp.gov.br* foi considerado como que da região sudeste, e o *.gov.br* como do governo federal. Avaliaram-se dois aspectos dessa participação:

a. O número total de sítios correspondentes a cada uma das cinco regiões do país e do governo federal;

b. A quantidade total de *bytes* por região e *Governo Federal*, considerando-se apenas os documentos HTML (isto é, imagens, vídeos e outros tipos de

arquivos não foram contabilizados na tabela abaixo. Vide figura 17 para uma análise de outros tipos de arquivos).

A Tabela 3.1 apresenta um sumário do número de sítios e da quantidade de *bytes* coletados para cada uma das grandes regiões brasileiras. Embora a coleta tenha chegado a um total de 18.796 sítios, foram considerados os que continham pelo menos um documento HTML nessa análise. Os sítios satisfazendo essas condições totalizam 11.856.

REGIÃO	VOLUME EM GIGABYTES	NÚMERO TOTAL DE SÍTIOS	PARTICIPAÇÃO DA REGIÃO NO TAMANHO TOTAL EM BYTES	PARTICIPAÇÃO DA REGIÃO NO NÚMERO TOTAL DE SÍTIOS .GOV.BR
SUL	26	3.416	18%	29%
SUDESTE	32	3.358	22%	28%
NORTE	7	816	5%	7%
NORDESTE	27	1.786	18%	15%
GOV.BR	38	1.668	26%	14%
CENTROOESTE	17	812	11%	7%
TOTAL	**148**	**11.856**	**100%**	**100%**

Tabela 3.1 – Quantidade de sítios e tamanho em Gigabytes por região geográfica

A distribuição percentual dos dados apresentados na Tabela 3.1 pode ser analisada no gráfico apresentado na Figura 3.1.

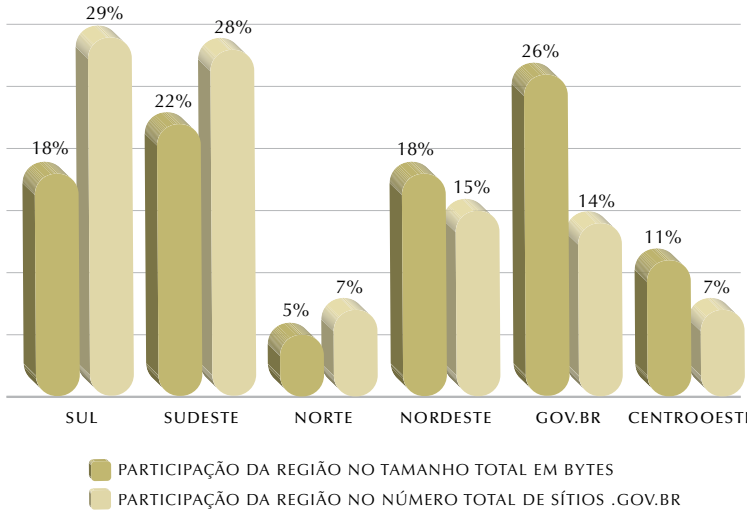

Figura 3.1 – Participação das regiões e do governo federal na composição da *Web* governamental

Observou-se uma maior participação em relação ao tamanho em *bytes* na *Web* governamental no agrupamento composto pelos sítios *Governo Federal*, 26%. Em número absoluto de sítios, a região Sul apresentou 33% dos 18,7 mil sítios coletados.

Em relação à participação das unidades da federação na composição da *Web* governamental, o domínio ***pr.gov.br,*** pertencente ao Estado do Paraná, foi o que apresentou a maior participação em número absoluto de sítios de todos os sítios brasileiros de governo coletados, cerca de 17%, conforme mostrado na Figura 3.2. O *Governo Federal* representado pelos sítios com domínio *.gov.br* vem em segundo lugar, empatado com o Estado de São Paulo (***sp.gov.br)***. Estes dois últimos participam, cada um, com 14% dos sítios sob a *Web* governamental brasileira.

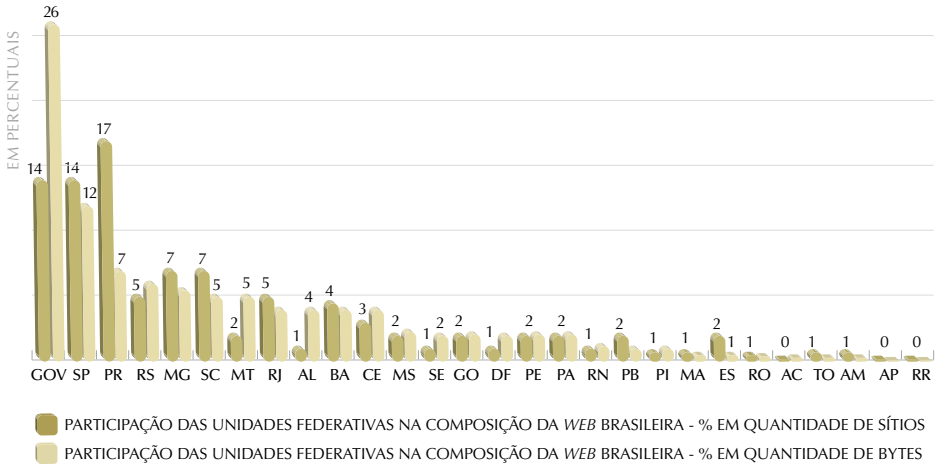

Figura 3.2 - Distribuição percentual do número de sítios por UF, incluindo o Governo Federal (gov)

Em relação ao tamanho do conteúdo em *bytes* dos sítios brasileiros de governo, o *Governo Federal* é o que apresenta o maior índice, com 26% do total verificado, seguido pelo Estado de São Paulo, com 12%. Os demais estados apresentam índice inferior a 10% do total de quantidade em *bytes*.

A relação entre o número de sítios ou eventualmente o número de páginas e a quantidade em *bytes* pode também ser uma abordagem de análise a ser considerada no futuro. Observando-se a Figura 3.2, verifica-se que os estados, em geral, guardam a mesma participação percentual em relação ao total tanto em número de *bytes* de seus sítios como na quantidade de sítios na *Web* governamental brasileira, indicando que possuem um tamanho médio de sítios em *bytes* equivalentes. Poucos estados fogem dessa regra. De um lado, os domínios do *Governo Federal* apresentam maior conteúdo

em *bytes* em relação aos seus sítios. De outro, o Estado de Paraná (PR), que apresenta maior quantidade de sítios em números absolutos e relativos à participação no total da *Web* governamental brasileira. Investigar as razões pode mais do que demonstrar existir quantidade de domínios governamentais na *Web* sem conteúdos significativos, ou ainda domínios governamentais com conteúdo excessivo, revelando uma dificuldade no acesso à informação relevante e pública.

Outros idiomas na *Web* governamental

Das 3.182.202 páginas que puderam ter seu idioma identificado através de análise automatizada, 97% estão em português. O *software* utilizado para identificar a linguagem compara o texto contido nas páginas com dicionários com palavras-chave dos idiomas português, inglês, espanhol e francês, contabilizando as palavras que aparecem num determinado documento. Caso o número de palavras-chave de um dos idiomas testados ultrapasse um determinado limite inferior, e não haja ambiguidade (mais de um idioma com palavras-chave suficientes), ele é considerado identificado. Os resultados para os idiomas estrangeiros estão ilustrados na Figura 3.3.

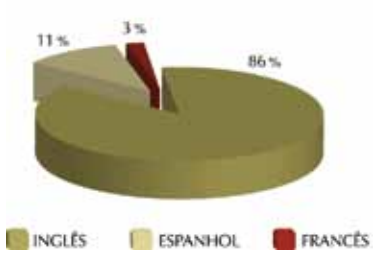

Figura 3.3 – Outros idiomas encontrados na *Web* governamental

É importante lembrar ainda que em 47% das 6,3 milhões de páginas em HTML coletadas o idioma não pode ser identificado por meio do procedimento utilizado. A amostra é significativa e pode revelar que essencialmente os conteúdos da *Web* governamental brasileira visam atender ao público interno, tendo pouca quantidade de informação em outros idiomas. Por outro lado, sem uma abordagem semântica não é possível verificar se informações relevantes em outros idiomas são oportunas, sejam, por exemplo, para conteúdo de relações exteriores ou para os fins de turismo. Identificar qual informação pôde e convém estar em outro idioma não foi ainda objeto de análise.

