UNIDADE 3

A *WEB* COMO REPOSITÓRIO UNIVERSAL DE INFORMAÇÕES

3.1 OBJETIVO GERAL

Apresentar os conceitos de *web* visível e profunda e a forma como se dá o processo de indexação da web pelos mecanismos de busca, mostrando as dificuldades enfrentadas, primeiro, devido à existência de conteúdos não estruturados em bases de dados e, segundo, pelas limitações do atual esquema de endereçamento de páginas.

3.2 OBJETIVOS ESPECÍFICOS

Esperamos que, ao final desta Unidade, você seja capaz de:

a) diferenciar *web* visível de *web* profunda ou invisível;

b) identificar mecanismos, padrões e tecnologias que permitem que uma página *web* possa ser indexada corretamente pelos indexadores dos mecanismos de busca;

c) reconhecer problemas ligados à identificação permanente de recursos na *web* e as ferramentas existentes para tentar evitá-los.

3.3 ALÉM DA PONTA DO *ICEBERG*

A experiência da maioria das pessoas que usam a *web* no seu dia a dia é a de que os mecanismos de busca são capazes de responder a qualquer questão. No entanto, o que poucas pessoas sabem é que a *web* se assemelha a um *iceberg*: uma pequena porção fica visível acima do mar, no entanto, sua maior parte permanece invisível, submersa (Figura 43). A cobertura dos mecanismos de busca limita-se a cerca de 20% dos conteúdos disponíveis na *web*. A porção não coberta pelos mecanismos de busca é chamada de *web* invisível ou *web* profunda.

Figura 43 – A *web* é como um *iceberg*, o que os mecanismos de busca nos fornecem como resposta é apenas uma pequena parte do todo

Fonte: *Flickr*.[12]

Nesta Unidade, vamos aprender mais sobre como os mecanismos de busca são capazes de recuperar tanta coisa publicada na *web*, mas também como muito do que existe nela não é encontrável por meio desses mecanismos. Vamos aprender também padrões, normas e tecnologias que tornam nossos conteúdos – por exemplo, a página da biblioteca – mais facilmente encontráveis, tanto hoje como daqui a 200 anos.

3.4 *WEB* VISÍVEL X *WEB* PROFUNDA

Com o surgimento da *web*, qualquer indivíduo passou a poder publicar conteúdos sob a forma de páginas hipertextuais, ou seja, interligadas por *links*, fazendo com que encontrar a informação desejada, à medida que as páginas *web* se multiplicavam, se tornasse uma questão crítica.

Hipertexto

Designação de "um processo de escrita/leitura não linear e não hierarquizada e que permite o acesso ilimitado a outros textos de forma instantânea. Possibilita ainda que se realize uma trama, ou rede, de acessos, sem seguir, necessariamente, sequências ou regras (FACHINETTO, 2005)".

[12] Autor: *National Ocean Service Image Gallery.* Disponível em: <https://www.flickr.com/photos/usoceangov/8290528771>. Acesso em: 30 jul. 2021.

O diretório *Yahoo* e a *Internet Public Library* (IPL) (<http://www.ipl.org/>) foram as primeiras tentativas de solucionar esse problema, ainda nos anos 1990. Esses SRIs coletavam, descreviam e organizavam sistematicamente páginas *web*. O trabalho era feito por profissionais de informação, o que lhe garantia uma qualidade muito grande. No entanto, a *web* crescia mais rápido que a capacidade desses SRIs de indexarem as novas páginas.

Além disso, um problema cada vez mais sério, que tornava frágeis serviços como o diretório *Yahoo*, a IPL e outros semelhantes, era a necessidade de *constante atualização dos endereços das páginas* nos índices desses serviços.

Assim, à medida que o tempo passava, mudanças de endereço de páginas, devidas à administração das pastas dos servidores que hospedavam as páginas *web* ou às mudanças nas próprias instituições que mantinham as páginas (mudanças de nomes de domínio), se tornavam constantes. Essas mudanças passaram a ser uma dor de cabeça para qualquer serviço de informação que trabalhava com esses endereços, como o diretório *Yahoo* ou a IPL, já mencionados. Veremos, mais adiante, outras questões relativas à alteração de endereço das páginas *web* e suas consequências para a economia da *web*.

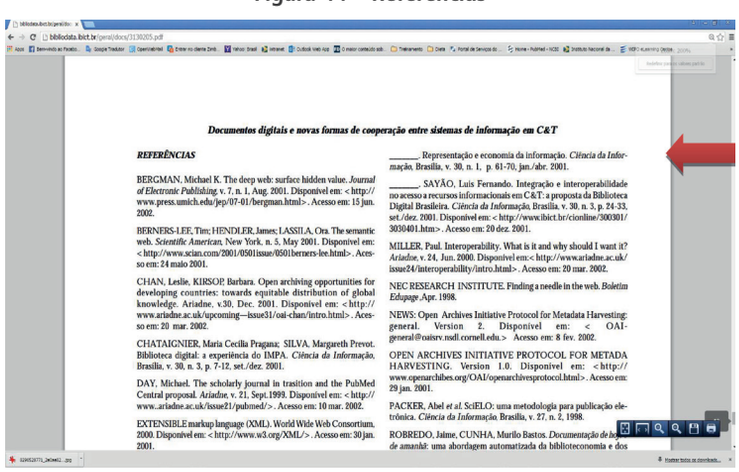
3.4.1 **Atividade**

Para realizar esta atividade, você precisará de um dispositivo com acesso à internet. A figura a seguir apresenta as referências bibliográficas de um artigo, a saber:

MARCONDES, C. H.; SAYÃO, L. F. Documentos digitais e novas formas de cooperação entre sistemas de informação em C&T. **Ci. Inf.**, Brasília, v. 31, n. 3, p. 42-54, 2002.

Veja, na Figura 44, que várias das referências listadas são de artigos digitais:

Figura 44 – Referências



Documentos digitais e novas formas de cooperação entre sistemas de informação em C&T

REFERÊNCIAS

BERGMAN, Michael K. The deep web: surface hidden value. *Journal of Electronic Publishing*, v. 7, n. 1, Aug. 2001. Disponível em: < http://www.press.umich.edu/jep/07-01/bergman.html>. Acesso em: 15 jun. 2002.

BERNERS-LEE, Tim; HENDLER, James; LASSILA, Ora. The semantic web. *Scientific American*, New York, n. 5, May 2001. Disponível em: < http://www.scian.com/2001/0501issue/0501berners-lee.html>. Acesso em: 24 maio 2001.

CHAN, Leslie, KIRSOP, Barbara. Open archiving opportunities for developing countries: towards equitable distribution of global knowledge. *Ariadne*, v.30, Dec. 2001. Disponível em: < http://www.ariadne.ac.uk/upcoming—issue31/oai-chan/intro.html>. Acesso em: 20 mar. 2002.

CHATAIGNIER, Maria Cecília Pragana; SILVA, Margareth Prevot. Biblioteca digital: a experiência do IMPA. *Ciência da Informação*, Brasília, v. 30, n. 3, p. 7-12, set./dez. 2001.

DAY, Michael. The scholarly journal in trasition and the PubMed Central proposal. *Ariadne*, v. 21, Sept.1999. Disponível em: < http://www..ariadne.ac.uk/issue21/pubmed/ > . Acesso em: 10 mar. 2002.

EXTENSIBLE markup language (XML). World Wide Web Consortium, 2000. Disponível em: < http://www.w3.org/XML/ > . Acesso em: 30 jan. 2001.

