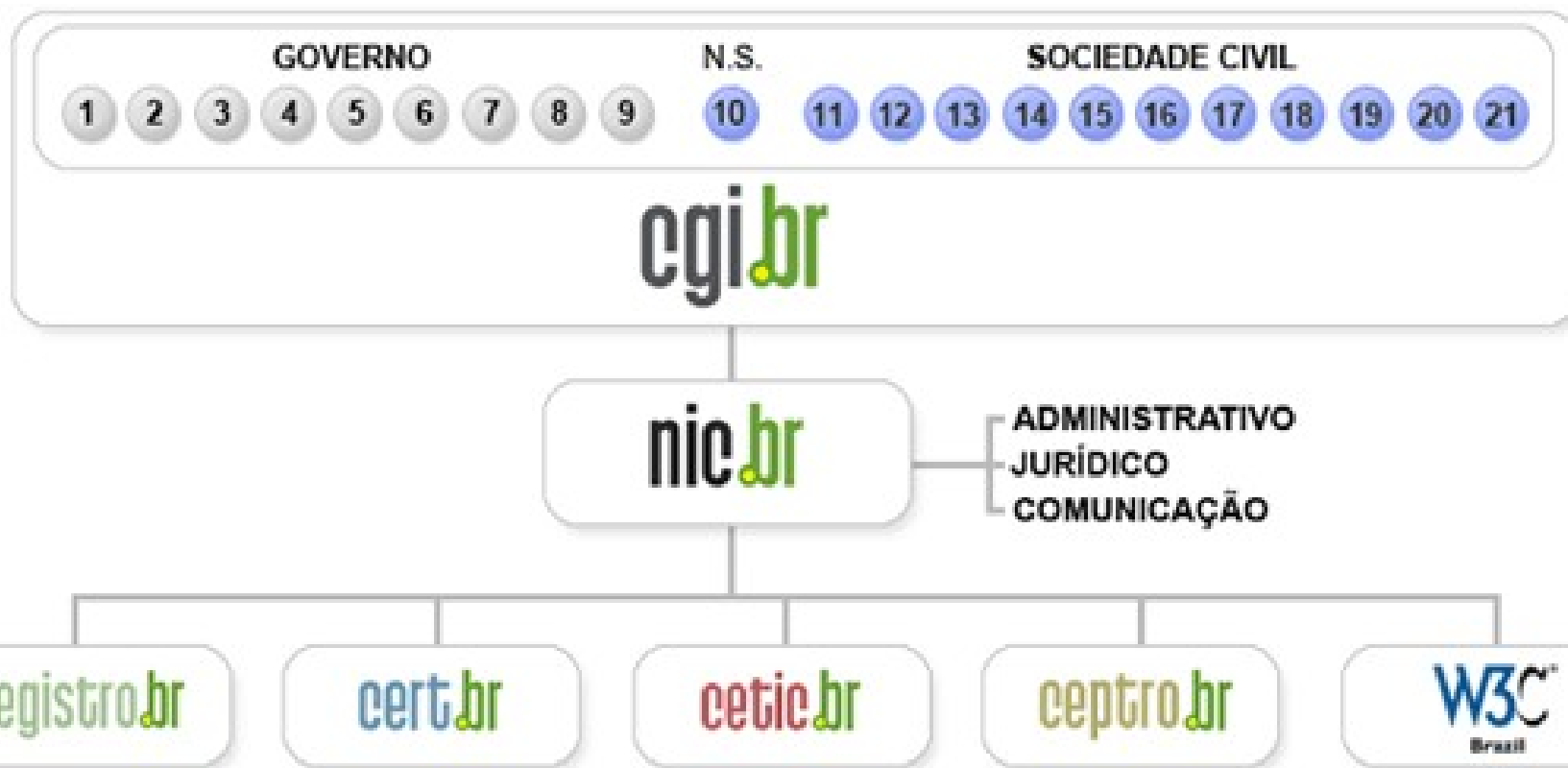


dimensões e características da

Web

brasileira: um estudo do .gov.br



Agenda:

- Introdução
- Objetivos da pesquisa
- Desafios técnicos para o estudo da Web

Parceiros

winweb

Instituto Nacional de Ciência e Tecnologia para a Web



Secretaria de Logística e
Tecnologia da Informação

Colaboradores:

- ATI – Agência de Tecnologia da Informação. Governo do Estado de Pernambuco
- Caixa Econômica Federal
- CIASC – Centro de Informática e Automação do Estado de Santa Catarina
- FEA – faculdade de Economia e Administração, Universidade de São Paulo
- PRODEB – Companhia de Processamento de Dados do Estado da Bahia
- PRODERJ – Centro de Informação e Comunicação do Rio de Janeiro
- Secretaria de Gestão Pública do Estado de São Paulo
- SERPRO – Serviço Federal de Processamento de Dados

Introdução

O que é o Projeto Censo da Web

- >> *Coleta de páginas html*
- >> *Análise de suas características*



<http://vperreiro.files.wordpress.com/2010/05/dominio.jpg>

O objetivo geral deste projeto é criar e divulgar indicadores de todos os sítios hospedados sob o domínio “.br”, considerando os requisitos definidos por um conselho consultivo representando diversos segmentos interessados, respeitando as boas práticas de privacidade e confidencialidade

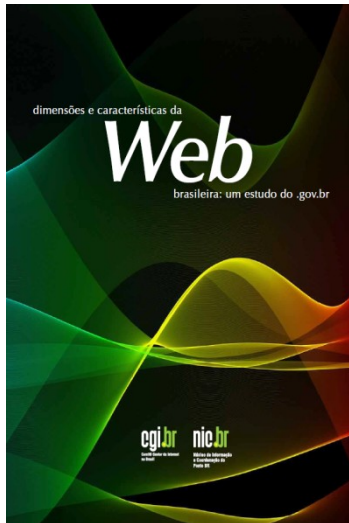
Objetivos específicos

- Tamanho total da Web brasileira: número de sítios e páginas da Web, tamanho em Gigabytes
- Proporção de sítios Web utilizando IPv6
- Distribuição do uso de idiomas na Web brasileira
- Proporção de páginas da Web aderentes aos padrões HTML do W3C
- Proporção de páginas da Web aderentes aos padrões de acessibilidade Ases

Objetivos específicos

- Proporção de tipos de objetos usados nas páginas da Web
- Proporção de tipos de tecnologias usadas nas páginas da Web
- Idade das páginas
- Geolocalização dos servidores IP
- Sincronização de tempo dos servidores web

Produtos



- Resultado do censo da Web GOV.BR
- Base de dados para divulgação e fins de pesquisa
- Publicação de ferramentas