Aderência aos padrões HTML do W3C

Identificou-se a avaliação da aderência das páginas HTML aos padrões do W3C através da aplicação de um *software* validador projetado pelo próprio consórcio. Tal como propugna o W3C e as boas práticas de desenvolvimento *Web*, a aderência aos padrões *Web* é indicador importante da universalidade de acesso por qualquer dispositivo conectado à *Web*, bem como por qualquer ambiente operacional. Quanto mais aderente aos padrões, melhor a página será acessada por qualquer usuário, independente do dispositivo e de seu ambiente operacional. Por outro lado, páginas não aderentes terão acessos restritos a alguns dispositivos ou sistemas operacionais, donde pressupõe seu caráter de não universalidade. Considera-se que, principalmente para conteúdos da *Web* governamental, a aderência aos padrões e a universalidade do acesso devem ser constantemente consideradas e exigidas.

Para essa análise, verificou-se a contagem do número de incorreções de acordo com o padrão encontrado pelo *software* validador. Dos 6,3 milhões de páginas HTML coletadas, cerca de 91% apresentaram mais de uma incorreção de aderência, apenas 5% estão completamente de acordo com o padrão, e 4% não puderam ser avaliadas, conforme mostra a Figura 3.4.

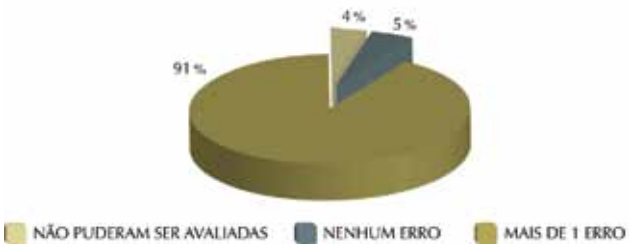

Figura 3.4 – Aderência aos padrões HTML do W3C

Aderência aos padrões de acessibilidade ASES

A aderência a padrões de acessibilidade visa garantir o acesso universal aos sítios *Web*, mesmo para portadores de deficiência. Os critérios de acessibilidade são separados em 3 níveis de acessibilidade ou conformidade, definidos pelo padrão WCAG. O nível de conformidade A é considerado mandatório para que um sítio seja considerado acessível. O nível de conformidade AA consiste em práticas de que deveriam ser seguidas, indo além das mais

básicas, e o nível de conformidade AAA, em práticas opcionais, porém melhorariam ainda mais a acessibilidade do sítio.

O governo brasileiro criou o e-MAG – Modelo de Acessibilidade de Governo Eletrônico, dentro dos padrões internacionais: consiste em um conjunto de recomendações a ser considerado para que o processo de acessibilidade dos sítios e portais do governo brasileiro seja conduzido de forma padronizada e de fácil implementação. Criou ainda o ASES, *software* que auxilia o desenvolvedor *Web* na construção de sítios acessíveis, em conformidade com o e-MAG.

Fez-se a avaliação da aderência das páginas HTML coletadas aos padrões de acessibilidade através dos mesmos testes utilizados no ASES. O processo de avaliação consiste da contagem de conformidades das páginas.

Dos 6,3 milhões de páginas HTML coletadas, 98% não apresentaram nenhuma aderência aos padrões de acessibilidade conforme mostrado na Figura 3.5.

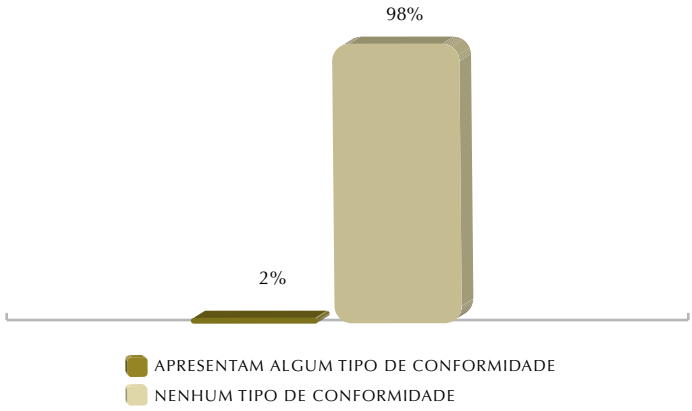

Figura 3.5 – Proporção de páginas aderentes aos padrões de acessibilidade ASES

Tecnologias utilizadas para servir arquivos na *Web* governamental

Dentre os resultados obtidos da coletada de dados do *.gov.br*, pode-se destacar o mapeamento das tecnologias de disponibilização e armazenagem de informações. A seguir, apresenta-se o gráfico relativo à participação das principais tecnologias servidoras de documentos na *Web* governamental (Figura 3.6).

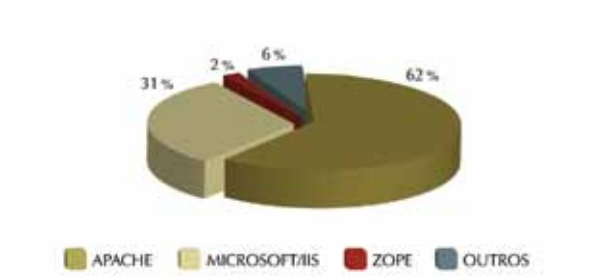

Figura 3.6 – Tecnologias utilizadas para servir arquivos na *Web*

As tecnologias baseadas em *software* de código aberto foram encontradas em mais de 60% das páginas coletadas. Plataformas proprietárias ocupam pouco mais de 30% da fatia de sistemas servidores de documentos na *Web* governamental.

Tecnologias utilizadas para servir arquivos nas cinco regiões brasileiras

Conforme ilustrado no gráfico a seguir, as páginas coletadas sob subdomínios relativos a unidades federativas da região Sul apresentam a maior incidência de servidores de *Web* baseados em sistemas de código aberto, e também o menor percentual de uso de sistemas proprietários, considerando inclusive as páginas do *Governo Federal*.

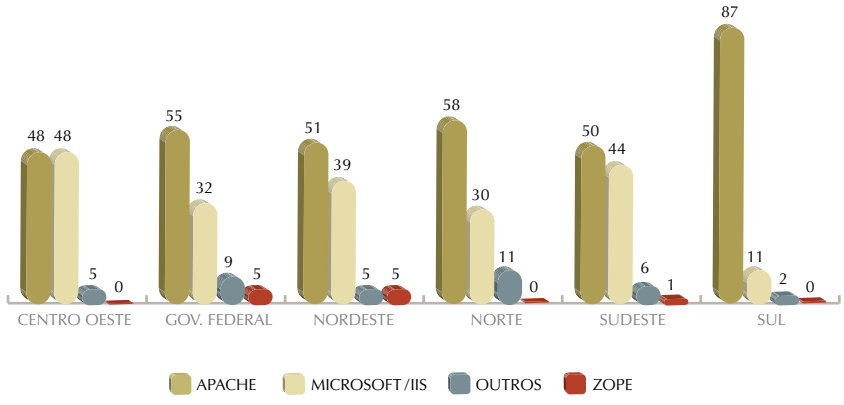

Figura 3.7 – Tecnologias utilizadas para servir arquivos por região

As tecnologias utilizadas para servir arquivos nas UFs

Também foi verificada a utilização do tipo de plataforma servidora por unidade da federação. O gráfico da Figura 3.8 apresenta o uso do tipo de plataforma servidora de *Web*, em relação ao total de sítios daquela UF. O Amapá e o Paraná são os primeiros colocados em uso relativo de sistema de código aberto para servir conteúdo na *Web*. Em relação ao uso de *software* proprietários, verifica-se que o DF é o estado que mais utiliza esse tipo de sistema para servir conteúdo dentre os demais.

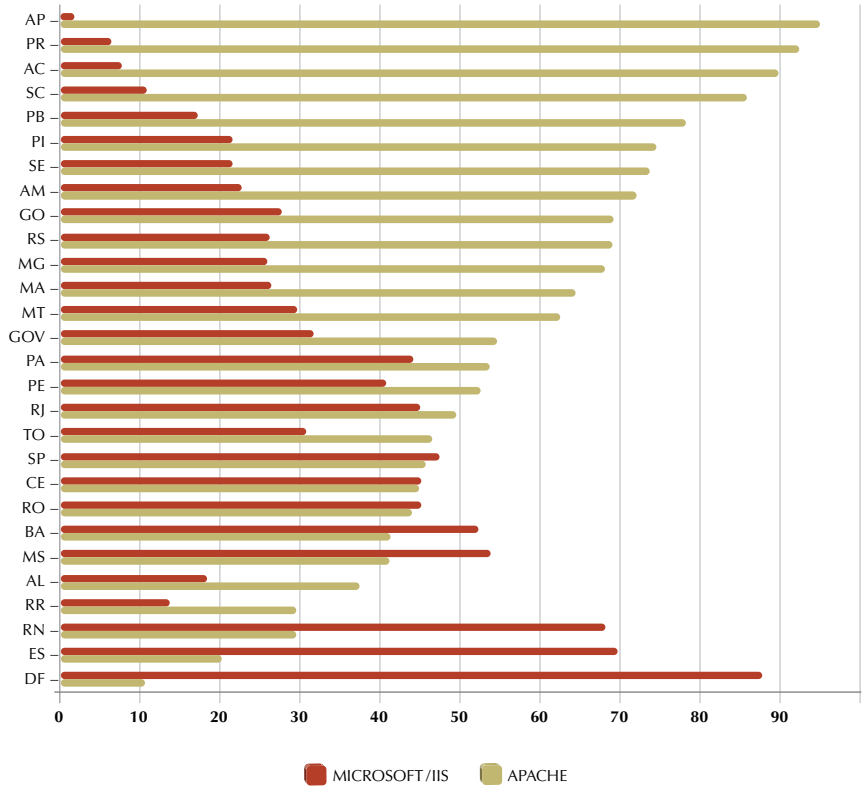

Figura 3.8 – Uso dos principais servidores de *Web* nas UFs brasileiras

Verifica-se, ainda, que a maioria dos sítios do *Governo Federal* está hospedada em servidores Apache, que é uma tecnologia aberta.