_____. Representação e economia da informação. *Ciência da Informação*, Brasília, v. 30, n. 1, p. 61-70, jan./abr. 2001.

_____. SAYÃO, Luis Fernando. Integração e interoperabilidade no acesso a recursos informacionais em C&T: a proposta da Biblioteca Digital Brasileira. *Ciência da Informação*, Brasília, v. 30, n. 3, p. 24-33, set./dez. 2001. Disponível em: < http://www.cibict.br/cionline/300301/3030401.htm> . Acesso em: 20 dez. 2001.

MILLER, Paul. Interoperability. What is it and why should I want it? *Ariadne*, v. 24, Jun. 2000. Disponível em:< http://www.ariadne.ac.uk/issue24/interoperability/intro.html>. Acesso em: 20 mar. 2002.

NEC RESEARCH INSTITUTE. Finding a needle in the web. *Boletim Edupage* ,Apr. 1998.

NEWS: Open Archives Initiative Protocol for Metadata Harvesting: general. Version 2. Disponível em: < OAI-general@oaisrv.nsdl.cornell.edu.> Acesso em: 8 fev. 2002.

OPEN ARCHIVES INITIATIVE PROTOCOL FOR METADA HARVESTING. Version 1.0. Disponível em: < http://www.openarchibes.org/OAI/openarchivesprotocol.html>. Acesso em: 29 jan. 2001.

PACKER, Abel *et al.* SciELO: uma metodologia para publicação eletrônica. *Ciência da Informação*, Brasília, v. 27, n. 2, 1998.

ROBREDO, Jaime, CUNHA, Murilo Bastos. *Documentação de hoje e de amanhã*: uma abordagem automatizada da biblioteconomia e dos

Fonte: *Marcondes* (2002).

a) Primeiro, tente acessar o artigo indicado pela seta vermelha, usando o ***Uniform Resource Locator*** **(**URL**)**: <http://www.ibict.br/cionline/300301/3030401.htm>. O que você encontrou? O que você acha que aconteceu?

b) Agora, tente acessar o URL: <http://www.abobrinhas.com.br/>. O que você encontrou? O que você acha que aconteceu nesse caso?

URL

Sigla para *Uniform Resource Locator* que, em uma tradução livre para o português, significa *Localizador padrão de recursos.* Representa o endereço de um recurso disponível em uma rede de internet ou intranet.

Resposta comentada

Quando tentamos acessar um endereço desatualizado, nos deparamos com o famoso "erro 404 – página não encontrada" ou "Esta página não está disponível". O periódico digital para onde a URL do primeiro exemplo apontava, o *Ciência da Informação* versão *on-line*, mudou de URL, e agora pode ser encontrado em: <http://revista.ibict.br/ciinf/>; o novo URL daquele artigo agora é: <http://revista.ibict.br/ciinf/index.php/ciinf/article/view/149/128>.

No caso do URL do item "b)", tivemos um resultado semelhante, pois a página deixou de ser operante.

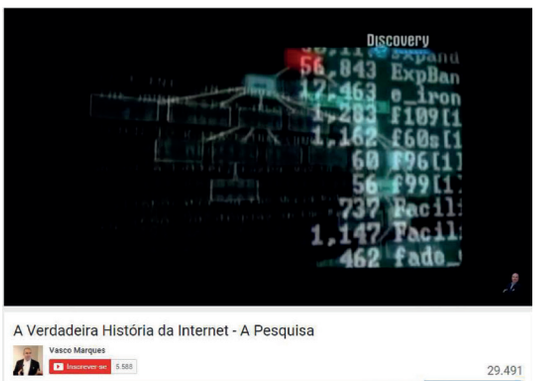
3.4.2 Indexação: costurando a grande "teia"

À medida que os serviços baseados na coleta, descrição e organização de páginas *web*, por profissionais de informação, mostravam suas limitações frente ao crescimento acelerado da rede, começaram a ser cogitadas formas de realizar esse trabalho automaticamente. Experiências com indexação automática, feitas por computadores que processavam o texto de documentos e dele extraíam palavras-chave usadas para indexar tais documentos, já existiam desde a década de 1960 (LUHN, 1960). Mas indexar toda, ou, pelo menos, grande parte da *web*, era um desafio muito maior. Em fins da década de 1990, isso começa a se tornar uma opção viável, com o surgimento dos primeiros mecanismos de busca.

Multimídia

Que tal saber mais sobre a história do surgimento dos mecanismos de busca? Se ficou curioso, assista ao documentário *A verdadeira história da internet – a pesquisa*, que pode ser encontrado no endereço: <https://www.youtube.com/watch?v=FHVPBueXudE>. Nele, você conhecerá a história dos mecanismos de busca da *web* e de como eles vieram a ser tornar um dos componentes mais importantes da economia da informação em nossos dias.

Figura 45 – Cena do vídeo

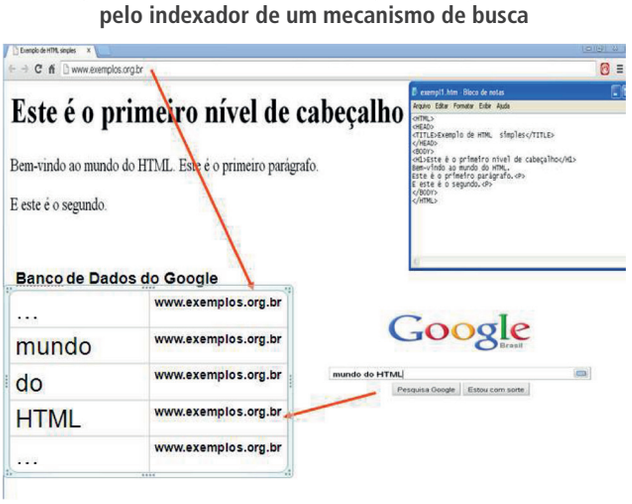


Fonte: captura de tela do vídeo.

Um mecanismo de busca é um serviço que tem dois componentes. Um deles é bem conhecido por todos: a página *web* onde nós podemos fazer buscas, por exemplo: <www.google.com.br>; <www.bing.com> ou <www.yahoo.com.br>. Mas, para que as buscas sejam possíveis, outro componente do mecanismo de busca deve atuar, ainda que de forma invisível para nós: o programa indexador de páginas *web*, muitas vezes também chamado de *crawler* ou *spider*. Ao acessar uma página *web*, o indexador coleta palavras existentes no texto da página e as armazena em uma base de dados, juntamente com o endereço dessa página e um sumário da mesma, geralmente, o seu texto inicial.

Na Figura 46, podemos observar uma página *web* fictícia, codificada em *Hypertext Markup Language* (HTML), mostrando as palavras do seu conteúdo, onde estão, entre outras, as palavras *mundo*, *do* e *HTML*, juntamente com as marcações da linguagem HTML (janela "Bloco de notas"). As palavras existentes na página são extraídas pelo programa indexador e armazenadas em uma base de dados (janela do banco de dados), juntamente com o endereço da página. É essa base de dados que consultamos que permite que façamos a busca pelas palavras "mundo do HTML" (na caixa de entrada do *Google*) e que encontremos essa página.

Figura 46 – Processo de indexação de páginas *web* pelo indexador de um mecanismo de busca



Fonte: produção do próprio autor.

Explicativo

A *Hypertext Markup Language* (HTML) é a linguagem por meio da qual são escritas as páginas da *web*. Ela permite inserir em uma página, simultaneamente, texto, imagens, recursos multimídia e também *links* para outras páginas, tornando a *web* uma teia virtual de conteúdos ligados por *links*. Veja mais detalhes em: <http://www.ipl.org/>. Acesso em 3 ago. 2021.

O processo de percorrer as páginas *web* para indexá-las, feito pelo programa indexador do mecanismo de busca, é chamado de *web crawling*. Outros mecanismos de busca têm um esquema semelhante para indexarem a *web*. Os detalhes dos esquemas de cada mecanismo de busca são segredos comerciais, porque respondem pelas prioridades, diferenciais e cobertura na indexação de cada mecanismo de busca.