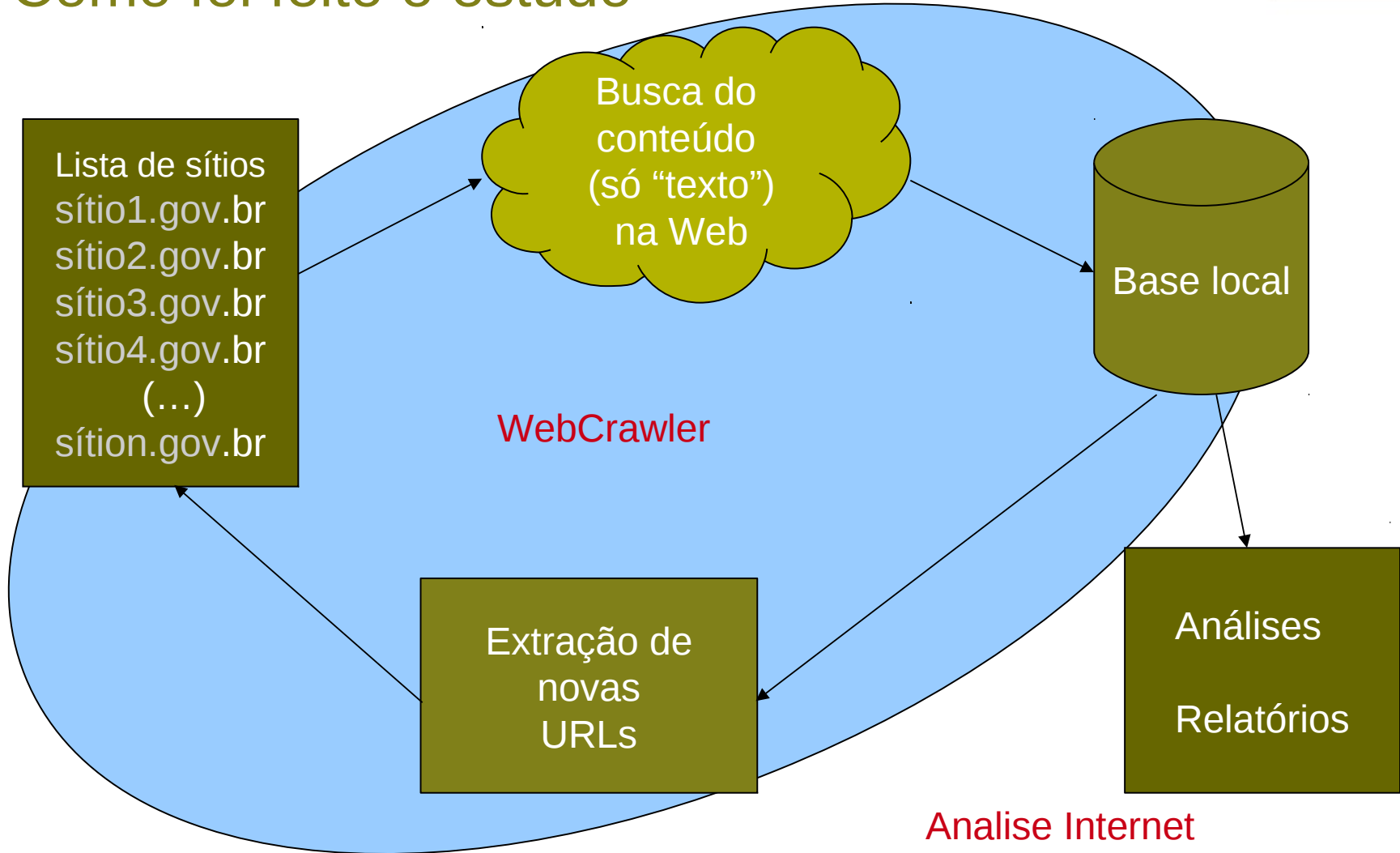
Desafios técnicos para o estudo da Web

A Web Brasileira

- Não existe uma definição para “Web brasileira”- a WWW (World Wide Web) é uma rede, como diz seu nome, de alcance mundial!
- Como restringir o estudo?
 - Localização geográfica
 - Idioma
 - Domínios



Como foi feito o estudo



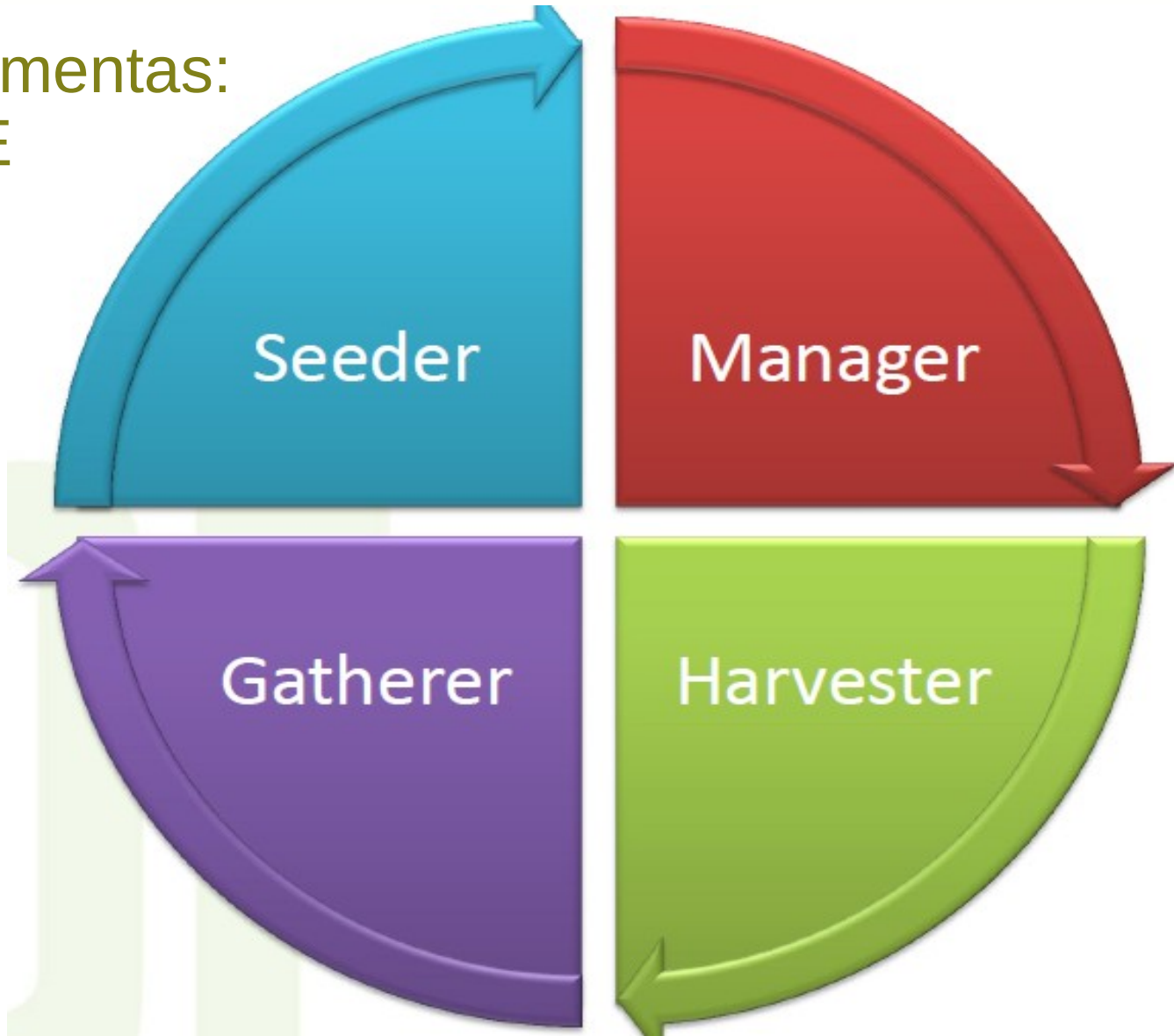
Ferramentas

- **WIRE = webcrawler**
 - Software livre
 - Desenvolvido no Chile para estudos acadêmicos sobre a Web
 - Fizemos correções e adaptações no NIC.br

- **Análise Internet**
 - Desenvolvido no NIC.br
 - Localização Geográfica, IPv6, Sincronização, Aderência a padrões, Tipos de documentos, Tipos de Servidores, Tempo de Resposta.