Domínios como sítios estruturados em páginas

Somente os domínios com algum conteúdo verificável podem ser considerados como sítios estruturados; para tanto, levantou-se quantos domínios apontavam pelo menos um documento HTML, comumente chamado de página HTML, ou seja, um arquivo que pode ser interpretado por um navegador *Web*, conhecido também como *browser*.

O termo sítio, equivalente a *Website* ou sítio eletrônico, refere-se a um conjunto de páginas HTML referenciadas por um mesmo nome (considerado aqui como o nome de domínio completo) na Internet. Por exemplo, http://www.prefeitura.sp.gov.br (considerou-se como URL tudo o que está depois do http:// e antes da primeira "/"). As URLs http://sítio.prefeitura.sp.gov.br/pagina1.html e http://sítio.prefeitura.sp.gov.br/calendario/evento.html fazem parte do mesmo sítio, para efeito desta pesquisa, enquanto http://www.prefeitura.sp.gov.br refere-se a um sítio diferente.

Do total inicial de 18.796 sítios, apenas 11.586 apresentaram essas características. Efetuou-se também o levantamento do total de arquivos digitais para cada grupo.

A coleta de dados realizada identificou 7.947.607 arquivos digitais; destes, 6.331.256 são documentos em formato HTML, criados ou não por sistemas automatizados de geração de conteúdo. Os demais 1.616.351 arquivos digitais não HTML são arquivos em outros formatos, como: TXT, SWF, EXE, ZIP, RAR.

Ainda segundo a coleta, o número médio de documentos HTML por sítio é de 534 documentos. Todos esses números descrevem de forma sucinta algumas características dos sítios de governo presentes na *Web* brasileira.

Objetos mais usados nas páginas da *Web* governamental

O levantamento indicou que entre todos os 192,2 milhões de *links* encontrados nas páginas da *Web .gov.br*, cerca de 89% correspondem a algum tipo de arquivo gráfico, 8,3% correspondiam a algum tipo de arquivo hipertexto e 2,5% algum tipo de arquivo de texto como .DOC, .PDF, .XML, .ODT ,conforme apresenta a Figura 3.9.

Figura 3.9 - Objetos mais freqüentes na *Web* governamental
Base utilizada: 192.247.032 links *analisados*

Tecnologias utilizadas para disponibilização de dados e de conteúdo na *Web* governamental

As tecnologias empregadas na distribuição de informação de maneira automatizada dos sítios governamentais brasileiro distribuem-se basicamente em dois tipos de tecnologia: PHP e ASP. As tecnologias baseadas em sistemas de código aberto, como o PHP, predominaram no conjunto total das páginas de governo coletadas. 70% das páginas HTML coletadas tinham a extensão .PHP.

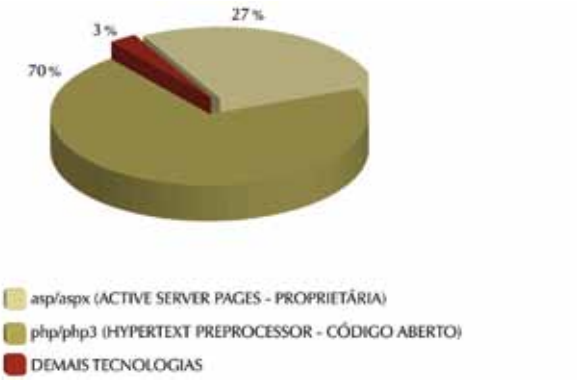

Figura 3.10 – Tipos de tecnologia utilizada para gerar documentos

Dos *links* para objetos gráficos identificados, cerca de 99% apontavam para imagens em formato .GIF, .JPG, .PNG ou .BMP. Os arquivos em formato .PDF representam 80% dentre todos os tipos de documentos coletados; já os arquivos em formato .DOC representam 13%. Esses resultados estão ilustrados nos gráficos das Figuras 3.11 e 3.12.

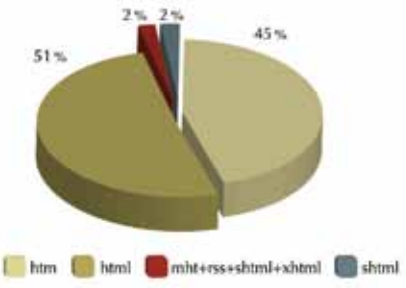

Figura 3.11 - Tipos de hipertexto mais utilizados
Base utilizada: 15.957.331 objetos coletados

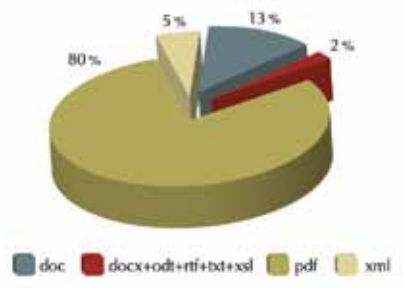

Figura 3.12 - Tipos de documentos mais utilizados
Base utilizada: 4.821.244 objetos coletados

É evidente, portanto, a opção pelo formato .PDF para publicação de documentos.

Dentre os conteúdos editáveis, o formato DOC é o mais publicado. A publicação de formato editável pode não ser uma boa prática, a não ser em casos de publicação de modelos utilizados pelos usuários para o envio de outras informações. Evidente também ainda a baixa utilização de arquivos .XML, formato apropriado para interoperação de dados.

Sincronização de tempo dos servidores brasileiros

A sincronização dos relógios dos servidores, estações de trabalho e outros dispositivos conectados à Internet é importante para o correto funcionamento de muitas aplicações, bem como em situações em que se necessita a análise dos registros (*logs*) feitos pelas aplicações para tratar incidentes de segurança e eventos correlatos. O NIC.br provê um serviço público e gratuito que fornece a Hora Legal Brasileira via Internet, em conjunto com o Observatório Nacional, instituição responsável pela sua definição. Oferece ainda um sítio *Web* com informações e instruções sobre como utilizar esse serviço: o http://ntp.br. O CGI.br recomenda formalmente a sincronização de todos os

dispositivos ligados à rede em sua resolução CGI.br/RES/2008/009/P (http://www.cgi.br/regulamentacao/resolucao2008-009.htm).

Para aferir a sincronização dos servidores que hospedam os sítios .gov.br, obteve-se o horário de seus relógios, via protocolo http comparado com a hora correta. O resultado mostra que apenas pouco mais da metade dos servidores está corretamente sincronizada, e o restante apresenta diferenças em relação à Hora Legal Brasileira entre 1 segundo e até mais de que duas horas, denotando a necessidade de revisão nas configurações.

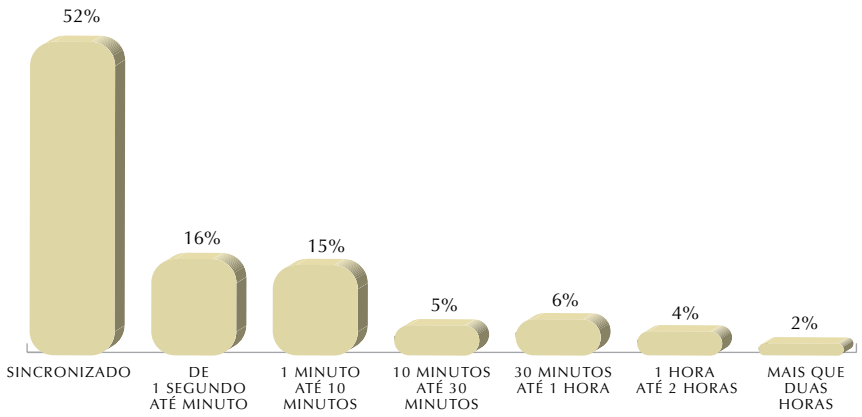

Figura 3.13 – Sincronização de tempo dos servidores

Geolocalização dos IPs

Estimou-se a geolocalização dos servidores com o uso da base de dados GeoLite, da empresa MaxMind[1], consultada a partir de seus endereços IP.