Explicativo

O *Google* criou o *Manual para Editores da web*, que explica o processo de *web crawling*, realizado pelo seu indexador, da seguinte forma: o *Google* tem um grupo de computadores – o *Googlebot* – que está continuamente rastreando (navegando) bilhões de páginas na *web*. Esse processo de rastreio é algorítmico, ou seja, os programas de computador determinam quais os *web sites* que devem ser rastreados, com que frequência e quantas páginas de cada *web site* devem ser analisadas.

O processo de rastreio do *Google* inicia com uma lista de URL de páginas da *web*. À medida que o *Googlebot* navega nesses *web sites*, ele detecta os *links* de cada página e os adiciona à sua lista de páginas a serem rastreadas. O *Googlebot* faz uma cópia de cada uma das páginas que rastreia, a fim de compilar um índice de grande proporção de todas as palavras que visualiza. Essa lista também indica o local onde cada palavra surge em cada página.

Fonte: GOOGLE, 2013.

Os mecanismos de busca são quase um mito hoje em dia, devido à crença na capacidade que eles teriam de responder tudo; usamos, frequentemente, expressões como "pergunte ao *Google*", ou "oráculo digital", de uma forma quase mítica, para nos referirmos a eles. Esse esquema de indexação dos mecanismos de busca pode parecer completo e infalível, mas, se o examinarmos de perto, perceberemos a existência de vários senões.

Por exemplo, o que dizer de conteúdos de páginas que não são textuais, como imagens, música e vídeo, que não têm palavras para

permitir que sejam indexadas? Ou de páginas que os mecanismos de busca *intencionalmente* não visitam, por exemplo, por conterem idiomas exóticos, por serem pouco consultáveis, por terem pouco interesse comercial ou por estarem situadas muito profundamente na hierarquia de páginas de um *site*? Ou ainda, o que dizer de páginas cujos administradores demandaram explicitamente que não fossem indexadas, por meio do arquivo robots.txt?

robots.txt

Arquivo no formato .txt (bloco de notas). Funciona como um filtro para os robôs dos *sites* de busca e faz com que os desenvolvedores de páginas da *web* controlem permissões de acesso a determinadas páginas ou pastas dos *sites*. O robots.txt controla qual informação de um *site* deve ou não ser indexada pelos *sites* de busca.

Fonte: SCHULTZE, [2013?].

O *Google* apresenta os resultados de uma busca em ordem de *relevância*. A relevância que o *Google* atribui a cada página é calculada, principalmente, por meio de um programa interno ou algoritmo denominado *Pagerank*. Além disso, os mecanismos de busca são um serviço comercial. Páginas (*sites*) pagam aos mecanismos de busca para que seus *links* sejam mostrados prioritariamente, à medida que os usuários buscam por palavras-chave que possam existir nessas páginas.

Multimídia

Como bibliotecário, é importante saber mais sobre o *Pangerank*, pois esse é um conhecimento importante para compreender como o *Google* funciona. Para mais detalhes sobre o seu funcionamento, leia o material disponível em: <http://pt.wikipedia.org/wiki/PageRank>. Acesso em: 3 ago. 2021.

Curiosidade

No topo do *ranking*

Uma estratégia de marketing usada por várias empresas é a de pagar aos serviços de busca para controlar as palavras-chave que promoverão seus negócios, bem como a posição em que elas aparecerão na página de resultados da busca.

Figura 47 – Topo

Fonte: *Pixabay*.[13]

[13] Autor: *Sebadelval*. Disponível em: <https://pixabay.com/pt/silhueta-caminhante-topo-da-montanha-74565/>. Acesso em: 3 ago. 2021.

O *Google* oferece esse serviço comercial com o nome de *Adwords* e você pode encontrar mais detalhes sobre ele no endereço: <https://adwords.google.com.br/>. Acesso em: 3 ago. 2021.

Outra questão, para que o esquema de indexação dos mecanismos de busca funcione, é que cada página *web* deve ser associada a seu endereço. Esse endereço não pode ser alterado entre o momento em que o indexador visita a página e aquele em que consultamos a base de dados do mecanismo de busca.

Imagine se o administrador do servidor que hospeda a página altera o nome da pasta onde a página está armazenada, ou o nome do arquivo da página (de exemplo.html para exemplo1.html) – coisas que acontecem com razoável frequência na administração de páginas *web*. Quando esse tipo de coisa acontece, o endereço que o mecanismo de busca tem na sua base de dados, associado às palavras do conteúdo da página, não vai mais corresponder ao endereço em que ela pode ser acessada. Ao clicarmos no *link* dos resultados apresentados pelo mecanismo de busca, teremos como resposta o famoso "erro 404 – página não encontrada".

Outra exigência para que esse esquema de indexação de páginas pelos mecanismos de busca funcione é que cada página tenha um endereço razoavelmente *permanente*, que não mude desde o momento em que o programa indexador visita a página até que uma consulta seja feita pelas palavras do seu conteúdo. Vimos como várias questões ligadas à administração do servidor e da página podem afetar esse endereço, fazendo com que ela não seja mais encontrada, pelo menos até que o programa indexador visite a página novamente, em seu novo endereço. O endereço ou *link* apresentado como resposta à nossa consulta pelos mecanismos de busca é o que nos permite acessar a página que contém as palavras-chave especificadas na consulta.

Mas será que todos os conteúdos publicados na *web* possuem um endereço razoavelmente permanente, como: <www.uff.br>; <www.bn.gov.br>; <www.caixa.gov.br> ou <www.google.com.br>, os quais, muitas vezes, guardamos de cabeça, de tanto que são acessados no nosso dia a dia? Na próxima seção, veremos mais detalhes sobre os endereços permanentes ou identificadores persistentes para as páginas *web*.

3.4.3 **O essencial pode ser invisível aos olhos (do indexador)**

Imagine que você consultará uma base de dados na *web*. Pode ser, por exemplo: uma visita à página da *Receita Federal*, para checar se sua devolução do imposto de renda já foi liberada; à página do *Departamento de Trânsito* (DETRAN) do seu estado, para verificar se seu carro tem alguma multa; uma consulta ao catálogo da *Biblioteca Nacional*, ou você pode imaginar, ainda, que consultará a sua conta bancária. A resposta a essas consultas é gerada de forma dinâmica pelo programa gerenciador de bases de dados e é publicada na *web*. O conteúdo é exibido *exclusivamente* para você, que está fazendo a consulta, em função do seu *Cadastro de Pessoas Físicas* (CPF) ou do *Registro Nacional de Veículos* (RENAVAM) do seu carro. Ele só existe no momento da consulta e se extingue quando você sai da página, por isso, *não tem um endereço permanente*.

De acordo com *Bergman* (2001), os primeiros sistemas gerenciadores de bases de dados (SGBDs) disponibilizados por meio da *web* surgiram a partir de 1996. Antes disso, eles já eram largamente empregados pelas empresas, universidades e pelo governo, para armazenarem seus dados. Com as facilidades da consulta através da *web*, o uso das bases de dados *on-line* só vem crescendo. A questão é que conteúdos armazenados em bases de dados que são publicados na *web* dinamicamente, como resultado de uma consulta feita, só existem enquanto páginas *web* no instante da consulta; eles não possuem um endereço permanente e, portanto, não podem ser indexados pelos mecanismos de busca.

Os conteúdos de alguma maneira disponíveis, publicados na *web*, mas que têm alguma das características citadas anteriormente, formam a chamada *web invisível*, porção dos conteúdos da *web* não visível pelos indexadores dos mecanismos de busca.