Ferramentas: WIRE



Além disso...

- Calcula a **quantidade e tamanho** das páginas
- Informa e **classifica** as URLs
- Informa a **idade** das páginas
- Diferencia páginas **estáticas e dinâmicas**
- Calcula diferentes índices de páginas como **pagerank** e **siterank**
- Identifica o **idioma** das páginas

Ferramentas: AnáliseInternet

- Realiza 5 tipos de teste:
 - Carregamento dos dados do WIRE
 - Validação de páginas
 - Testes sobre sítios e servidores
 - Testes sobre links encontrados
 - Download de Arquivos adicionais



Ferramentas: AnáliseInternet

- **Validação de páginas:**
 - Validador de HTML do W3C
 - Validados de acessibilidade ASES
- Arquitetura Distribuída
 - Aumenta a velocidade de processamento



<http://xenlights.com/images/SoftwareValidation.jpg>

ASES

Aderência à padrões HTML

- Universalidade do acesso:
 - Acesso sem barreiras
 - Compatibilidade
 - Acessibilidade
 - Ganho de desempenho
 - Economia de banda
 - Código mais simples e fácil de atualizar
 - Melhor visibilidade em ferramentas de busca
 - Evita instabilidade e versões de páginas



<http://cssti.files.wordpress.com/2009/06/26032008www.jpg?w=363&h=362>

Aderência a padrões de acessibilidade

- Garante acesso universal aos sítios Web:
 - O modelo de acessibilidade considerado foi o e-MAG
 - O e-MAG tem como referência as diretrizes de acessibilidade do W3C publicadas no WCAG
 - A ferramenta utilizada para a validação foi o ASES, criado pela SLTI do Ministério do Planejamento



Ferramentas: AnáliseInternet

- Testes sobre Sítios e Servidores: **Reposta**
 - Realiza uma requisição HEAD
 - Obtêm:
 - Tempo de resposta
 - Tipo de servidor
 - Diferença de tempo
 - Ipv4



http://www.superdownloads.com.br/imagens/materias/Rodrigo%20Uma%20mat%20de%20conex%20de-
0df1aueur

Ferramentas: AnáliseInternet

- Testes sobre Sítios e Servidores: **Domínio**
 - Decide sobre qual domínio um site se encontra
 - Considera **ccTLD**, **TLD** e **UF** em caso de sites .gov.br

<http://www.prefeitura.sp.gov.br/>

<http://www.nic.br/>

<http://www.google.com.br/>



IPv6

- Todo computador na Internet necessita de um endereço: o número IP
- A numeração usada atualmente (IP versão 4) está se esgotando. Previsões indicam que o estoque global acabará no 1o. Semestre de 2011.
- Uma nova versão de IP está sendo implantada na Internet: o IPv6
- O IPv6 é necessário para:
 - Garantir o crescimento e desenvolvimento da rede
 - Inclusão digital
 - Internet das coisas

Ferramentas: AnáliseInternet

- Testes sobre Sítios e Servidores: IPv6
 - Não é suficiente verificar se o domínio possui ipv6



ipv6.google.com.br

www.v6.facebook.com

- Utiliza variações do nome do site: [www6](#), [www.ipv6](#), [ipv6](#)
- Realiza ping6 e requisição GET ao endereço
- Verifica se o **NameServer** possui suporte a IPv6

Sincronização com a Hora Legal Brasileira

A Sincronização com a Hora Legal Brasileira (que equivale ao padrão mundial UTC) é recomendada pelo CGI.br:

<http://www.cgi.br/regulamentacao/resolucao2008-009.htm>

Todo servidor deve estar sincronizado e, na medida do possível, também computadores pessoais. Isso é importante para:

- segurança
- funcionamento correto das aplicações

Deve-se utilizar o NTP: <http://ntp.br>. É uma configuração simples de ser realizada, porém pouco conhecida. Os servidores de tempo são oferecidos pelo NIC.br, em conjunto com o Observatório Nacional.

A medida foi realizada obtendo-se a hora dos servidores, via HTTP, e comparando-a com a hora correta.

Testes de Sincronização de Tempo

Resposta do servidor Web

Depende da qualidade da rede e processamento do servidor

Precisão de ~ **seg**

Protocolo NTP

Mais preciso: ~ **μ seg**

Mais difícil de ser obtido



Localização Geográfica dos Servidores da Web governamental

O indicador mostra a proporção de servidores localizados no Brasil e no exterior. Usa dados de uma base especializada (GeoIP MaxMind) que tem cerca de 95% de exatidão.

Servidores fora do Brasil podem ser mais baratos, contudo implicam em maior lentidão no acesso, e no uso de canais de comunicação de internacionais, com alto custo, levando a um aumento nos custos de acesso no Brasil.

É importante que os servidores que hospedam sítios destinados aos internautas do Brasil estejam hospedados no Brasil.

Ferramentas: AnáliseInternet

- Testes sobre Sítios e Servidores: **Geolocalização**
 - Geo Localização de servidores por IP
 - Integra API do **GeoIP®**



Dimensões da web governamental

Participação das regiões na composição da web governamental - .gov.br

Número de sítios Web - 11.856

Número total de páginas HTML da Web – 6.331.256

Número médio de páginas HTML por sítio – 534,01

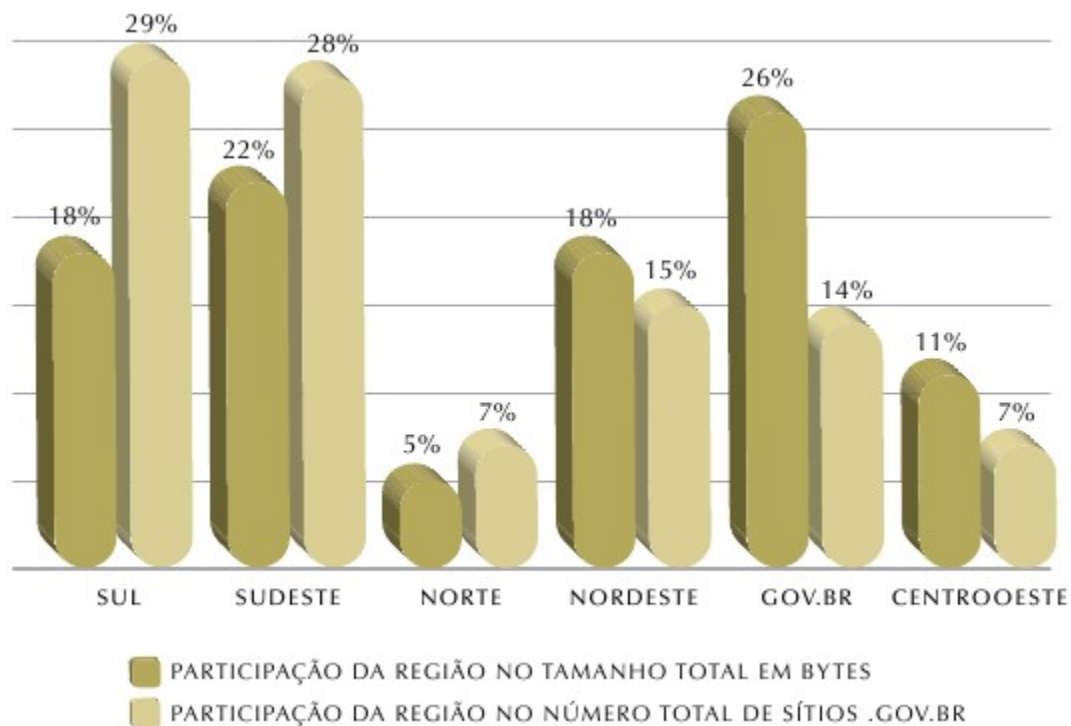
Aproximadamente 70% dos sítios tem menos de 100 páginas