Servidores *Web* hospedados em locais distantes dos usuários implicam uma velocidade menor de acesso, por conta do tempo de tráfego dos pacotes. A hospedagem dos servidores no exterior, além disso, colabora para o aumento dos custos de acesso à Internet no Brasil, já que implica maior utilização dos enlaces internacionais, com custo alto, pelas operadoras de telecomunicações.

Cerca de 6% dos sítios *.gov.br* estão hospedados fora do país.

1 "This product includes GeoLite data created by MaxMind, available from http://maxmind.com/"

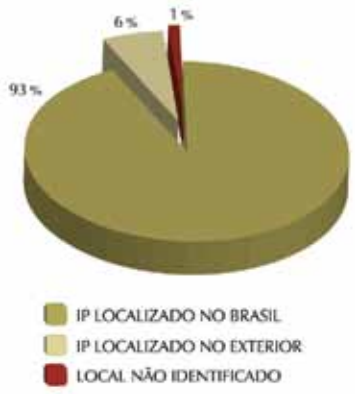

Figura 3.14 – Geolocalização dos IPs
Base: 11.856 sítios com pelo menos um documento HTML

Tempo médio de respostas dos servidores brasileiros

O tempo médio de resposta dos servidores, nesse estudo, consiste no tempo que levaram para responder uma consulta http simples (HEAD), incluindo o tempo de ida e volta dos pacotes de dados, mais o tempo de processamento do servidor. O teste é influenciado, portanto, pela localização do medidor na rede do NIC.br, em São Paulo.

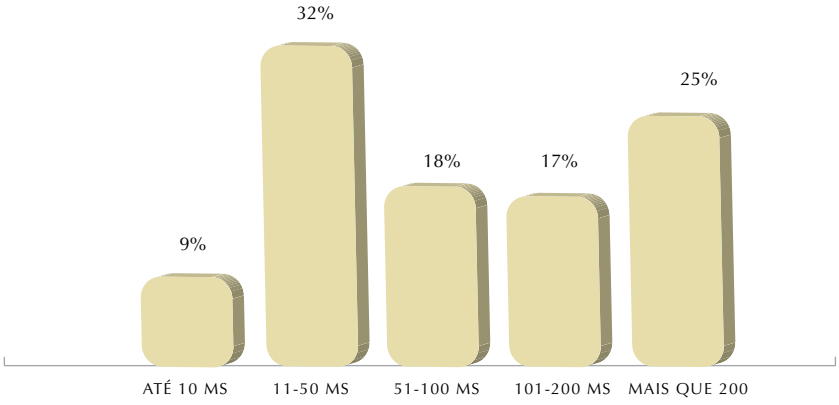

Figura 3.15 – Distribuição do tempo médio de resposta para sítios hospedados no Brasil

De forma simplificada, o indicador pode ser considerado uma medida de desempenho do sítio, do ponto de vista de um usuário localizado em São Paulo. Nota-se como os sítios hospedados fora do Brasil (Figura 3.16) têm

resultados piores do que os hospedados no país. Dos hospedados no país, aproximadamente 59% apresentaram tempos até 100ms, o que é um bom resultado, contudo aponta para a possibilidade de melhoria na infraestrutura dos demais servidores e na própria infraestrutura da Internet brasileira.

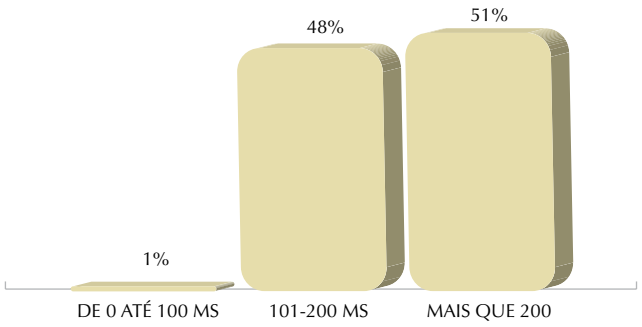

Figura 3.16 – Distribuição do tempo médio de resposta para os sítios hospedados no exterior.

Respostas dos sítios brasileiros de governo a consultas IPV6

O protocolo IP é a base tecnológica que sustenta a Internet; é quem torna possível a utilização inteligente da infraestrutura de telecomunicações, que faz da Internet uma rede muito mais ubíqua, versátil e acessível, em comparação aos serviços convencionais de telecomunicações. Todas as aplicações Internet, inclusive a *Web*, amparam-se, num nível mais básico, nesse protocolo. A informação na Internet divide-se em pacotes que trafegam de forma independente pela rede, e o IP permite que eles encontrem seu caminho, identificando cada dispositivo na Internet com um número, o "endereço IP".

A versão mais usada hoje do protocolo, o IPv4, tem perto de 4 bilhões de endereços possíveis, todavia cerca de 94% desse total já é utilizado. Com o IPv4, a Internet só consegue manter o atual ritmo de expansão por mais 1 ou 2 anos. Por isso, um novo protocolo, o IPv6, será introduzido na rede; ele deverá funcionar lado a lado com o IPv4 durante um período de transição e o substituirá a longo prazo, possibilitando a continuidade da expansão da Internet. Todos devem fazer a implantação de forma célere, pois quando o IPv4 esgotar-se, paulatinamente surgirão serviços e usuários que conseguirão comunicar-se utilizando apenas o IPv6.

O *Governo Federal* mostra compreensão sobre a gravidade da situação ao estabelecer no documento de referência da e-PING: "Os órgãos da Administração Pública Federal deverão se interconectar utilizando IPv4 e planejar sua futura migração para IPv6. Novas contratações e atualizações de redes devem prever suporte à coexistência dos protocolos IPv4 e IPv6 e a produtos que suportem ambos os protocolos.[2]"

Nenhum sítio estava disponível via protocolo IPv6 no censo da *Web para o .gov.br*.

Equipe técnica CETIC.br
Centro de Estudos sobre as Tecnologias da Informação e da Comunicação

[2] Disponível em: http://www.governoeletronico.gov.br/anexos/e-ping-versao-3.0. Acesso em 23 de julho de 2010.

CAPÍTULO 4

Indicadores e universo de dados

Indicadores e universo de dados

Esta seção apresenta os indicadores avaliados unicamente para o universo de domínios .gov.br, apresentando para cada um deles sua definição, propósito, metodologia utilizada para obtenção e apresentação dos resultados. Os indicadores avaliados no contexto dessa etapa do projeto foram os seguintes:

A1: Tamanho total da *Web* brasileira: número de sítios e páginas da *Web*
A2: Tamanho total da *Web* brasileira: tamanho em Gigabytes
B1: Proporção de sítios *Web* utilizando IPv6
B2: Proporção de sítios *Web* utilizando domínio alternativo IPv6 (ipv6.domínio)
B3: Proporção de sítios *Web* respondendo a *Ping* IPv6
B4: Proporção de sítios *Web* que respondem ao comando GET no endereço IPv6
C1: Distribuição do uso de idiomas na *Web* brasileira: proporção de idiomas
E1: Proporção de páginas da *Web* aderentes aos padrões HTML do W3C
F1: Proporção de páginas da *Web* aderentes aos padrões de acessibilidade Ases
G1: Proporção de tipos de objetos usados nas páginas da *Web*: percentual por tipo de objeto
G2: Proporção de tipos de tecnologias usadas nas páginas da *Web*

H1: Idade (ultima atualização) média das páginas da *Web* brasileira
H2: Proporção de páginas dinâmicas na *Web* brasileira
I1: Informação sobre sincronização de tempo dos servidores da *Web* brasileira
I2: Informação sobre tempo de resposta médio dos servidores da *Web* brasileira

Nesse projeto, são variáveis vinculadas à grandeza a ser medida (domínio):

- URL (*Uniform Resource Locator*): URL é um *string* compacto, definido por padrões do W3C que indica a localização de um recurso disponível na Internet. Todos os arquivos disponíveis na *Web* e documentos gerados dinamicamente são capturados via uma URL. Os documentos da *Web* são identificados pelo *substring* http:// no início da URL.
- Página da *Web* (Documento HTML): uma página da *Web* é um recurso escrito na linguagem HTML. Uma página pode conter *links* para outros tipos de recursos, tais como arquivos pdf, arquivos de imagens e recursos que disparam a execução de programas com parâmetros específicos e que geram páginas dinâmicas.
- Sítio *Web*: é o conjunto de uma ou mais URLs associadas a um determinado domínio.

