

# Utilizando ferramentas de software livre para estudar a Web Brasileira

1. O projeto
2. As Etapas
3. Ferramentas Utilizadas: Wire, Analise internet
4. Desafios

# O Projeto



## Proposta

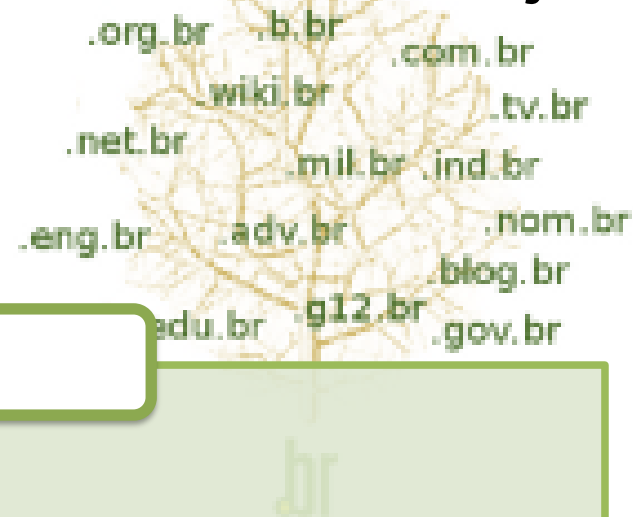
- Estudar características dos sites e páginas brasileiras.

# O Projeto

## Objetivos

- Criar e divulgar **indicadores** de todos os sítios brasileiros
- Medida de **tamanho e crescimento** da Web
- Classificação dos tipos de **tecnologia** mais utilizados
- Medida de **aderência** à padrões
- Localização de servidores de sites
- Medida de utilização de **IPv6**

# O Projeto



## Como

- Coleta
- Armazenamento
- Análise da Web brasileira

## Como definir a WEB brasileira?

- ✓ Sítios em português
- ✓ Sítios feitos para brasileiros
- ✓ Sítios feitos por brasileiros
- ✓ Sítios hospedados no Brasil
- ✓ Sítios registrados sob o domínio .br

### ✓ Problema

- Exclui grandes portais como [globo.com](http://globo.com) ou [msn.com](http://msn.com)

## Domínio .br + redirecionamento a partir de .br

# Fatores Limitantes

## ✓ A Web profunda

## ✓ Servidores

- Indisponibilidade
- Inexistência de robots.txt

## ✓ Conteúdo

- Conteúdo igual para URL's diferentes
- Caracterização da página
  - Encodificação
  - Idioma



## Etapas do Projeto

### Estudo de subconjuntos do .br

- Censo do **.gov.br**
- Estudo amostral do **.br**
- Estudo de outros domínios menores como **.org.br**

### Estimar recursos

- Tempo
- Espaço
- Banda

## Ferramentas Utilizadas

### ✓ WIRE

- **Web Crawler** open source desenvolvido pela desenvolvido inicialmente pelo “**Center for Web Research**”
- Adaptado pelo **NIC.br**
- **Linguagem C++**



### ✓ AnáliseInternet

- Realiza testes não incluídos no WIRE
- Consolida resultados em um banco de dados
- **Linguagem Java**
- Desenvolvido pelo **NIC.br**



## Ferramentas Utilizadas: WIRE

### Vantagens

- Fácil **recuperação** em caso de parada
- Boa **documentação**
- Funcionamento em **ciclos**
- Facilidade de **configuração**
- Boa **estruturação dos dados**