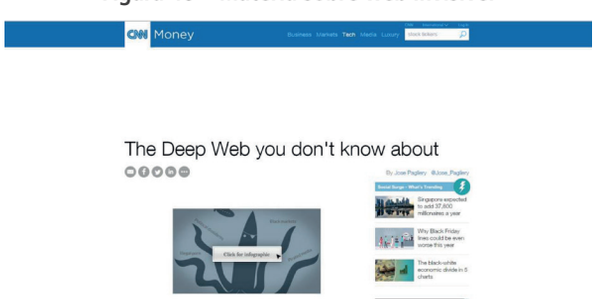
Multimídia

A *web* que você desconhece

Segundo *Bergman* (2001), um dos primeiros autores a estudar o fenômeno da *web* invisível, foi estimado que, no ano 2000, as dimensões da *web* invisível eram cerca de 1.000 a 2.000 vezes maiores que as da *web* indexável pelos mecanismos de busca. Dados mais recentes sobre esse tema podem ser encontrados em uma matéria da *Cable News Network* (CNN), no endereço: <http://money.cnn.com/2014/03/10/technology/deep-web/>. Acesso em: 26 nov. 2015.

Não deixe de ver o infográfico, é muito interessante!

Figura 48 – Matéria sobre *web* invisível



Fonte: captura de tela da página da *CNN*.

Como um *iceberg*, que tem sua maior parte debaixo da água, a maior parte da *web* é "submersa", ou seja, invisível aos indexadores dos mecanismos de busca (Figura 49).

Figura 49 – Este *iceberg* é uma analogia que permite visualizar a relação entre *web* visível X *web* profunda

Fonte: produção do próprio autor a partir de imagem da *Wikipédia*.[14]

Sherman e *Price* (2001), que também estudaram a *web* invisível, relacionaram e classificaram os motivos a seguir para a invisibilidade de muitos conteúdos da *web*.

a) *Web* opaca:
   - profundidade da indexação (páginas situadas muito profundamente na hierarquia de um *web site*);
   - n.º máximo de registros recuperados;
   - frequência da indexação;
   - páginas sem *links*, desconectadas;
   - tecnicamente indexáveis, mas ignoradas por razões políticas ou de negócios.

b) *Web* privada:
   - páginas protegidas por senhas ou por *firewalls*;
   - páginas reservadas por robot.txt;
   - páginas com *meta tags noindex* incluídas em seu código HTML.

c) *Web* proprietária:
   - páginas que exigem registro;
   - páginas que cobram taxas para acesso.

d) *Web* realmente invisível:
   - conteúdos em formatos não indexáveis (.MP3, .MP4, .WAV, .ZIP, *Flash*, *Shockwave*, *Javascript*);
   - páginas dinâmicas, que não têm endereço permanente, conteúdos armazenados em bases de dados (exemplos: devolução do imposto de renda, consulta ao saldo bancário, multas de trânsito etc.).

Firewall

Software ou *hardware* que verifica informações provenientes da internet ou de outra rede e, dependendo da configurações que ele possui, bloqueia ou permite que tais informações cheguem ao seu computador.

Fonte: AJUDA do Windows 10, c2017.

Meta tags noindex

Funcionam como etiquetas usadas para descrever o conteúdo de um *site* para os buscadores. Nelas, são inseridas as palavras-chave que facilitarão o encontro de determinada página em uma busca. Mas as *meta tags* também podem funcionar para que uma página não seja encontrada. Para isso, existe a *meta tag noindex*. No caso do *Google*, por exemplo, quando o *Googlebot* rastreia uma página e encontra a *metatag noindex*, ele exclui essa página inteiramente dos resultados da pesquisa *Google*, mesmo que outros *sites* apontem para ela.

Fontes: AJUDA do Search Console, c2017; SARTI, 2011.

[14] Autor: *Uwe Kils*. Disponível em: <https://en.wikipedia.org/wiki/Iceberg#/media/File:Iceberg.jpg>. Acesso em: 3 ago. 2021.

3.4.4 **Atividade**

Com base em alguns dos motivos de invisibilidade de conteúdos da *web*, segundo a classificação de *Sherman* e *Price* (2001), diferencie a *web* visível da *web* profunda (ou invisível).

Resposta comentada

Web visível:

– páginas que têm um endereço permanente.

Web profunda:

– páginas geradas dinamicamente quando um usuário consulta um banco de dados. Essas páginas não têm um endereço permanente;

– páginas situadas muito profundamente na hierarquia de um *website*;

– páginas protegidas por senhas, por terem conteúdos comerciais, ou com acesso impedido por *firewalls*;

– conteúdos em formatos não indexáveis (.MP3, .MP4, .WAV, .ZIP, *Flash*, *Shockwave*, *Javascript* etc.).

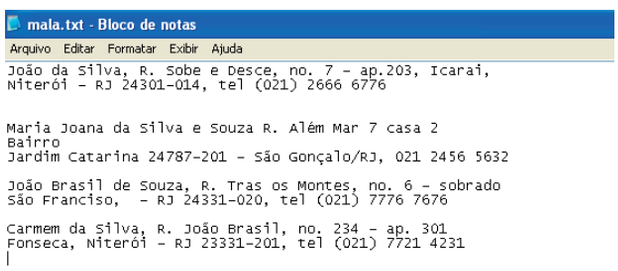
3.5 COMO A ESTRUTURAÇÃO DA INFORMAÇÃO AFETA OS MECANISMOS DE BUSCA

Vimos, na seção anterior, como os indexadores dos mecanismos de busca indexam as páginas da *web*, coletando palavras existentes nas páginas e montando uma base de dados com índices das palavras associadas ao endereço da página onde cada uma delas foi encontrada. Na

Unidade 2, entre os tópicos que discutimos, tratamos das unidades de significado, você se lembra? Falamos que os conceitos, muitas vezes, são formados por mais de uma palavra. A indexação por palavras é bastante imprecisa semanticamente, já que uma palavra pode ter vários significados, muitas palavras podem ter o mesmo significado e conceitos podem ser formados de diversas palavras.

Vamos entender esses problemas que afetam a indexação com um exemplo. Observe a Figura 50, é a imagem de um arquivo textual criado com o editor de textos *Notepad*.

Figura 50 – Informações não estruturadas em um arquivo .TXT



Fonte: produção do próprio autor.

Notepad

Em português, significa bloco de notas. É um editor de texto que vem em todas as versões do *Microsoft Windows*. Seu uso mais comum é o de exibir ou editar arquivos de texto e, por ser bem simples, com poucos recursos, acaba sendo muito utilizado por programadores, para escreverem seus códigos em ambientes mais limpos.

Percebemos que esse arquivo guarda informações do que seria uma mala direta (mala.txt), com nomes e endereços de clientes. Repare que o termo "*João Brasil*" aparece duas vezes no arquivo, uma vez como o nome de um cliente (8ª linha) e outra como o nome de uma rua (10ª linha). Se esse arquivo fosse indexado por palavras, como fazem os indexadores dos mecanismos de busca, e se fizéssemos uma pesquisa por um cliente que tivesse como nome "*João Brasil*", encontraríamos duas respostas, uma correta e outra incorreta. Em um arquivo textual, não há como especificar que queremos "*João Brasil*" sendo o nome de um cliente.

Curiosidade

A evolução dos editores

É importante compreendermos como um computador armazena conteúdos. Os primeiros computadores, ainda em fins da década de 1940 e início da década de 1950, eram máquinas que faziam, fundamentalmente, cálculos complexos. Desde a década de 1960 até os nossos dias, o computador é, cada vez mais, uma máquina que, além de fazer cálculo, armazena conteúdos. Este, aliás, é o principal uso dos computadores atualmente. Exemplos disso são a quantidade de conteúdos que armazenamos em nossos *smartphones* (músicas, fotos, mensagens etc.) e que as empresas armazenam em seus bancos de dados.

Figura 51 – Bloco de notas

Fonte: *Wikimedia Commons*[17].