Temos os seguintes universos de dados nesse projeto:

- Universo de sítios da *Web*: todos os sítios encontrados na coleta de dados, independente de terem ou não uma ou mais páginas HTML válidas.
- Universo de sítios da *Web* com páginas HTML válidas: existe pelo menos uma página da *Web* válida, ou seja, obtida com sucesso (código *http-status* igual a 2XX).
- Universo de Páginas HTML válidas: universo de páginas HTML requisitadas e obtidas com sucesso (código *http-status* igual a 2XX).

Existem indicadores relacionados a apenas um universo; outros, a mais de um. No caso dos indicadores gerais de tamanho da *Web*, são gerais, sem vínculo a um ou outro universo. Os indicadores serão apresentados divididos de acordo com o universo ao qual pertencem.

As subseções a seguir apresentam a avaliação de cada um dos indicadores para a coleta de domínios .gov.br.

A1: Tamanho total da *Web* brasileira - número de sítios e páginas da *Web*

Definição do indicador

Total de sítios coletados sob o domínio .gov.br e de suas respectivas páginas, incluindo aquelas páginas fora do domínio .gov.br, redirecionadas a partir de um domínio .gov.br.

Propósito

Identificar o número total de sítios e de páginas da *Web* brasileira para os diferentes universos de dados, ou seja, considerando o valor quantitativo de sítios e de páginas obtidos e aqueles que responderam de forma válida à requisição HTTP realizada (ou seja, tiveram um código de retorno igual a 2XX).

Metodologia

Execução de um *crawler* que percorre as páginas que satisfazem a definição anterior, a partir de um conjunto inicial de sementes fornecidas manualmente.

Apresentação dos resultados

As Tabelas 4.1, 4.2 e 4.3 apresentam os resultados obtidos para esse indicador, utilizando recortes incrementalmente restritos. Enquanto a Tabela 4.1 considera objetos quaisquer encontrados pelo coletor durante o processo, a Tabela 4.2 limita-se ao conjunto de páginas HTML e os servidores que as contêm. Em seguida, a Tabela 4.3 restringe esse conjunto aos sítios dentro do subdomínio .gov.br.

NÚMERO DE OBJETOS E SÍTIOS DA *WEB*		
NÚMERO DE SÍTIOS *WEB*	NÚMERO TOTAL DE OBJETOS DA *WEB*	NÚMERO MÉDIO DE OBJETOS POR SÍTIO
18.796	7.947.607	422,84

Tabela 4.1: Indicador A1 - Número de objetos e sítios da *Web*, considerando todos os objetos encontrados pelo coletor.

NÚMERO DE PÁGINAS HTML E SÍTIOS DA *WEB*		
NÚMERO DE SÍTIOS *WEB*	NÚMERO TOTAL DE PÁGINAS HTML DA *WEB*	NÚMERO MÉDIO DE PÁGINAS HTML POR SÍTIO
12.891	6.334.054	491,35

Tabela 4.2: Indicador A1 - Número de Páginas HTML e sítios da *Web* encontrados pelo coletor .

NÚMERO DE PÁGINAS HTML E SÍTIOS DA *WEB* - .GOV.BR		
NÚMERO DE SÍTIOS *WEB*	NÚMERO TOTAL DE PÁGINAS HTML DA *WEB*	NÚMERO MÉDIO DE PÁGINAS HTML POR SÍTIO
11.856	6.331.256	534,01

Tabela 4.3: Indicador A1 - Número de Páginas HTML e sítios da *Web* encontrados pelo coletor com sufixo .gov.br

SUBDOMÍNIO	SÍTIOS	SUBDOMÍNIO	PÁGINAS	PÁGINAS/SÍTIO
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)**		

Tabela 4.4: Distribuição de Sítios (a) e Páginas (b) do gov.br por Unidade da Federação

SÍTIOS	PÁGINAS
1 (0,01%)	>12000
29 (0,22%)	10000 - 12000
181 (1,40%)	9000 - 10000
61 (0,47%)	8000 - 9000
38 (0,29%)	7000 - 8000
55 (0,43%)	6000 - 7000
51 (0,40%)	5000 - 6000
71 (0,55%)	4000 - 5000
108 (0,84%)	3000 - 4000
176 (1,37%)	2000 - 3000
443 (3,44%)	1000 - 2000
12.891 (100,00%)	TOTAL
(a)	

SÍTIOS	PÁGINAS
70 (0,60%)	900 - 1000
93 (0,80%)	800 - 900
123 (1,05%)	700 - 800
191 (1,64%)	600 - 700
299 (2,56%)	500 - 600
187 (1,60%)	400 - 500
269 (2,30%)	300 - 400
411 (3,52%)	200 - 300
855 (7,32%)	100 - 200
9.179 (78,61%)	<100
11.677 (100,00%)	TOTAL
(b)	

Tabela 4.5: Distribuição de páginas por sítio na coleta realizada, para todos os sítios (a) e para sítios com menos de 1000 páginas (b).

A2: Tamanho total da *Web* brasileira - tamanho em Gigabytes

Definição do indicador

Soma do tamanho das páginas sob o domínio .gov.br., considerando as premissas adotadas nesse projeto.

Propósito

Calcular o volume ocupado pelos sítios *Web* e páginas da *Web* brasileira.

Metodologia

Execução de um *crawler* que percorre as páginas que satisfazem a definição anterior, a partir de um conjunto inicial de sementes fornecidas manualmente.

Apresentação dos resultados

As Tabelas 4.6, 4.7 e 4.8, apresentam os resultados obtidos para esse indicador, utilizando recortes incrementalmente restritos. Enquanto a Tabela 4.6 considera objetos quaisquer encontrados pelo coletor durante o processo, a Tabela 4.7 limita-se ao conjunto de páginas HTML e os servidores que as contêm. Em seguida, a Tabela 4.8 restringe esse conjunto aos sítios dentro do subdomínio .gov.br.

TAMANHO DA *WEB* .GOV.BR.		
TAMANHO DA *WEB* (VOLUME EM GB)	TAMANHO MÉDIO DOS SÍTIOS *WEB* (EM MB)	TAMANHO MÉDIO DAS PÁGINAS DA *WEB* (EM KB)
169,7	9,43	24,17

Tabela 4.6: Indicador A2: Tamanho Total da *Web* sob o domínio .gov.br.
(Tamanho em *GigaBytes*)

VOLUME EM *BYTES* DAS PÁGINAS COLETADAS		
TAMANHO DA *WEB* (VOLUME EM GB)	TAMANHO MÉDIO DOS SÍTIOS *WEB* (EM MB)	TAMANHO MÉDIO DAS PÁGINAS DA *WEB* (EM KB)
148,37	11,79	24,56

Tabela 4.7: Volume em bytes nas páginas HTML coletadas
e número de sítios encontrados pelo coletor

VOLUME EM *BYTES* DAS PÁGINAS COLETADAS - .GOV.BR		
TAMANHO DA *WEB* (VOLUME EM GB)	TAMANHO MÉDIO DOS SÍTIOS *WEB* (EM MB)	TAMANHO MÉDIO DAS PÁGINAS DA *WEB* (EM KB)
148,33	12,81	24,57

Tabela 4.8: Volume em bytes nas páginas HTML coletadas e número de sítios
onde estas páginas foram encontradas, considerando somente sítios com sufixo .gov.br

SUBDOMÍNIO	VOLUME (GB)	PÁGINAS	VOLUME/PÁGINAS(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

Tabela 4.9: Distribuição do volume em *Gigabytes* no domínio .gov.br. por Unidade da Federação

SITIOS	TAMANHO
1 (0,01%)	>2.100 M
1 (0,01%)	1.300 -2.100 M
2 (0,02%)	1.000 -1300 M
1 (0,01%)	800 -1.000 M
2 (0,02%)	750 -800 M
1 (0,01%)	700 -750 M
2 (0,02%)	650 -700 M
1 (0,01%)	600 -650 M
8 (0,06%)	550 -600 M
4 (0,03%)	500 -550 M
11 (0,09%)	450 -500 M
11 (0,09%)	400 -450 M
9 (0,07%)	350 -400 M
22 (0,17%)	300 -350 M
54 (0,42%)	250 -300 M
60 (0,47%)	200 -250 M
99 (0,77%)	150 -200 M
138 (1,07%)	100 -150 M
227 (1,76%)	50 -100 M
12.237 (94,93%)	<50 M
12.891 (100,00%)	TOTAL
(a)	

SITIOS	TAMANHO
22 (0,18%)	48 -50 M
26 (0,21%)	46 -48 M
22 (0,18%)	44 -46 M
20 (0,16%)	42 -44 M
28 (0,23%)	40 -42 M
31 (0,25%)	38 -40 M
23 (0,19%)	36 -38 M
27 (0,22%)	34 -36 M
35 (0,29%)	32 -34 M
34 (0,28%)	30 -32 M
51 (0,42%)	28 -30 M
44 (0,36%)	26 -28 M
63 (0,51%)	24 -26 M
36 (0,29%)	22 -24 M
69 (0,56%)	20 -22 M
71 (0,58%)	18 -20 M
74 (0,60%)	16 -18 M
105 (0,86%)	14 -16 M
120 (0,98%)	12 -14 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
12.237 (100,00%)	TOTAL
(b)	

Tabela 4.10: Distribuição do volume em bytes por sítio
na coleta realizada para todos os sítios (a) e para sítios com menos de 50 MBytes (b).