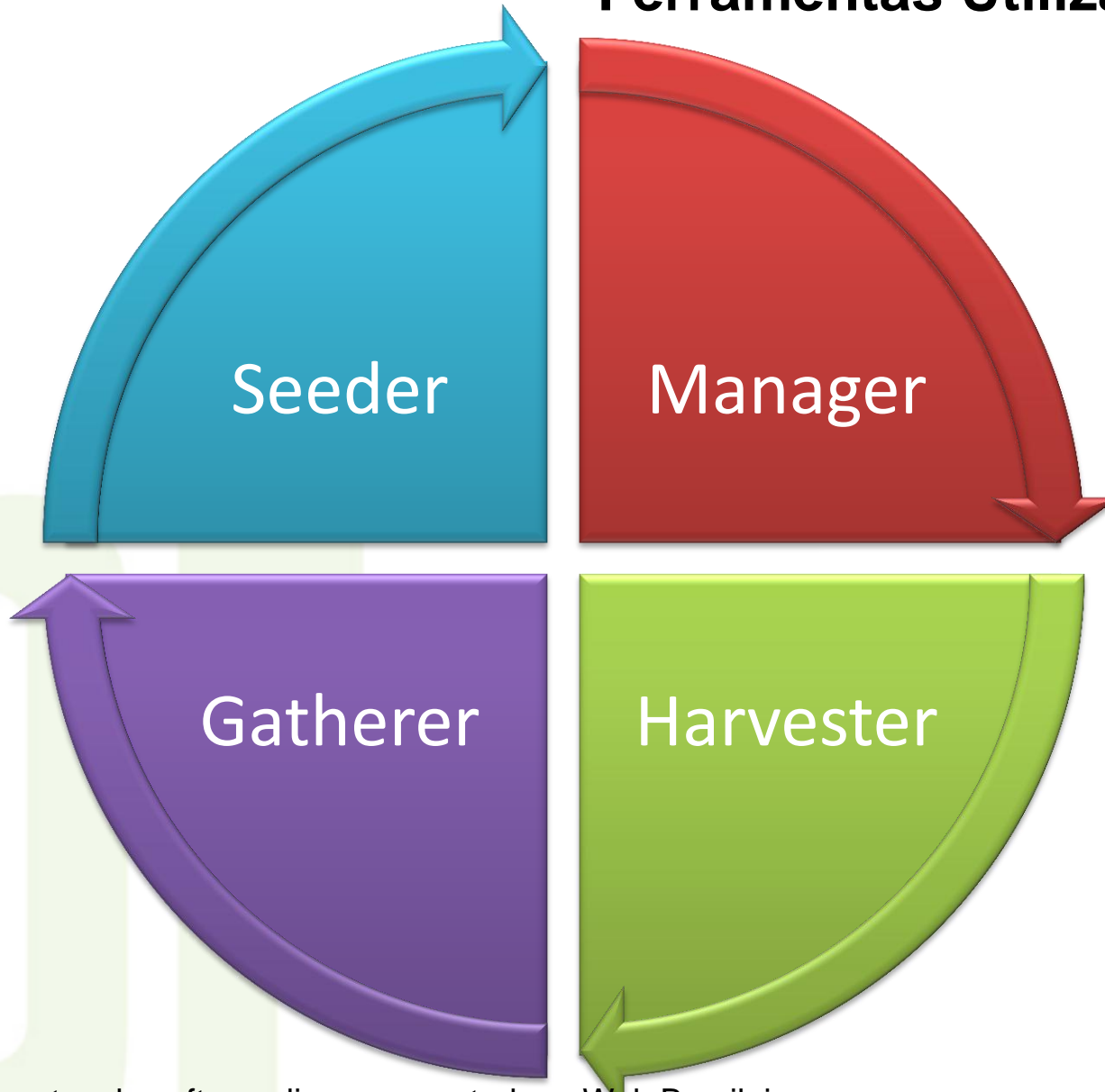
### Desvantagens

- Arquitetura **unithread**
- Baixa **velocidade de download**
- Baixa **escalabilidade**
- Difícil de ser **testado**

### Outros projetos considerados

- **Heritrix**
- **Apache Nutch**

# Ferramentas Utilizadas: WIRE



## Ferramentas Utilizadas: 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**
- Diferentes rankeamentos de páginas **pagerank** e **siterank**
- Identifica o **idioma** das páginas

## Ferramentas Utilizadas: WIRE

**Algumas modificações foram necessárias...**

## Ferramentas Utilizadas: WIRE

### ✓ Armazenamento de páginas

- Garantir a **integridade do HTML**
- Aumentar a **escalabilidade** do sistema
- Salva as páginas uma **estrutura hierárquica** de páginas
- Utiliza o **sistema de arquivos**