O *Notepad* é um editor de textos simples. Em um editor de textos sofisticado, como o *MS WORD*, os caracteres são gráficos, desenhados ponto a ponto, podendo ter diferentes fontes (*Times New Roman, Courier, Verdana* etc.), diferentes tamanhos, cores e podem, ainda, receber destaque em negrito, itálico e sublinhado. Por sua vez, o *Notepad* só trabalha com caracteres do código *American Standard Code for Information Interchange* (ASCII), que é o padrão de código para representação de conteúdos mais antigo da indústria de computação (década de 1960 do século XX). No código ASCII, cada tecla do teclado do seu computador gera um código de 8 bits (= 1 *byte*) para identificar cada caractere; quando acionamos as teclas *Shift* + A, o computador armazena na memória o código 01000001, que é o código ASCII correspondente à letra A. Oito bits permitem um total de 256 combinações, ou seja, o código ASCII só tem 256 caracteres diferentes. Parece pouco, mas, na década de 1960, isso era o máximo.

Veja, agora, a Figura 52. Ela mostra a tabela de um sistema gerenciador de bases de dados, no caso, o *Microsoft Access*, com as mesmas informações sobre clientes contidas em nosso arquivo textual da Figura 50. Entretanto, neste caso, as informações estão separadas por campos da base de dados, a saber: nome, logradouro, número, complemento, CEP, bairro, cidade, estado, DDD e telefone.

Figura 52 – Informações estruturadas em uma tabela de banco de dados

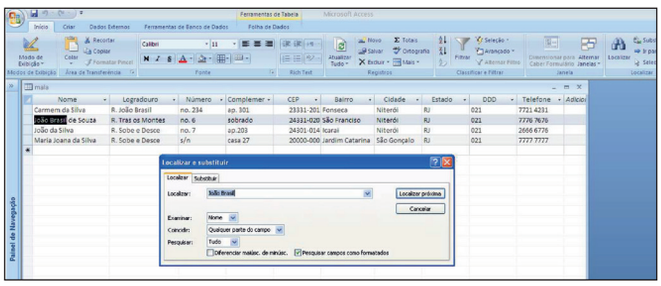
Nome	Logradouro	Número	Complemento	CEP	Bairro	Cidade	Estado	DDD	Telefone
João da Silva	R. Sobe e Desc	no. 7	ap.203	24301-014	Icaraí	Niterói	RJ	021	2666 6776
Maria Joana da Silva	R. Sobe e Desc	s/n	casa 27	20000-000	Jardim Catarina	São Gonçalo	RJ	021	7777 7777
João Brasil de Souza	R. Tras os Mon	no. 6	sobrado	24331-020	São Francisco	Niterói	RJ	021	7776 7676
Carmem da Silva	R. João Brasil	no. 234	ap. 301	23331-201	Fonseca	Niterói	RJ	021	7721 4231

Fonte: produção do próprio autor.

Poderíamos, então, utilizar as facilidades do sistema gerenciador de bases de dados e fazer uma consulta por clientes que tenham como nome "*João Brasil*". Veja a Figura 53, na qual usamos as ferramentas de recuperação de informações do *Microsoft Access* para fazer essa consulta.

[15] Em domínio público. Disponível em: <https://commons.wikimedia.org/wiki/File:Notepad_icon.svg>. Acesso em: 3 ago. 2021.

Figura 53 – Um gerenciador de banco de dados nos dá a possibilidade de fazer consultas contextualizadas sobre informações estruturadas em uma tabela



Fonte: produção do próprio autor.

Um sistema gerenciador de bases de dados nos permite maior precisão na especificação de nossas consultas porque, ao contrário de um arquivo textual, ele nos deixa *contextualizar* a informação que desejamos obter como resposta. O contexto é dado ao especificarmos em qual campo da base de dados queremos que a informação seja recuperada. Na Figura 53, podemos ver a maneira de especificar uma consulta como "Encontre *João Brasil* no campo Nome". A diferença entre um arquivo textual e uma tabela, num sistema gerenciador de bases de dados, é que aquele contém informações que diríamos ser *não estruturadas* em campos, ao passo que esta contém informações *estruturadas* e contextualizadas em campos.

Informações digitais não estruturadas, ou seja, documentos formados essencialmente por textos, como documentos administrativos, artigos científicos, manuais, normas, leis e textos em geral são largamente usados em nossa sociedade e na economia da informação em que estamos inseridos. O problema de armazenarmos informações não estruturadas em meio digital é que, como já vimos, há impactos na indexação automática para a construção de índices de palavras pelos mecanismos de busca, o que dificulta buscas com maior precisão semântica. Por outro lado, como visto anteriormente nesta Unidade, informações estruturadas em tabelas de bancos de dados são invisíveis aos indexadores dos mecanismos de busca, sendo uma das partes da *web* invisível. A solução para essa questão vem sendo a inserção de informação estruturada em documentos textuais.

3.5.1 Metadados: ajudando a estruturar a informação

Se você entrar na página da *Universidade Federal Fluminense* (UFF), <www.uff.br>, poderá visualizar o código-fonte da página, por meio de comandos no seu navegador. Ao realizar esses comandos (Figura 54), é possível encontrar, na codificação da página em linguagem HTML, a seguinte linha:

Figura 54 – Comandos

```
<meta name="keywords" content="UFF, Academia,
Federal Fluminense, Vestibular, Niterói, Faculdade" />
```

Fonte: produção do próprio autor.

Essa linha (e várias outras semelhantes, encontradas na codificação da página) é a marcação *meta* da linguagem HTML. Ela serve para inserir metadados na codificação da página, um conjunto de palavras-chave que descrevem o seu conteúdo, como: UFF, Academia, Federal Fluminense, Vestibular, Niterói, Faculdade. Em outra linha, é possível encontrar a marcação meta (Figura 55) utilizada para inserir o metadado especificando o idioma do conteúdo da página:

Figura 55 – Marcação meta

```
<meta name="content-language" content="pt-br" />
```

Fonte: produção do próprio autor.

As marcações meta são uma das maneiras de se inserir informação estruturada no conteúdo das páginas *web* em linguagem HTML. Veja que, se estivéssemos recuperando páginas *web* sobre vestibular, da mesma forma que na recuperação da informação em um banco de dados, poderíamos especificar o contexto usando palavras-chave (*keywords*) do conteúdo que queremos recuperar (vestibular), ou a data (*date*) em que a página foi criada, ou o idioma (*language*) da página (português brasileiro, pt-br).

As marcações meta da linguagem HTML fazem parte do conteúdo da página, mas quando esta é exibida normalmente, em um navegador, elas *não são mostradas*. As marcações meta são invisíveis para as pessoas que estão navegando na página, mas visíveis aos indexadores dos mecanismos de busca, a maioria dos quais têm capacidade de entendê-las e, dessa forma, conseguem indexar com mais qualidade uma página *web* que faça uso desse recurso.

A inserção de informação estruturada, semanticamente mais rica e inteligível por programas, nas páginas *web*, permite que estes processem os conteúdos dessas páginas de forma mais precisa. Esses são os programas indexadores dos mecanismos de busca, os programas navegadores, também chamados de *browsers*, como o *Google Chrome*, o *Mozilla Firefox* ou o *Internet Explorer*. Esses programas são chamados, em geral, de agentes inteligentes.

Multimídia

Que tal uma leitura para saber mais sobre programas agentes? Como sugestão, temos o estudo a seguir, que apresenta o conceito de agentes de *software* e suas aplicações:

NERI, Edmilson Lucena. Agentes de *software*: delegando decisões a programas. **RAE eletrônica**, São Paulo, v. 4, n. 1, art. 3, jan./jun. 2005. Disponível em: <http://www.scielo.br/pdf/raeel/v4n1/v4n1a03.pdf>. Acesso em: 7 fev. 2016.