Cerca de 10% dos sítios tem mais de 1 mil páginas

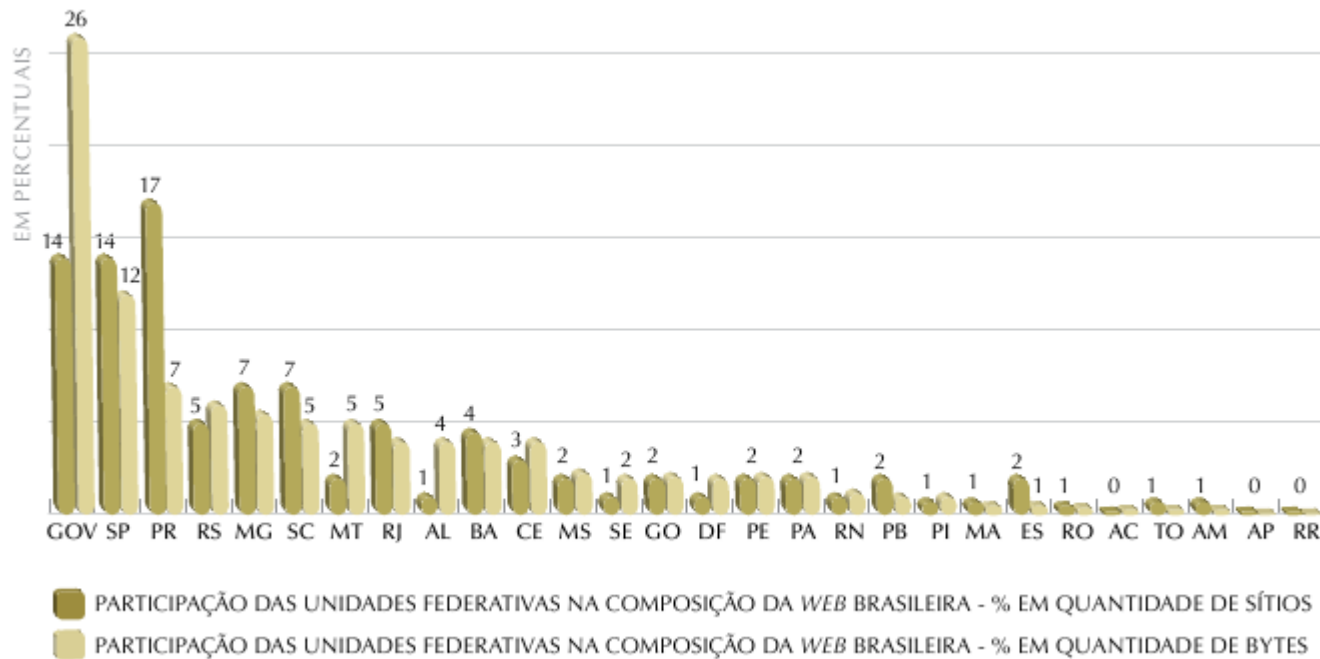
Participação das regiões na composição da web governamental

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%

Participação das regiões na composição da web governamental



Participação das regiões na composição da web governamental – por UF

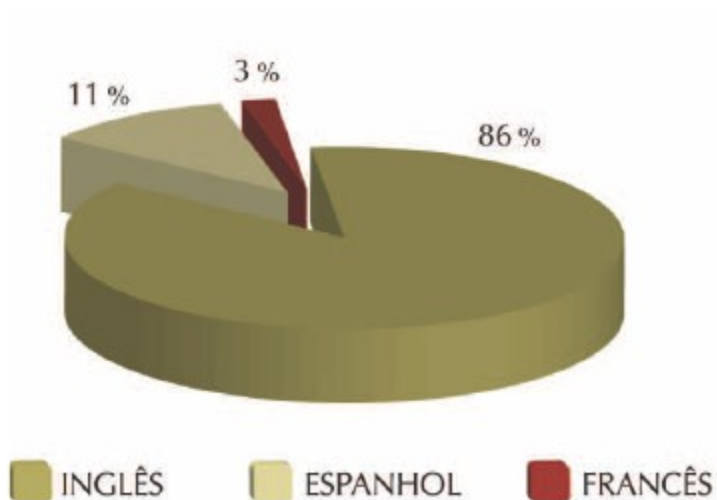


Outros idiomas na web governamental

Identificação da presença de outros idiomas nas páginas

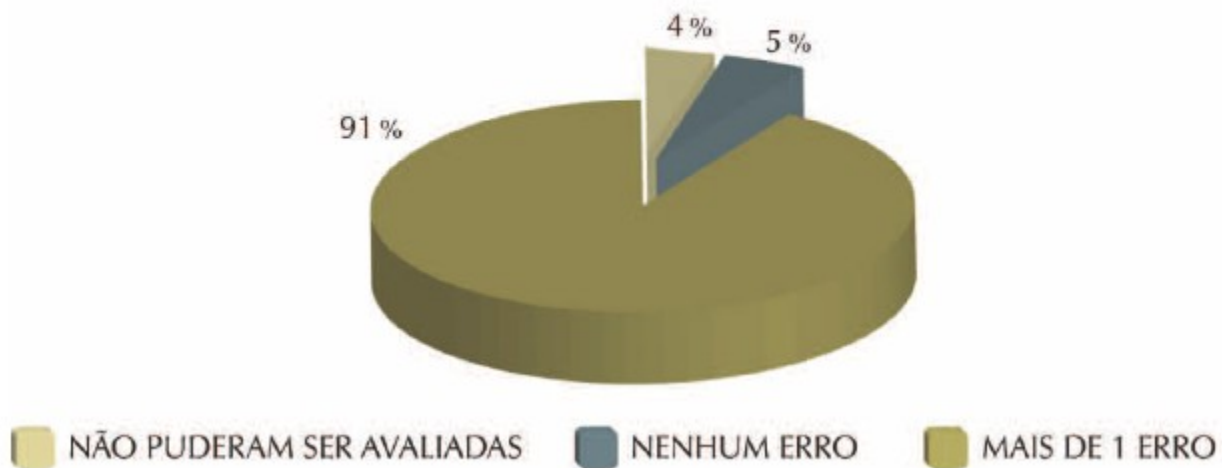
- verificados 4 idiomas por meio de palavras-chaves – Português, Inglês, Espanhol e Francês
- Verificadas as páginas HTML válidas
- 97 % estão em Português