C1: Distribuição do uso de idiomas na *Web* brasileira - Proporção de idiomas

Definição do indicador

Valor percentual da quantidade de páginas do domínio .gov.br., de acordo com uma relação pré-definida de idiomas.

Propósito

Obter uma distribuição da quantidade relativa de páginas do domínio .gov.br., de acordo com o seu idioma.

Metodologia

Execução de um *crawler* que percorre as páginas que satisfazem a definição anterior, a partir de um conjunto inicial de sementes fornecidas manualmente. O *crawler* utilizado baseia-se na frequência de ocorrência de palavras em um dado idioma, de acordo com dicionários pré-construídos de um conjunto de idiomas pré-determinado.

Apresentação dos resultados

A Tabela 4.11 apresenta a distribuição de quatro idiomas pré-definidos nas páginas do domínio .gov.br: Português, Inglês, Espanhol, e Francês. Cabe ressaltar que existe um universo de páginas HTML para as quais não foi possível identificar o idioma a partir da técnica utilizada. Esse universo corresponde a 2.912.597 (47,8% do total de páginas).

	PERCENTUAL DE PÁGINAS DA *WEB* PARA CADA TIPO DE IDIOMA DE UM CONJUNTO PRÉ-DETERMINADO		
DISTRIBUIÇÃO DOS IDIOMAS UTILIZADOS PELAS PÁGINAS DO DOMÍNIO	Português	3.088.680	97,05 %
	Inglês	80.726	2,54 %
	Espanhol	10.623	0,33 %
	Francês	2.623	0,08 %

Tabela 4.11: Distribuição dos Idiomas das páginas no domínio .gov.br

E1: Proporção de páginas da *Web* aderentes aos padrões HTML do W3C

Definição do indicador

Valor percentual de páginas HTML da *Web* brasileira ".br" que atendem aos padrões W3C, de acordo com o seu tipo de documento.

Propósito

Avaliar a qualidade das páginas HTML da *Web* brasileira ".br" em relação à conformidade com o padrão HTML especificado pelo W3C.

Metodologia

Foi executado um validador W3C de documentos que identifica o tipo de documento e informa o número de erros obtidos de acordo com esse tipo. O validador de documentos retorna o número total de erros obtidos a partir da análise de concordância com as normas do W3C.

Apresentação dos resultados

A partir da validação das páginas da *Web* feita com o programa validador do W3C, foi realizada a consolidação dos valores retornados pelo validador, indicando o número de incorreções encontrado na página.

A Tabela 4.12 apresenta os resultados gerais de validação das páginas *Web*, utilizando a ferramenta de validação da W3C.

VALOR RETORNADO PELA FERRAMENTA	VALOR ABSOLUTO	VALOR PERCENTUAL (%)
NÃO FOI POSSÍVEL VALIDAR	267.137	4,24
PÁGINAS VÁLIDAS	316.501	5,02
APRESENTAM INCORREÇÕES >0	5.717.315	90,74

Tabela 4.12: Quantidade e percentual de páginas da *Web* governamental aderentes aos padrões W3C

QUANTIDADE DE INCORREÇÕES	PÁGINAS DA *WEB*	
	VALOR ABSOLUTO	VALOR PERCENTUAL (%)
≤ 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

Tabela 4.13: Aderência da *Web* governamental aos padrões W3C – Distribuição das incorreções por faixa

UF	VALIDAÇÃO W3C (PÁGINAS HTML DA *WEB*)		
	NÃO CONFORMIDADE	CONFORMI-DADE	% CONFORMIDADE
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
Outros domínios	955.212	74.199	7,2

Tabela 4.14: Aderência da *Web* governmanetal aos padrões W3C - Recorte por Unidade Federativa

F1: Proporção de Páginas da *Web* aderentes aos padrões de acessibilidade ASES

4.5.1 Definição do indicador

Valor percentual de Páginas HTML válidas, compatíveis com os padrões determinados de acessibilidade, considerando os níveis de conformidade A, AA, AAA.

4.5.2 Propósito

Avaliar a qualidade das páginas HTML em relação à conformidade com os padrões de acessibilidade WCAG 1.0 (W3C) e eMAG (Governo Brasileiro).

4.5.3 Metodologia

Para se avaliar a acessibilidade, realizaram-se:

- a coleta dos dados de páginas, seguindo o procedimento padrão de coleta adotado nesse projeto.
- a execução do validador ASES de acessibilidade, que atribui um valor de acessibilidade (A, AA, AAA ou não conformidade) para cada página HTML coletada.

Para definição das formas de avaliação da acessibilidade, adotou-se como documento de referência o WCAG 1.0 - *Web Content Accessibility Guidelines 1.0*, para explicitar as conformidades de acessibilidade de A, AA e AAA. Para isso, foram utilizados os níveis de prioridade e a definição descrita a seguir.

O grupo de trabalho atribuiu a cada ponto de verificação um nível de prioridade, com base no respectivo impacto, em termos de acessibilidade. Esses níveis são descritos a seguir:

- Prioridade 1: Pontos que os criadores de conteúdo *Web* devem satisfazer inteiramente. Se não o fizerem, um ou mais grupos de usuários ficarão impossibilitados de acessar as informações contidas no documento. A satisfação desse tipo de pontos é um requisito básico para que determinados grupos possam acessar documentos disponíveis na *Web*.
- Prioridade 2: Pontos que os criadores de conteúdos na *Web* deveriam satisfazer. Se não o fizerem, um ou mais grupos de usuários terão dificuldades em acessar as informações contidas no documento. A satisfação desse tipo

de pontos promoverá a remoção de barreiras significativas ao acesso a documentos disponíveis na *Web*.

- Prioridade 3: Pontos que os criadores de conteúdos na *Web* podem satisfazer. Se não o fizerem, um ou mais grupos poderão ter dificuldades para acessar informações contidas nos documentos. A satisfação desse tipo de pontos irá melhorar o acesso a documentos armazenados na *Web*.

Alguns pontos de verificação especificam um nível de prioridade que poderá mudar sob determinadas condições (explicitadas). Assim, as conformidades de acessibilidade para cada página da *Web* ficaram definidas da seguinte forma:

- Nível de conformidade "A": foram satisfeitos todos os pontos de verificação de prioridade 1;
- Nível de conformidade "AA": foram satisfeitos todos os pontos de verificação de prioridades 1 e 2;
- Nível de conformidade "AAA": foram satisfeitos todos os pontos de verificação de prioridades 1, 2 e 3;
- Não conformidade: não foram satisfeitos nenhum ponto de verificação por completo; logo, não existe conformidade para a página da *Web* analisada.

Cabe ressaltar que a página *Web* de nível A não é nem AA e nem AAA, bem como AA não é AAA.

Apresentação dos resultados

A Tabela 4.15 apresenta os resultados obtidos para o indicador de acessibilidade, considerando páginas HTML da *Web* do universo .gov.br. Os dados apresentados na tabela são referentes a 6.279.206 páginas HTML. Outras 54.848 páginas (0,86%) não foram classificadas, uma vez que o validador não retornou um resultado esperado.

A Tabela 4.17 apresenta os resultados obtidos para o indicador de acessibilidade, considerando páginas HTML, fazendo um recorte por Unidade Federativa (UF), realizado a partir da identificação da UF na URL da página HTML (http://...uf.gov.br/).

	CONFORMIDADE COM OS NÍVEIS DE PRIORIDADE (PÁGINAS HTML DA *WEB*) - RECORTE POR UNIDADE			
	PRIORIDADE 3	PRIORIDADE 2	PRIORIDADE 1	NÃO CONFORMIDADE
QUANTITATIVO	39.440	14.662	71.628	6.153.476
PERCENTAGEM	0,63	0,23	1,14	98,00

Tabela 4.15: Indicador F1 – Conformidade das páginas *Web* governamental com os níveis de prioridade (Páginas HTML da *Web*)

UF	Níveis de prioridade (páginas HTML da *Web*)				
	3	2	1	Não Conformidade	
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%
Outros domínios	2.508	7.937	47.390	1.139.676	95,17%

Tabela 4.16: Conformidade com os níveis de prioridade (Páginas HTML da *Web*) - recorte por Unidade Federativa

G1: Proporção de tipos de objetos usados nas páginas da *Web* - percentual por tipo de objeto

Definição do indicador

Valor percentual dos tipos de objetos usados nas páginas da *Web* brasileira ".br", de acordo com uma classificação categórica (imagens, *scripts*, vídeos etc.).