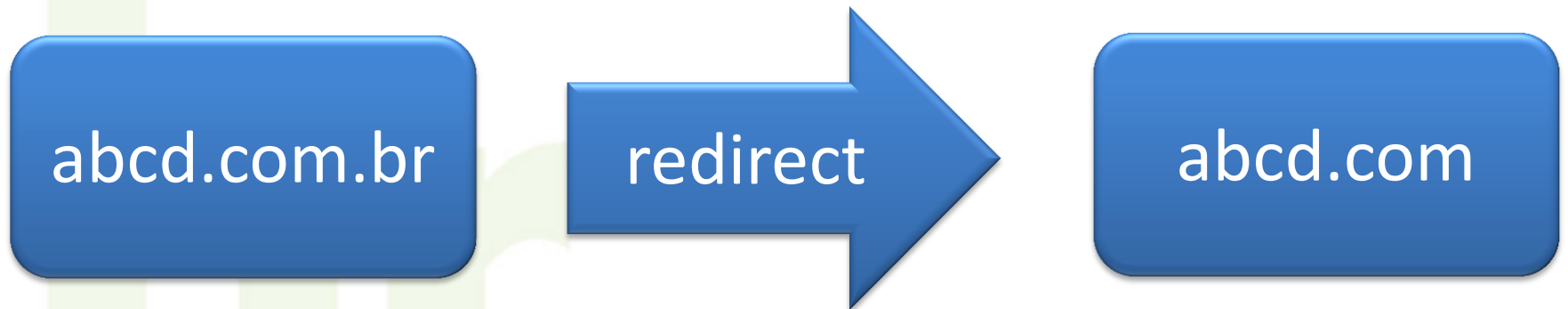
# Ferramentas Utilizadas: WIRE

## ✓ Tratamento de redirect



## Ferramentas Utilizadas: WIRE

### ✓ Tratamento de redirect



**Deve-se salvar todas as paginas que seriam vistas por uma URL com domínio .br**

## Ferramentas Utilizadas: WIRE

- ✓ Lista de domínios
  - Finalidade:
    - Quantidade de domínios percorridos
    - Coletas sem lateralidade
  - Antes:
    - Domínios configurados diretamente **XML de configuração**
  - Depois:
    - Arquivo com a lista de domínios

## Ferramentas Utilizadas: WIRE

- ✓ Aleatorização da ordem de download das páginas
  - Evita que as coletas sigam uma ordem fixa de download
  - Entre cada ciclo embaralha a lista de URLs é embaralhada



## Ferramentas Utilizadas: WIRE

- ✓ Correção da ferramenta de análise de idomas
  - Finalidade:
    - Aumentar a acertividade
  - Ações:
    - Melhora do **parsening**
      - estruturação de tags
      - delimitação de comentários
      - exclusão de scripts nas páginas
    - Correções de Bugs



## Ferramentas Utilizadas: WIRE

✓ Normalização de URLs segundo as normas da **RFC3986**

- **Finalidade:**

- Redução de páginas duplicadas
- Facilitar a estrutura de arquivos

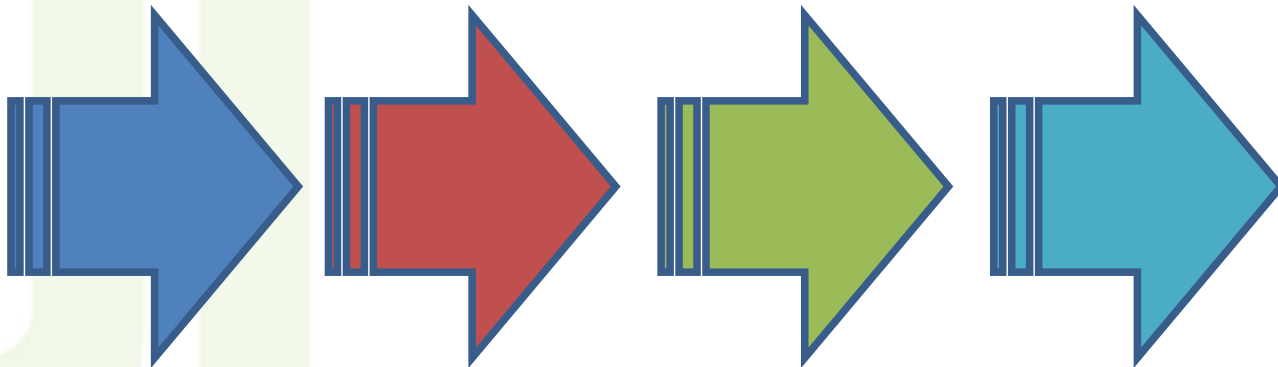
- **Principais casos:**

- /./ /./, /, espaços no fim
- Percent encoding, para caracteres especiais: %20

## Ferramentas Utilizadas: WIRE

### ✓ Chunked encoding

- Encodificação para transferência de dados com protocolo **HTTP 1.1**
- Sem suporte o conteúdo das páginas fica prejudicado
- Aumenta a integridade das páginas