Outros idiomas na web governamental

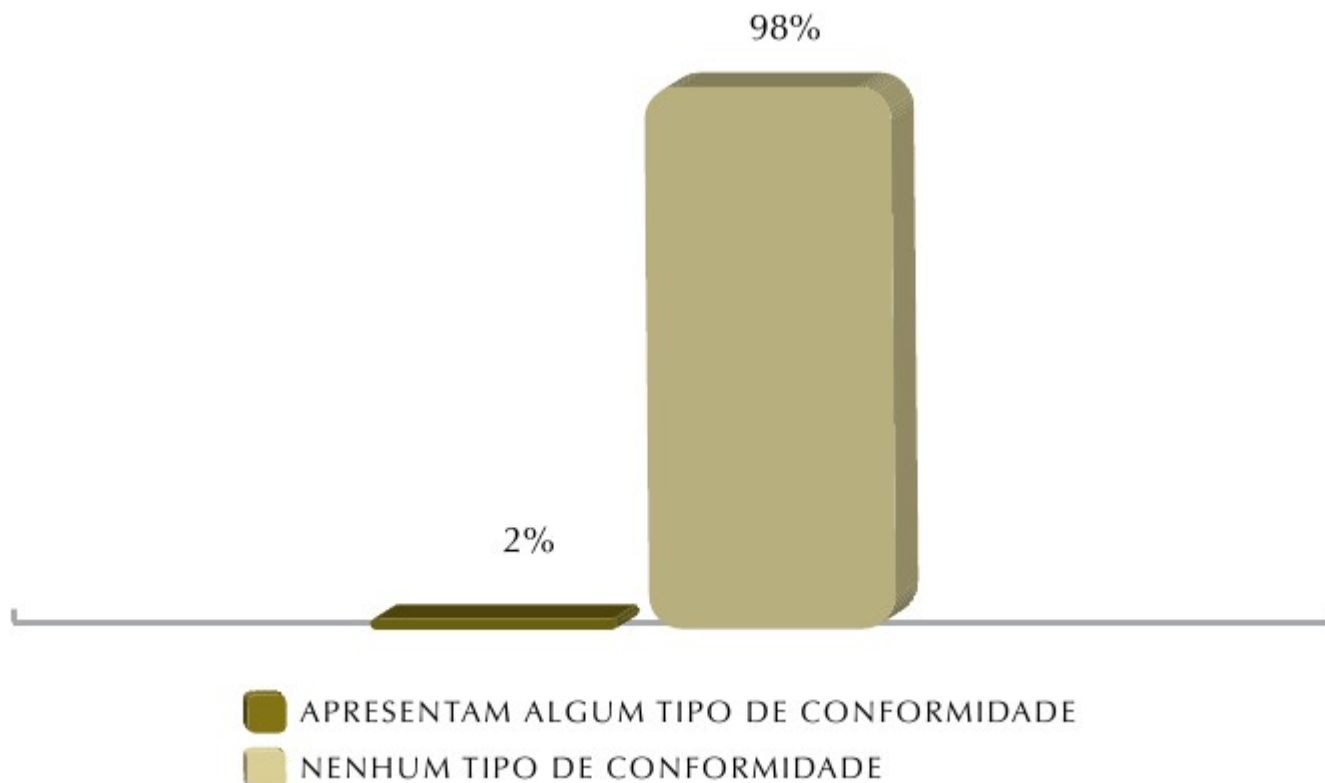


Distribuição dos idiomas diferentes do Português nas páginas web .gov.br

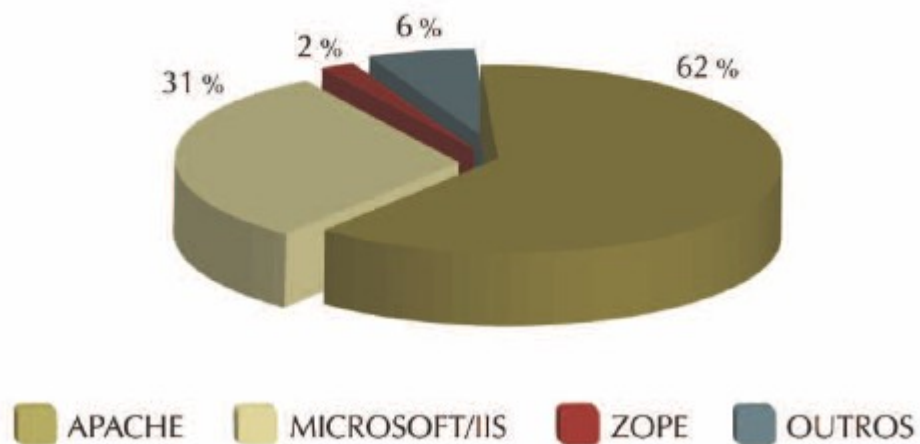
Aderência aos padrões HTML do W3C na web governamental



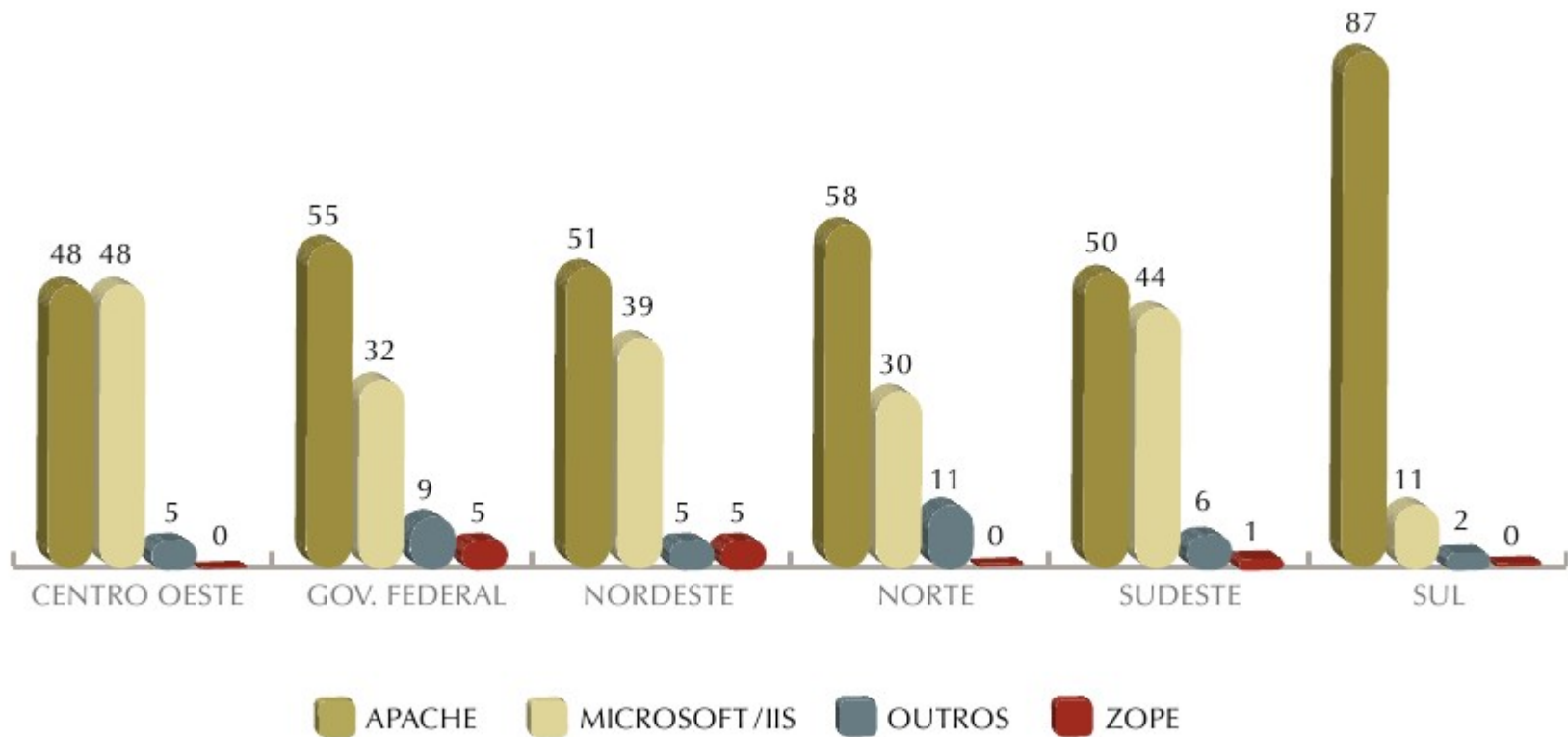
Aderência aos padrões de acessibilidade ASES na web governamental