3.5.2 **Microformatos**

Além das marcações meta do HTML, existe outra forma de se inserir informação estruturada no conteúdo das páginas *web*: são os microformatos. Os microformatos são conjuntos de conteúdos padronizados que usam, além da marcação meta do HTML, outras marcações das versões mais recentes do HTML ou do XHTML, para indicarem metadados específicos incluídos nas páginas *web*. Veja o exemplo a seguir, do microformato *hCards* (Figura 56), que especifica, em formato inteligível por programas, informações sobre contatos pessoais em páginas *web*.

Figura 56 – *hCard microformat*

***hCard microformat*:**

```
<p class="vcard">
<span class="fn">Oli Studholme</span>
<a class="url" href="http://oli.jp/">http://oli.jp</a>
, or
<a class="url" href="http://twitter.com/boblet">
follow me on Twitter (@<span class="nickname">boblet</span>)</a>
.</p>
```

Fonte: produção do próprio autor.

Quando uma página com a codificação desse microformato do exemplo é exibida no navegador, o que o usuário vê, na tela, é o seguinte (Figura 57):

Figura 57 – Microformato

Oli Studholme — http://oli.jp, or follow me on Twitter (@boblet).

Fonte: produção do próprio autor.

Veja outro exemplo, agora do microformato *rel-lincense* (Figura 58), usado para inserir metadados, que serão legíveis para programas, sobre o tipo de licença sob a qual pode ser disseminado o conteúdo da página *web*.

Figura 58 – *Rel-license microformat*

rel-license microformat

```
<small>This article is licensed under a
<a rel="license"href="http://creativecommons.org/licenses/by-nc-sa/2.0/">
Creative Commons Attribution Non-commercial Share-alike
(By-<abbr>NC</abbr>-<abbr>SA</abbr>) license</a>.
</small>
```

Fonte: produção do próprio autor.

Uma página *web* com a codificação do exemplo é assim exibida no navegador (Figura 59):

Figura 59 – Microformato

This article is licensed under a Creative Commons Attribution Non-commercial Share-alike (By-NC-SA) license.

Fonte: produção do próprio autor.

Essa codificação explica que a página em questão pode ser distribuída, copiada etc., segundo a licença de uso *Creative Commons*.

Creative Commons

Organização sem fins lucrativos que permite o compartilhamento e o uso da criatividade e do conhecimento através de licenças jurídicas gratuitas.

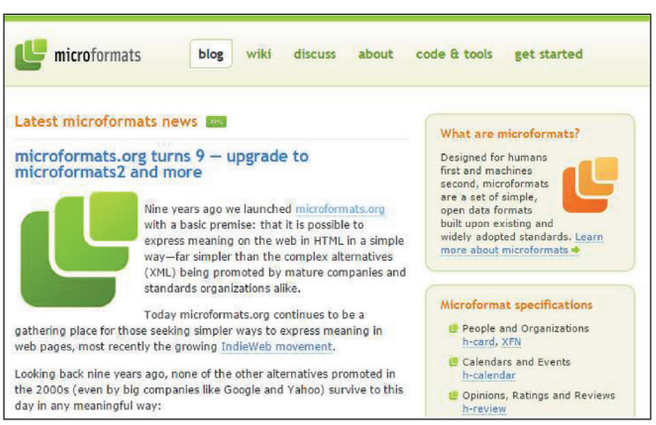
Fonte: SOBRE, [201-?].

Multimídia

Indo mais fundo no microformato

Se você ficou interessado em saber mais sobre essa ferramenta chamada microformato, acesse: <http://microformats.org/>, se aprofunde no tema e, inclusive, aprenda sobre os diferentes padrões existentes atualmente. Mas, atenção: o *site* está na língua inglesa.

Figura 60 – Página *Microformats*



Fonte: captura de tela da página.

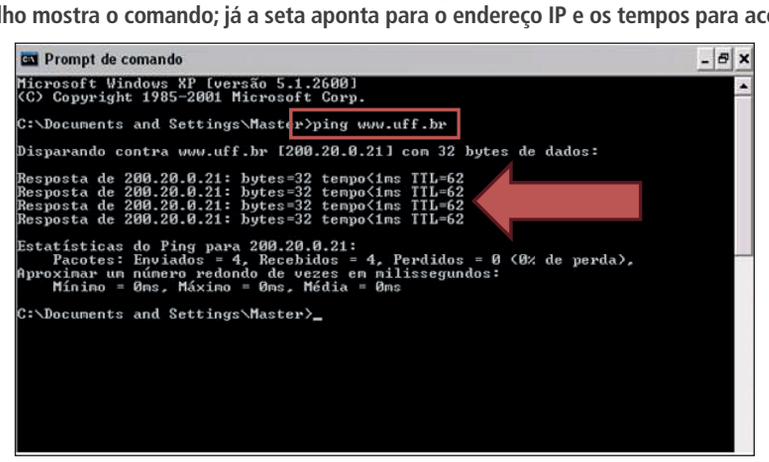
Em resumo, podemos dizer que informação não estruturada e informação estruturada referem-se a conteúdos inteligíveis por programas, e são esses programas que irão indexar, com maior ou menor precisão, os conteúdos da *web*. Informação não estruturada refere-se a conteúdos puramente textuais, cuja única estrutura é a dada pela linguagem, ou seja, palavras separadas por espaços. É assim que um programa vê um conteúdo textual: palavras e espaços. Já a informação estruturada ou se refere àquela armazenada em bancos de dados onde os campos dão o contexto, ou à informação descrita por metadados ou microformatos.

3.6 IDENTIFICADORES PERSISTENTES PARA GARANTIR O ACESSO PERMANENTE AOS RECURSOS NA *WEB*

Os endereços de páginas *web* como nós os conhecemos, do tipo <www.uff.br>, <www.bn.gov.br>, <www.caixa.gov.br> ou <www.google.com.br>, são como nomes fantasia, ou seja, nomes que facilitam a memorização, e são chamados tecnicamente de *nomes de domínio*. Mas, na verdade, qualquer página *web* possui um endereço numérico chamado endereço *Internet Protocol* (IP). Podemos descobrir o endereço IP de uma página entrando no *prompt* de comando de qualquer computador com sistema operacional *Windows* e usando o comando *ping*. Esse comando verifica se uma página esta "no ar" e tenta se comunicar com ela.

Para ilustrar o uso do *ping*, vamos a um exemplo. Imagine que você entrou no *prompt* de comando do *Windows*. O primeiro passo é entrar com o comando *ping*, depois, com o endereço de uma página (por exemplo: <www.uff.br>), assim: C:> ping www.uff.br. Em seguida, deve teclar *enter*. O comando *ping* informará o tempo de resposta gasto para se comunicar com a página em questão, só que, em vez de mostrar o endereço da página como ele foi informado, ele dá seu endereço IP e o tempo gasto para contactá-la, como podemos ver na Figura 61:

Figura 61 – Execução e resposta do comando ping no prompt de comando. O retângulo vermelho mostra o comando; já a seta aponta para o endereço IP e os tempos para acessá-lo



Fonte: produção do próprio autor.

Podemos constatar que o endereço IP correspondente a <www.uff.br> é 200.20.0.21. Esse é o endereço IP usado pelo protocolo da internet, o *Transmission Control Protocol* (TCP/IP), para acessar a página <www.uff.br>. Os nomes fantasia e seus endereços IP correspondentes são requisitados e gerenciados por meio do serviço Registro.br.

Prompt de comando

Recurso do *Windows* que oferece um ponto de entrada para a digitação de comandos do *Microsoft Disk Operating System* (MS-DOS) e de outros comandos do computador. Ao digitá-los, você pode executar tarefas no computador sem usar a interface gráfica do *Windows*. Esse recurso é normalmente usado apenas por usuários avançados.