Propósito

Obter uma distribuição dos tipos de objetos usados nas páginas da *Web* brasileira ".br", de acordo com uma categoria pré-definida (imagens, *scripts*, vídeos, etc.).

Metodologia

As páginas foram coletadas usando o *Web crawler*, considerando tanto a URL de cada página coletada quanto as URLs presentes em cada página coletada. Todas as extensões foram convertidas para caixa baixa. A taxonomia de tipos de documentos foi extraída da e-Ping, Padrões de Interoperabilidade de Governo Eletrônico, Documento de Referência Versão 2.0, 11 de Dezembro de 2009.

Apresentação dos resultados

Os resultados estão na Tabela 4.17.

GRUPO	QUANTIDADE POR GRUPO	%	TIPO	QUANTIDADE POR TIPO	%
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
DOCUMENTO	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
PLANILHA	156.623	0.08	ods	331	0.21
			xls	156.240	99.76
			xlsx	52	0.03
APRESENTAÇÃO	28.533	0.01	odp	158	0.55
			ppt	28.302	99.19
			pptx	73	0.26
BANCO DE DADOS	6.531	0.00	csv	6.405	98.07
			myd	63	0.96
			myi	63	0.96
GRÁFICOS	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
ÁUDIO E VÍDEO	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			

Tabela 4.17: Quantidade e percentual de objetos nas páginas HTML, por tipos de documentos

G2: Proporção de tipos de tecnologias usadas nas páginas da *Web* - percentual por tipo de tecnologia

Definição do indicador

Valor percentual dos tipos de tecnologias usadas nas páginas da *Web* brasileira ".br".

Propósito

Obter uma distribuição dos tipos de linguagens usadas nas páginas da *Web* brasileira ".br", de acordo com uma lista de valores pré-determinados (PHP, ASP, ASPX, JSF, JSP, etc.).

Metodologia

A determinação de tecnologias usadas é um desafio, porque uma página coletada não possui obrigatoriamente informações sobre a tecnologia que a gerou. Uma opção adotada foi se basear nas eventuais extensões de arquivo presentes na URL.

Para determinar as tecnologias, partiu-se de um dicionário de 406 extensões de arquivos e processamos o arquivo de páginas válidas (OK), verificando em cada URL listada naquele arquivo a ocorrência de uma extensão válida. Uma extensão válida deve ocorrer antes da primeira "?" da URL e a partir da ultima "/" que antecede essa "?"´ . O processo verifica então, por casamento de padrões, a ocorrência das extensões na cadeia delimitada por "/" e "?" da URL. Para extensões que tenham o mesmo radical (p.ex., asp e aspx), considera-se a mais longa.

O ponto de partida da metodologia foi apurar quais as possíveis extensões, como medida das tecnologias utilizadas. Nesse caso, buscou-se uma lista de 406 extensões de arquivos, a partir do sítio http://www.file-extensions.org. Com base nessas extensões, analisou-se a URL de cada página coletada, de forma a identificar quais extensões ocorriam na URL.

Feita a identificação de extensões, há três casos possíveis. O primeiro caso: nenhuma extensão encontrada na URL, o que impede estimar qual a tecnologia utilizada. O segundo caso: há exatamente uma extensão, caso no qual a tecnologia, se for o caso, é associada diretamente. O terceiro caso: mais

de uma extensão associada à URL e se faz necessário estimar qual extensão detectada é a mais pertinente.

Utilizou-se dois critérios para se detectar a extensão mais pertinente. O primeiro critério é a posição onde a extensão ocorre na URL, com base na premissa de que a extensão do primeiro arquivo que ocorre identifica a sua tecnologia base. Esse critério foi validado em uma porção significativa dos casos. O segundo critério é, para extensões que ocorram na mesma posição, escolher a maior, por ser naturalmente mais discriminativa. Por exemplo, considerar que as extensões php e php3 são detectadas a partir da mesma posição em uma URL, o que se explica pelo fato de php ser parte de php3. Nesse caso, a extensão selecionada será php3, pois ela é a maior e a mais discriminativa.

O último passo da metodologia é selecionar, dentre as extensões identificadas, aquelas que são associadas à tecnologias. Este processo é feito manualmente, verificando as extensões que efetivamente ocorreram e as suas respectivas descrições.

Apresentação dos resultados

Os resultados são apresentados na Tabela 4.18.

TEC	QUANTIDADE	%	DESCRIÇÃO
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**	

Tabela 4.18: Quantidade e proporção de tecnologias utilizadas na *Web* brasileira

H1: Idade (última atualização) média das páginas da *Web* brasileira

Definição do indicador

Valor médio da idade das páginas da *Web* brasileira ".br", considerando a data da última atualização da página da *Web*.

Propósito

Obter a idade média das páginas da *Web*, considerando a sua última data de atualização.

Metodologia

O *software* de coleta utilizado procura registrar a idade das páginas coletadas, indicando a diferença entre a data e hora em que uma URL é coletada e a data e hora reportadas pelo servidor, por meio da última atualização da página em questão. Essa informação (data da ultima atualização de cada página) não é fornecida por todos os servidores, nem para todo tipo de conteúdo. Por não se tratar de informação obrigatória, muitas vezes ela não está presente na coleta. Além disso, erros na configuração da hora nos servidores *Web* podem levar a erros na informação de data e hora por eles fornecida. Nos dados da coleta, páginas para as quais a informação de data de alteração não foi fornecida ficaram sem registro de idade.

As páginas com informação de idade foram consideradas em termos de dias, a fim de se simplificar a análise.

Apresentação dos resultados

O resultado é apresentado no quadro a seguir.

IDADE MÉDIA DAS PÁGINAS	IDADE MÉDIA DAS PÁGINAS DA *WEB* GOVERNAMENTAL BRASILEIRA
	656 dias

Observação: do total de 6.331.256, pouco menos de 10% (614.770) apresentaram informação de idade válida.

CONJUNTO	TOTAL DE PÁGINAS	QUANTIDADE DE PÁGINAS C/ IDADE	%	IDADE MÉDIA EM DIAS
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 estados	4.823.688	427.997	9	607,55
Outros .gov.br	1.507.568	186.773	12	768,31
Total .gov.br	6.331.256	614.770	10	656,24
Outras páginas	2.798	1.110	40	440,07
TOTAL	**6.334.054**	**615.880**	**10**	**655,85**

Tabela 4.19: Idade das páginas da *Web* governamental brasileira por estado

H2: Proporção de páginas dinâmicas na *Web* brasileira

Definição do indicador

Valor percentual de páginas consideradas dinâmicas na *Web* governamental brasileira (.gov.br). Uma página dinâmica em geral é a referência do uso de linguagens de programação *server-side*, tal como PHP, ASP, JSP, ColdFusion entre outras, no desenvolvimento de um sítio ou de aplicações para intranet e extranet. Ela recebe esse nome por ter sido gerada em tempo de execução, produzindo o conteúdo estático que o usuário visualiza no momento de sua solicitação, via requisição HTTP.

Propósito

Ter uma medida percentual da quantidade de conteúdo dinâmico gerado a partir das páginas da *Web* governamental brasileira (.gov.br).

Metodologia

O coletor utilizado tem um conjunto de regras internas para determinar se uma página é dinâmica ou estática. Essas regras consideram o tipo de terminação utilizada para o arquivo de conteúdo (por exemplo, terminações como .jsp ou .php são associadas a documentos dinâmicos), bem como a existência de parâmetros associados à URL.

Com base nessa informação, o coletor armazena, para cada página consultada com sucesso, a natureza do conteúdo a ela associado (estático ou dinâmico).

Apresentação dos resultados

O resultado é apresentado no quadro a seguir.

PERCENTUAL DE PÁGINAS DINÂMICAS	PERCENTUAL DE PÁGINAS DINÂMICAS DA *WEB* GOVERNAMENTAL BRASILEIRA
	74,8 %

CONJUNTO (XX.GOV.BR)	TOTAL DE PÁGINAS ENCONTRADAS	TOTAL DE PÁGINAS DINÂMICAS	PORCENTAGEM
.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 estados	4.851.779	3.863.533	80
Outros .gov.br	1.479.477	873.269	59
Total .gov.br	6.331.256	4.736.802	75
Outras págs	2.798	1.499	54
Total	6.334.054	4.738.301	75

Tabela 4.20: Porcentagem de páginas dinâmicas na *Web* governamental bra**sileira**

B1: Proporção de sítios *Web* utilizando IPv6

Definição do indicador

Valor percentual de sítios *Web* no universo de servidores que respondem pelas páginas da *Web* brasileira ".br", preparados operacionalmente para responder seguindo o protocolo IPv6.