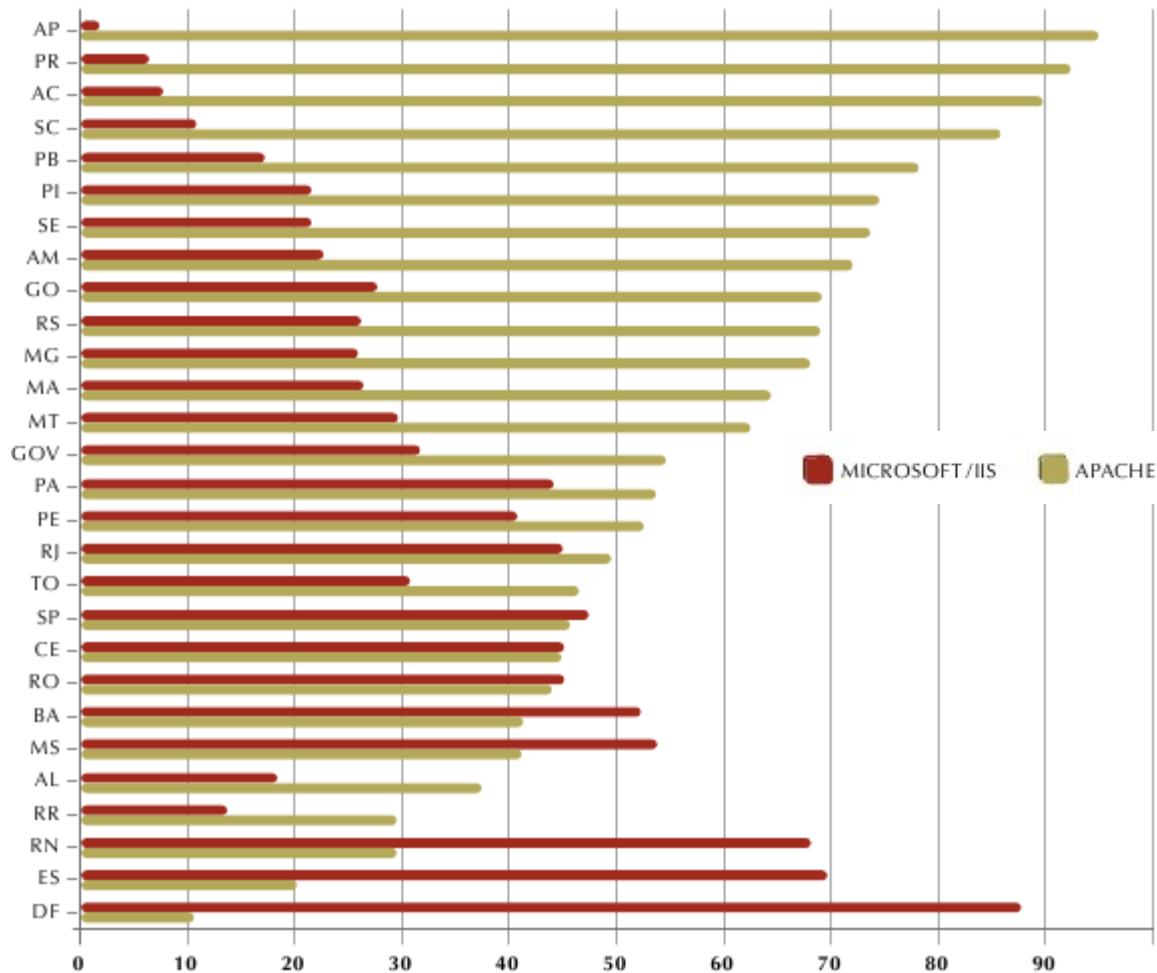
Tecnologias utilizadas para servir arquivos na web governamental



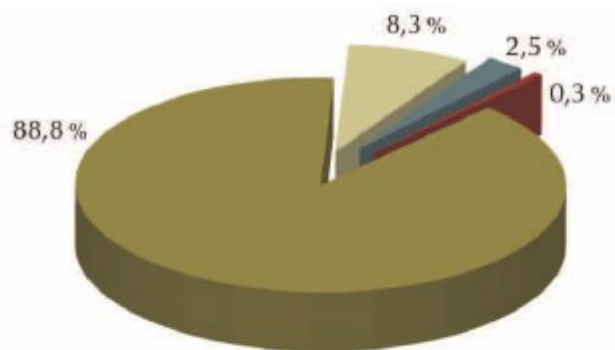
Tecnologias utilizadas para servir arquivos na web governamental – por região



Tecnologias utilizadas para servir arquivos na Web governamental – por UF

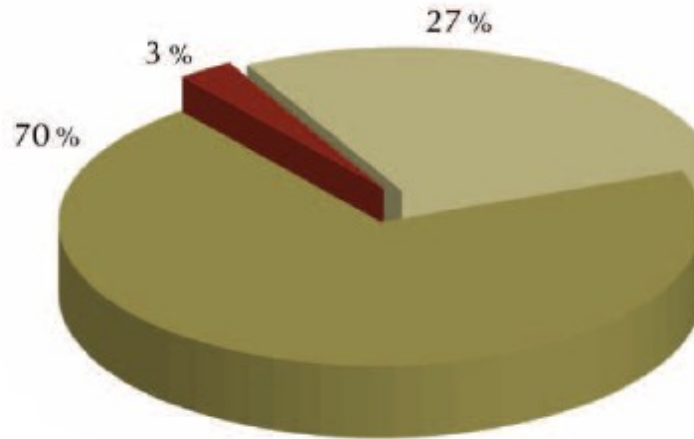


Objetos mais usados nas páginas da Web governamental



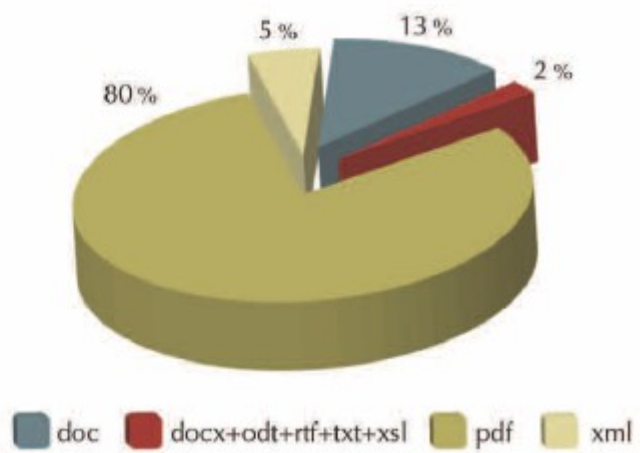
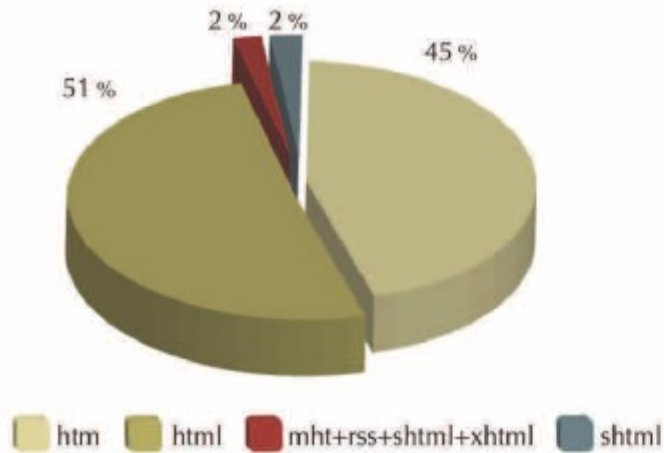
- HIPERTEXTO
- DOCUMENTO
- PLANILHA + APRESENTAÇÃO + BANCO DE DADOS + ÁUDIO + VÍDEO
- GRÁFICOS