Fonte: AJUDA do Windows 10, c2017.

Protocolo TCP/IP

Para que dois computadores se comuniquem, eles precisam "falar a mesma língua". O TCP/IP é um conjunto de protocolos que permitem que computadores conversem entre si. TCP é a sigla de *Transmission Control Protocol*, que, em português, significa Protocolo de controle de transmissão. IP é a sigla para *Internet Protocol*, ou seja, Protocolo de internet.

Fonte: MARTINS, 2012.

Explicativo

No serviço oferecido pelo Registro.br, uma instituição que deseja ter uma página na *web* deve registrar seu nome de domínio (por exemplo: <www.uff.br>), verificando, antes, se ele já existe, para enfim obter seu próprio endereço IP. Para saber mais sobre esse serviço, visite o endereço: <http://registro.br/>. Acesso em: 3 ago. 2021.

Fica, ainda, a questão da criação de endereços que permaneçam inalteráveis, que não ocasionem o erro 404. Essa é uma questão muito séria para a internet. Uma vez que ela consiste em conteúdos "interlinkados", se os *links* são frágeis e se frequentemente saem de uso, toda a economia da rede vai por água abaixo. Seria preciso que ela dispusesse de endereços *inalteráveis*.

O URL foi criado no início da internet, quando esta era formada por uns poucos milhares de páginas. Paulatinamente, esse número foi aumentando, algumas surgiram, outras desapareceram, outras ainda tiveram seus URLs alterados, por diversas razões. Logo, o URL como esquema de endereçamento de recursos *web* mostrou suas deficiências. A mais óbvia delas resulta de sua natureza, já que ele é um endereço, não um identificador. É como se você fosse identificado perante a sociedade pelo seu endereço residencial, não por seu CPF; se mudar de endereço, você "desaparecerá" perante a sociedade. Isso é o que acontece com as páginas identificadas por URLs.

Veja os exemplos dos seguintes URLs:

a) <http://catalogos.bn.br/scripts/odwp032k.dll?t=nav&pr=livros_pr&db=livros&use=CS0&rn=1&disp=card&sort=off&ss=22978160&arg>;

b) <http://basesdedados.casaruibarbosa.gov.br/scripts/odwp032k.dll?t=bpr=crb_apes_pr&db=crb_apes_db&use=ch&disp=list&ss=NEW&arg=rb-rbic|13>;

c) <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-19652001000100010&lng=en&nrm=iso&tlng=pt>;

d) <http://www.professores.uff.br/marcondes/index.html>.

URLs como os dos exemplos acima misturam, em uma mesma cadeia de caracteres, elementos transitórios, como:

a) nome da máquina servidora;

b) método ou programa pelo qual o recurso é acessado (scielo.php);

c) tecnologias usadas;

d) caminho (*path*) do documento no disco rígido do servidor (<www.professores.uff.br/marcondes>);

e) nome do arquivo, que nem sempre é único (quantos documentos são identificados por INDEX.HTML, por exemplo?).

Para suprir as deficiências do URL, outros esquemas – não mais de endereçamento, mas de *identificação* de recursos *web* – vêm sendo propostos, alguns deles já estão, inclusive, em uso, embora ainda não de forma tão ampla quanto os URLs. Pode-se dizer que estamos passando por um momento de transição do esquema de endereçamento da *web*.

Existem vários esquemas de identificação persistente de recursos *web*, entre eles o *Uniform Resource Identifier* (URI), o *Digital Object Identifier* (DOI) e o *handle system*. O DOI, por exemplo, é o identificador de artigos de periódicos eletrônicos de um consórcio de editores científicos. Cada editor recebe um conjunto de DOI e, a cada artigo digital publicado, um DOI é assinalado. Referências a outros artigos também são identificadas por DOI e, com isso, ao clicar em uma referência bibliográfica, nunca deveria acontecer o erro 404.

Multimídia

Identificação persistente

Se você estiver curioso para saber detalhes sobre essas ferramentas mais recentes de identificação persistente de endereços da *web*, vale a pena ler sobre cada uma das que citamos na Unidade. Mais informações podem ser encontradas nos endereços que listamos a seguir:

a) URI: <https://en.wikipedia.org/wiki/Uniform_Resource_Identifier>;

b) DOI: <http://www.doi.org/>;

c) *handle system*: <http://www.handle.net/>.

Figura 62 – Página não encontrada

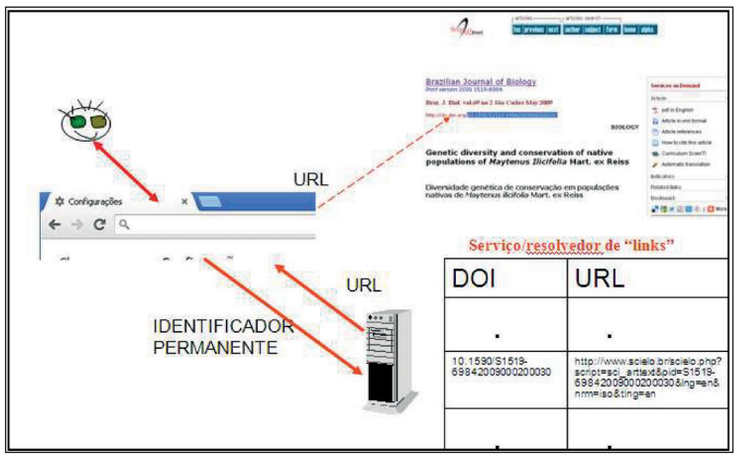
Fonte: *Wikimedia Commons*.[16]

Naturalmente, todos os esquemas de identificação persistente, por não serem endereços através dos quais um recurso pode ser acessado diretamente, necessitam de serviços – chamados serviços resolvedores. Os programas navegadores, para acessarem um recurso identificado por um URI, um DOI ou um *handle system*, primeiro recorrem a um serviço resolvedor. Esse serviço converte o identificador persistente na URL do recurso, devolve essa URL para o programa navegador e, aí sim, ele pode acessar

[16] Autor: *MdeVicente*. Disponível em: <https://commons.wikimedia.org/wiki/File:Notfound.png#/media/File:Notfound.png>. Acesso em: 3 ago. 2021.

o recurso. A Figura 63 apresenta um esquema que ilustra o processo que acabamos de descrever.

Figura 63 – Processo de resolução de um identificador persistente em um URL para acesso ao recurso



Fonte: produção do próprio autor.

3.6.1 Atividade

Para realizar esta atividade, você precisará ter disponível um dispositivo conectado à internet.

Na página da *DOI Foundation*, pode-se conhecer um serviço resolvedor. Lá, existe uma janela na qual se digita um DOI e o serviço resolvedor o converte em URL, fornecendo o acesso ao artigo correspondente.

Nossa proposta, então, é que você experimente acessar um artigo digital disponível no portal *The Scientific Eletronic Library Online* (SciELO). O artigo que queremos encontrar tem como título: *Genetic diversity and conservation of native populations of* Maytenus Ilicifolia *Mart. ex Reiss*. O URL desse artigo é: <http://www.scielo.br/scielo.php?pid=S1519-69842009000200030&script=sci_arttext> e seu DOI é: 10.1590/S1519-69842009000200030.

Primeiro, entre na página da *DOI Foundation*: <http://www.doi.org/>. Lá, você encontrará uma caixa com o título "*Resolve a DOI Name*" (na parte superior direita da tela); copie/escreva nela o DOI do artigo que queremos acessar.

Agora, responda às perguntas que se seguem.

a) O que aconteceu quando você digitou o DOI?

b) Você consegue acessar o artigo usando a URL que lhe oferecemos?