Propósito

Ter uma medida da atual quantidade de servidores *Web* operacionalmente prontos para se comunicarem utilizando IPv6.

Metodologia

Execução de consulta específica para o protocolo IPv6 aos servidores *Web*. A resposta indica se o servidor está operacionalmente preparado para responder a requisição IPv6.

Apresentação dos resultados

Somente 4 (quatro) dos 12.891 *hosts* da primeira coleta filtrada respondem ao protocolo IPv6. Os *hosts* estão listados na Tabela 7.1. Observa-se que, segundo a definição do Projeto Censo *Web*, todos os *hosts* coletados a partir de redirecionamentos da *Web* governamental brasileira são considerados como pertencentes a esse subconjunto da *Web*. Portanto, os *hosts* listados a seguir, embora não sejam do subdomínio .gov.br, são considerados, para efeito desse Projeto, pertencentes à *Web* governamental brasileira.

A Tabela 4.21 apresenta a relação de *hosts* que responderam à consulta via protocolo IPv6.

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

Tabela 4.21: *Hosts* que responderam ao Protocolo IPv6

Portanto, o percentual dos *hosts* que responderam a consulta realizada via protocolo IPv6 na *Web* governamental brasileira é de 4/12.891 = 0, 031%.

B2: Proporção de sítios *Web* utilizando domínio alternativo IPv6 (ipv6.dominio)

Definição do indicador

Valor percentual de sítios *Web* no universo de servidores que respondem pelas páginas da *Web* brasileira ".br", que atendem à requisição IPv6 em um domínio alternativo (ipv6.dominio).

Propósito

Ter uma medida da atual quantidade de servidores *Web* testados para comunicar-se utilizando IPv6.

Metodologia

Programa de *software* específico para realizar uma consulta IPv6 aos servidores *Web*, a fim de obter uma resposta que permita saber se está operacionalmente preparado para responder a requisição IPv6.

Apresentação dos resultados

Nenhum sítio respondeu a consulta ao domínio alternativo. Portanto, a proporção é zero.

B3: Proporção de sítios *Web* respondendo a ping IPv6

Definição do indicador

Valor percentual de sítios *Web* no universo de servidores que hospedam as páginas da *Web* brasileira ".br", que respondem a um PING nos endereços IPv6.

Propósito

Ter uma medida da atual quantidade de servidores *Web* ativos, que respondam a PING no domínio IPv6.

Metodologia

Programa de *software* específico para realizar uma consulta IPv6 aos servidores *Web* e obter uma resposta que permita saber se o servidor está operacionalmente preparado para responder a requisições IPv6.

Apresentação dos resultados

URL
www.itu.int

Tabela 4.22: *Hosts* que responderam ao PING via protocolo IPv6

Identificou-se que apenas um sítio da *Web* governamental brasileira respondeu a um PING nos endereços IPv6. A proporção é, então, 1/12.891 = 0, 008%.

B4: Proporção de sítios *Web* que respondem ao comando GET no endereço IPv6

Definição do indicador

Valor percentual de sítios *Web* no universo de servidores que hospedam as páginas da *Web* brasileira ".br", que respondem a um comando GET na porta 80 do endereço com protocolo IPv6.

Propósito

Ter uma medida da atual quantidade de servidores *Web* ativos e respondendo do GET na porta 80 do endereço IPv6.

Metodologia

Programa de *software* específico para realizar uma consulta IPv6 aos servidores *Web*, a fim de obter uma resposta que permita saber se está operacionalmente preparado para responder a requisição IPv6.

Apresentação dos resultados

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

Tabela 4.23: Hosts que responderam ao HTTP GET via protocolo IPv6

I1: Informação sobre sincronização de tempo dos servidores da *Web* brasileira

Definição do indicador

Valor estimado da diferença de sincronização de tempo dos servidores da *Web* brasileira (.gov.br) em relação a hora certa mundial, conhecida como tempo UTC (*Coordinated Universal Time*).

Propósito

Estimar o grau de sincronismo dos servidores da *Web* governamental brasileira (.gov.br) em relação a hora certa mundial.

Os computadores podem sincronizar o tempo, utilizando um servidor de tempo e um protocolo. Normalmente, adota-se o NTP (*Network Time Protocol*), que converte o tempo para uma linguagem compreensível ao servidor. Esse mecanismo é fundamental para garantir o correto registro das transações realizadas na *Web*, bem como as diferentes comunicações que ocorrem entre servidores na rede.

Metodologia

Foi realizada uma requisição HTTP ao servidor pelo método HEAD. O servidor respondeu com a data e a hora no campo *Date*. Foi medido o RTT (*round-trip time*) da consulta. A hora marcada pelo servidor foi estimada da seguinte forma: tempo dado pelo campo *Date* somado à metade do RTT. Observa-se que o RTT é dado em milissegundos e o campo *Date*, em segundos. A estimativa de sincronização foi feita pelo cálculo do módulo da diferença entre o tempo estimado do servidor e o tempo marcado no relógio

da máquina que fez o experimento, sincronizada via NTP. O resultado final é dado em segundos.

Apresentação dos resultados

Foram obtidos 12836 tempos válidos, dos 12891 servidores consultados. A Tabela 4.24 apresenta as principais estatísticas referentes aos tempos encontrados.

MIN	MEDIANA	MÉDIA	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

Tabela 4.24: Diferença absoluta entre a hora do servidor e o UTC em segundos

UF	MEDIANA	MÉDIA	MÁXIMO	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

Tabela 4.25: Indicador I1 - Estatísticas da sincronização por unidade da federação: tempo em segundos

I2: Informação sobre tempo de resposta médio dos servidores da *Web* brasileira

Definição do indicador

Valor do tempo de resposta médio para os servidores da *Web*, considerando cada sítio *Web* identificado na coleta de dados do universo .gov.br.

Propósito

Este indicador visa oferecer uma noção acerca do tempo de resposta médio dos sítios da *Web* do universo .gov.br.

Metodologia

Para realizar a coleta da informação de tempo de resposta de um determinado sítio da *Web*, é feita uma consulta específica ao servidor do sítio *Web*, onde se registra o tempo gasto (em unidade milisegundos) para concretizar a resposta do servidor.

Portanto, trata-se de um método simples, que fornece apenas uma ideia aproximada do tempo necessário para acesso ao servidor, contudo permite ter uma avaliação geral acerca desse indicador de qualidade no tempo de resposta a uma requisição.

Apresentação dos resultados

A Tabela 4.26 apresenta os resultados obtidos para o indicador de tempo de resposta médio para os sítios da *Web* do universo .gov.br.

Os dados apresentados na tabela são referentes a 12.871 sítios que tiveram pelo menos uma página HTML com resposta válida. Outros 20 sítios (0.15%) não foram contemplados nessa análise, visto que a consulta de tempo de resposta a eles não obteve resultado (o que ocorre devido ao servidor do sítio não aceitar este tipo de consulta ou a algum erro de indisponibilidade).

TEMPO DE RESPOSTA (MILISEGUNDOS)	SÍTIOS *WEB*	
	VALOR ABSOLUTO	PERCENTAGEM
≤ 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

Tabela 4.26: Tempo de resposta médio dos sítios *Web*

MÉDIA	MIN	MAX	MEDIANA	DESVIO PADRÃO	CV	Q1	Q2	Q3	P90	P97	P99
190,20	1	8313	71	368,78	1,94	27	71	201	475	1049	1595

Tabela 4.27: Tempo de Resposta - Análise Estatística

D2: Proporção de países que hospedam os sítios *Web* brasileiros

Definição do indicador

Valor percentual da quantidade de sítios da *Web* brasileira de acordo com o país que é hospedeiro desse sítio.

Propósito

Obter uma distribuição percentual dos sítios da *Web* brasileira de acordo com o país que o hospeda

Metodologia

Estimou-se a geolocalização dos servidores com o uso da base de dados da GeoLite, da empresa MaxMind, consultada a partir de seus endereços IP.

Apresentação dos resultados

A tabela 4.28 apresenta os resultados para o indicador, obtidos a partir de um universo de 11.856 sítios com domínios ".gov.br" e com pelo menos um documento HTML válido.

LOCALIZAÇÃO	BRASIL	EXTERIOR	NÃO IDENTIFICADO
PROPORÇÃO DE SÍTIOS HOSPEDADOS	93%	6%	1%

Tabela 4.28: Proporção dos servidores hospedados no Brasil e em outros países