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

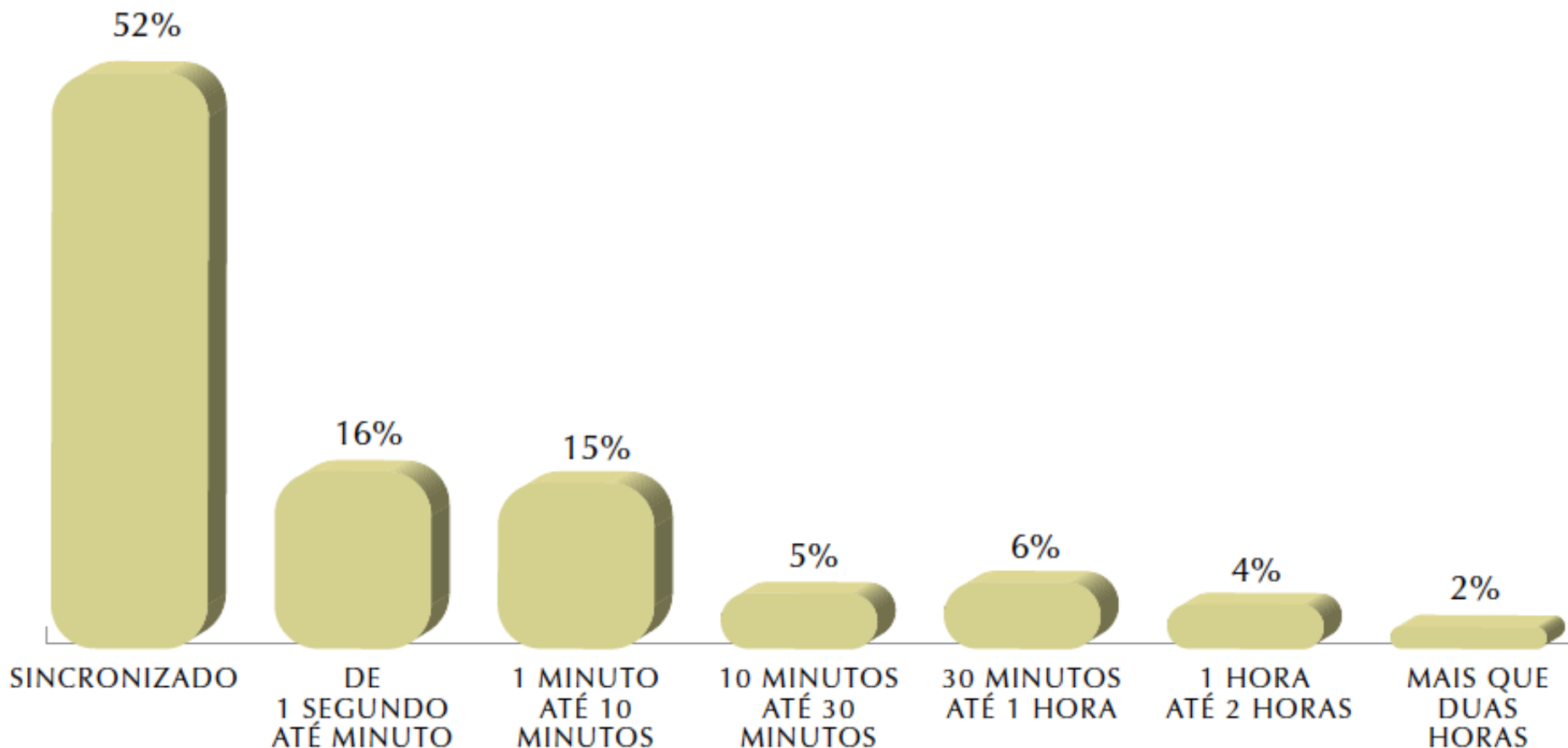


- asp/aspx (ACTIVE SERVER PAGES - PROPRIETÁRIA)
- php/php3 (HYPERTEXT PREPROCESSOR - CÓDIGO ABERTO)
- DEMAIS TECNOLOGIAS

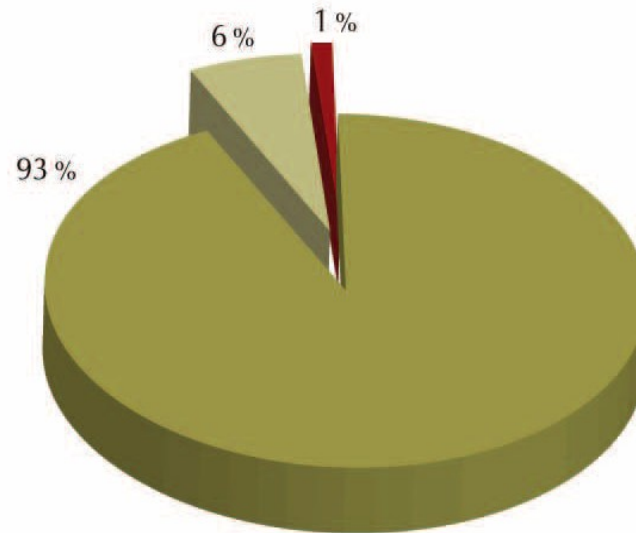
Tecnologias mais utilizadas para disponibilização de dados e conteúdo na web governamental