SciELO

Sigla de *The Scientific Eletronic Library Online* (Biblioteca Eletrônica Científica *On-line*). Como o nome já diz, é uma biblioteca eletrônica que abrange uma coleção selecionada de periódicos científicos brasileiros. Seu projeto prevê o desenvolvimento de uma metodologia comum para a preparação, o armazenamento, a disseminação e avaliação da produção científica em formato eletrônico.

Fonte: SCIELO, [20--?].

c) Como o serviço resolvedor utilizou o DOI que você digitou?

d) De onde veio esse número que você digitou na caixa do serviço resolvedor?

e) Qual a utilidade desse tipo de serviço?

Resposta comentada

A página da *DOI Foudation* mantém um serviço "resolvedor de DOI", ou seja, quando um DOI específico é fornecido, o programa resolvedor procura esse DOI na sua base de dados e devolve a URL correspondente ao programa navegador, para que este possa acessar o recurso, conforme ilustrado na Figura 63.

3.7 CONCLUSÃO

A *web* é um gigantesco repositório de informações, onde é possível encontrar respostas para praticamente tudo. No entanto, ela não é administrada de forma centralizada. A *web* tem pouco mais de 25 anos de criação, com padrões e tecnologias que evoluem constantemente, muitas vezes de forma não coordenada. Assim, o surgimento de um padrão ou tecnologia tem impacto sobre o anterior, criando um efeito dominó. O *World Wide Web Consortium (W3C)* é uma das organizações que tentam coordenar o desenvolvimento de padrões e tecnologias para a *web*.

Os mecanismos de busca gerais, usados por qualquer adolescente para encontrar as mais diversas informações na *web*, não são tão eficientes e infalíveis quanto a maioria das pessoas crê. É preciso saber que vários tipos de conteúdos, por diversos motivos, não são indexados.

Outro problema sério para o acesso aos recursos *web* no longo prazo são os endereços que mudam com frequência. Mecanismos de identificação permanente dos recursos *web* vêm sendo adotados para remediar esse problema, mas a *web* é gigante e eles ainda não são largamente usados.

RESUMO

Mecanismos de busca gerais, como o *Google*, o *Yahoo* e o *Bing*, indexam as páginas e os conteúdos existentes na *web*, mas sua cobertura não é completa e eles não indexam recursos em outro formato que não seja o textual – como imagens, sons etc. Essa parte da *web* não indexada pelos mecanismos de busca é chamada de *web invisível*.

Os mecanismos de busca também não indexam conteúdo armazenado em bases de dados. Isso porque esses conteúdos não têm um endereço permanente para serem referenciados. O uso de metadados e de microformatos pode ajudar os indexadores dos mecanismos de busca a indexarem melhor uma página *web*.

Outra questão são os endereços convencionais das páginas *web*, os URLs, que são instáveis e estão sempre sendo alterados. Uma mudança em um URL faz com que a página referenciada por ele não seja mais encontrada, resultando no famoso "erro 404 – página não encontrada". Esquemas de identificadores permanentes, como URI, DOI e *handle system*, são propostas para remediar essa questão, substituindo endereços (URLs) por identificadores permanentes. Esses identificadores, por não serem um endereço em si, têm que ser decodificados em URL para que os recursos identificados sejam acessados, o que é feito por novos serviços, invisíveis para os usuários, chamados de *serviços de resolução de identificadores*.

REFERÊNCIAS

AJUDA do Search Console. **Google**, [S.l.], c2017. Disponível em: <https://support.google.com/webmasters/answer/93710?hl=pt-BR>. Acesso em: 19 abr. 2021.

AJUDA do Windows 10. **Microsoft**, [S.l.], c2017. Disponível em: <http://windows.microsoft.com/pt-br/windows/what-is-firewall#1TC=windows-7>. Acesso em: 19 abr. 2021.

BERGMAN, Michael K. The deep web: surfacing hidden value. **Journal of Electronic Publishing**, [S.l.], 2001. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.12.363&rep=rep1&type=pdf>. Acesso em: 21 jul. 2010.

FACHINETTO, Eliane Arbusti. **O hipertexto e as práticas de leitura**. Santa Cruz do Sul: UNISC, 2005. Disponível em: <http://unisc.br/portal/images/stories/mestrado/letras/coloquios/ii/hipertexto_praticas.pdf>. Acesso em: 25 nov. 2015.

GOOGLE. **Tire o máximo partido do seu conteúdo**: um manual para editores da *web*. [S.l.]: Google, 2013. Disponível em: <http://www.i9socialmedia.com/e-book-oficial-do-google-para-produtores-de-conteudo-na-web/>. Acesso em: 25 nov. 2015.

LUHN, H., Keyword in context index for technical literature. **American Documentation**, Washington, v. 11, n. 4, p. 288-295, 1960.

MARTINS, E. O que é TCP/IP? **Tecmundo**, [S.l.], 2012. Disponível em: <http://www.tecmundo.com.br/o-que-e/780-o-que-e-tcp-ip-.htm>.

NERI, Edmilson Lucena. Agentes de *software*: delegando decisões a programas. **RAE eletrônica**, São Paulo, v. 4, n. 1, art. 3, jan./jun. 2005. Disponível em: <http://www.scielo.br/pdf/raeel/v4n1/v4n1a03.pdf>. Acesso em: 7 fev. 2016.

SARTI, E. Meta *tags*: o que são e como utilizá-las. **InfoWester**, [S.l.], 2011. Disponível em: <http://www.infowester.com/metatags.php>. Acesso em: 26 nov. 2015.

SCIELO. **SciELO**, São Paulo, [20--?]. Disponível em: <http://www.scielo.br/>. Acesso em: 19 abr. 2021.

SOBRE. **Creative Commons BR**, [S.l., 201-?]. Disponível em: <https://br.creativecommons.org/>. Acesso em: 19 abr. 2021.

SCHULTZE, Bernhard. Robots.txt: aprenda a configurá-lo: saiba como evitar que determinadas páginas do *site* apareçam no Google. **SEO Marketing**, Santo Amaro, [2013?]. Disponível em: <http://www.seomarketing.com.br/robots.txt.php>. Acesso em: 26 nov. 2015.

SHERMAN, Chris; PRICE, Garry. The invisible web: uncovering sources search engines can't see. **Library Trends**, Illinois, v. 52, n. 2, p. 282-298, Fall 2003. Disponível em: <https://www.ideals.illinois.edu/bitstream/handle/2142/8528/librarytrendsv52i2h_opt.pdf>. Acesso em: 13 nov. 2012.

Sugestão de Leitura

DIGIMIND. **Discover and exploit the invisible web for competitive intelligence**. [S.l.]: Digimind, 2007. Disponível em: <http://www.digimind.com/publications/white-papers/331-discover-and-exploit-the-invisible-web-for-competitive-intelligence.htm>.

GOOGLE. **Tire o máximo partido do seu conteúdo**: um manual para editores da web. [S.l.]: Google, 2013. Disponível em: <http://www.i9socialmedia.com/e-book-oficial-do-google-para-produtores-de-conteudo-na-web/>. Acesso em: 25 nov. 2015.

HAKALA, Juha. **Persistent identifiers**: an overview. Finland: TWR, 2010. Disponível em: <http://metadaten-twr.org/2010/10/13/persistent-identifiers-an-overview/>. Acesso em: 20 mar. 2012.

SAYÃO, Luís Fernando. Interoperabilidade das bibliotecas digitais: o papel dos sistemas de identificadores persistentes – URN, PURL, DOI, Handle System, CrossRef e OpenURL. **Transinformação**, Campinas, v. 19, n. 1, 2012. Disponível em:

<http://www.researchgate.net/publication/26462972_Padres_para_bibliotecas_digitais_abertas_e_interoperveis/file/9c960528a595537b92.pdf>. Acesso em: 13 maio 2010.