# Ferramentas Utilizadas: AnáliseInternet

✓ Quatro classes de teste:

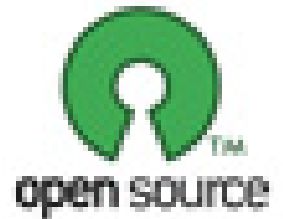
1. **Carregamento** dos dados do WIRE
2. Testes sobre as **páginas**
3. Testes sobre os **servidores / sites**
4. **Download** de arquivos **XML**

## Ferramentas Utilizadas: AnáliseInternet

- ✓ Desenvolvido sobre plataforma **Java**
  - Hibernate
  - Dao
  - Singleton
  
- ✓ Arquitetura **Multitarefa**
  - Classe que gerencia o teste
  - Classe executa a tarefa do teste
  - Número de threads configurável
  
- ✓ Fácil criação de novos testes

# Ferramentas Utilizadas: Análise Internet

- ✓ Testes sobre páginas
  - ✓ Validar **HTML** do W3C
  - ✓ Validador de **Acessibilidade** ASES
- ✓ **Arquitetura Distribuída**
  - Realiza os testes em ordem aleatória
  - Aumenta a velocidade de processamento

The logo for ASES (Automated Accessibility Evaluation System) features the letters 'A', 'S', and 'E' in blue and 'S' in green, all in a bold, sans-serif font with a slight shadow effect.The logo for W3C (World Wide Web Consortium) consists of the letters 'W3C' in a blue, serif font, with a registered trademark symbol (®) to the upper right.The logo for Open Source features a green circular icon with a white keyhole shape inside, followed by the text 'open source' in a lowercase, sans-serif font.

# Ferramentas Utilizadas: AnáliseInternet

✓ Testes sobre sítios web

## ■ Resposta

- Realiza uma requisição **HEAD** ao servidor
- Obtêm:
  - RTT
  - Tipo de servidor
  - Diferença de tempo
  - Ipv4

# Ferramentas Utilizadas: Análise Internet

## ✓ Testes sobre sítios web

### ■ 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/>

## Ferramentas Utilizadas: Análise Internet

✓ Testes sobre sítios web

### ■ 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

# Ferramentas Utilizadas: Análise Internet

✓ Testes sobre sítios web

The logo for ntp.br, featuring the text 'ntp.br' in a stylized, lowercase font. The 'n' and 't' are in a light green color, while the 'p' and 'br' are in a darker green. The logo is set against a light gray rectangular background.

## ■ NTP

■ Utiliza comando **ntpdate** para obter:

- NTP **Delay**
- NTP **Stratum**
- NTP **Offset**



# Ferramentas Utilizadas: Análise Internet

✓ Testes sobre sítios web

■ Geo Localização de servidores por IP

• Integra API do **GeoIP®**



## Acelerar a execução das coletas

- Tornar o WIRE mais eficiente em termos de banda e utilização de CPU

## Tratamento de armadilhas

- Páginas dinâmicas
- Calendários
- Páginas com conteúdo duplicado

## Análise completa dos links

- Verificação de tamanho dos links
- Indicação de interligação das páginas
- Extração de extensões e domínios

Utilização de ferramentas de Data Warehouse e Data Mining para

Delimitação dos recursos necessários para a realização de uma coleta completa

- ✓ Nic.br: <http://www.nic.br/>
- ✓ Ceptro: <http://www.ceptro.br/>
  
- ✓ WIRE: <http://www.cwr.cl/projects/WIRE/>
- ✓ Heritrix: <http://crawler.archive.org/>
- ✓ Apache Nutch: <http://nutch.apache.org/>
  
- ✓ ASES: <http://www.governoeletronico.gov.br/>
- ✓ W3C: <http://W3C.org/>

# Obrigado!

Perguntas, Sugestões, Comentários...

Contato: [heitor@nic.br](mailto:heitor@nic.br)