Sincronização com a Hora Legal Brasileira



Localização Geográfica dos servidores da Web governamental



- IP LOCALIZADO NO BRASIL
- IP LOCALIZADO NO EXTERIOR
- LOCAL NÃO IDENTIFICADO

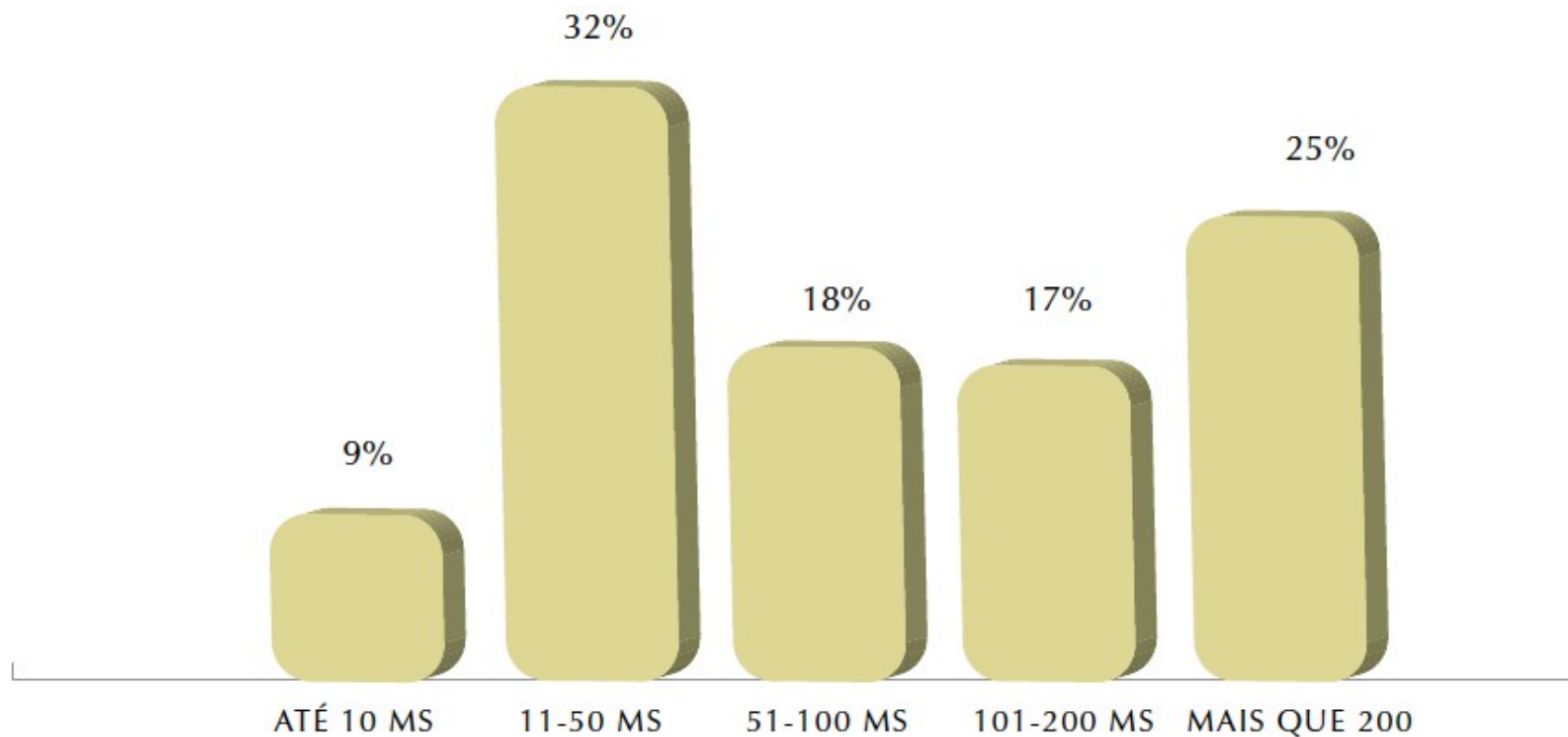
Tempo de resposta dos Servidores da Web governamental

Medida de desempenho do sítio, do ponto de vista de um usuário localizado em São Paulo

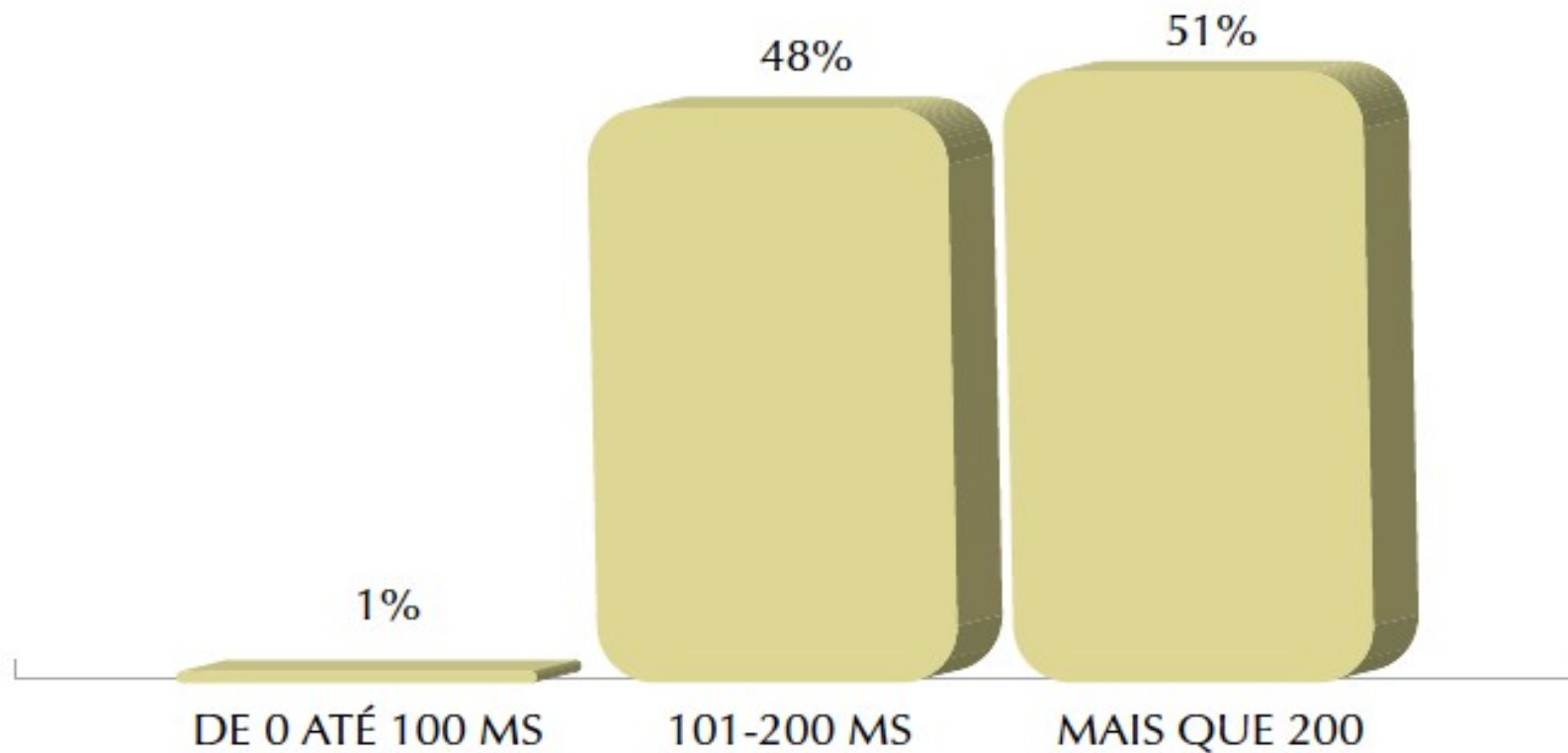
Mede o tempo de trânsito da informação, mais o tempo de processamento do servidores, para uma requisição de dados muito simples

Relação direta com a infraestrutura da Internet

Tempo de resposta – servidores no Brasil



Tempo de resposta – servidores no exterior



IPv6 na Web governamental

- No documento de referência da e-PING, temos: *“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”*

(<http://www.governoeletronico.gov.br/anexos/e-ping-versao-3.0>)

No estudo realizado, nenhum sítio “.gov.br” estava disponível via IPv6.

2011

- **Estudo de Novos Domínios**
 - **com.br**
 - **Profissionais liberais**
 - **Domínios educacionais**
- **Novos Projetos**
 - **Analizador de sites populares internacionais e brasileiros**
 - **Ferramenta de monitoramento de sites pessoais**

Obrigado !

- **Contatos:**

- web@ceptro.br
- moreiras@nic.br
- heitor@nic.br
- phadek@nic.br

- **Links úteis**

- <http://ceptro.br/CEPTRO/MenuCEPTROSPCensoWeb>
- <http://w3c.br/>
- <http://www.cwr.cl/projects/WIRE/>

- **Perguntas???**