As dobras semióticas do ciberespaço: da *web* visível à invisível

The semiotic fold of cyberespace: from the visible to the invisible web

Silvana Drumond MONTEIRO[1]
Marcos Vinicius FIDENCIO[2]

Resumo

Após a instituição do ciberespaço, na década de 1990, a *Web* tornou-se o seu principal constructo e vem dobrando e desdobrando-se em vários sentidos: *Web* Invisível, *Web* Visível, *Web* Semântica, *Web* Pragmática, *Web* Social ou 2.0, entre outras. Em relação à *Web* Invisível, de acordo com os autores que escrevem sobre o tema, existe a inquietante questão: que nome dar a esse (des)território? *Web* Invisível, Profunda, Oculta, Escura? A partir da compreensão do conceito de dobra, criado por Leibniz e ressignificado por Deleuze, que explica os agenciamentos maquínicos e a visão pragmática dos aspectos técnicos e materiais das semióticas, fez-se uma prospecção conceitual da *Web* Invisível e de alguns mecanismos de busca que fazem a dobra com essa *Web* no ciberespaço. Para além da literatura, descobriu-se uma *Web* verdadeiramente escura, a *DarkWeb*, paralela e *underground* utilizada para o bem e para o mal, como previsível da espécie humana.

Palavras-chave: Ciberespaço. Dobra semiótica. *Web* invisível. *Web* visível.

Abstract

After the institution of cyberspace in the 1990s, the Web has become its main construct and has been folding and unfolding in several directions: Invisible Web, Visible Web, Semantic Web, Pragmatic Web, Web 2.0 or Social, among others. In particular, the Invisible Web, according to the authors who write on the subject, there is a disturbing question: what to call this (un)territory? Invisible Web, Deep, Hidden, Dark? From understanding the concept of fold, created by Leibniz and reframed by Deleuze, which explains the machinic assemblages and pragmatic view of the technical aspects of semiotics and material, a conceptual exploration of the Invisible Web and of some search engines was made that make the fold with these webs in cyberspace. Going beyond the literature, a Web truly dark was discovered, DarkWeb, parallel and underground, used for good and for evil, as expected of the human species.

Keywords: *Cyberspace. Semiotics fold. Invisible web. Visible web.*

Introdução

A partir da proposta epistemológica de estudar os agenciamentos maquínicos na organização do conhecimento e da informação no ciberespaço, uma categorização dos mecanismos foi elaborada e estudada objetivando comprovar as múltiplas sintaxes de organização, tendo como aporte teórico a heterogeneidade e a multiplicidade dos regimes de signos - princípios filosóficos do Rizoma (Deleuze; Guattari, 1995) e as matrizes da linguagem-pensamento (Santaella, 2005).

[1] Professora Doutora, Universidade Estadual de Londrina, Departamento de Ciência da Informação. Rod. Celso Garcia, PR 445, km 380, *Campus* Universitário, 86055-900, Londrina, PR, Brasil. Correspondência para/*Correspondence to*: S.D. MONTEIRO. *E-mail*: <silvanadrumond@gmail.com>.

[2] Acadêmico, Universidade Estadual de Londrina, Departamento de Ciência da Informação, Curso de Biblioteconomia, Londrina, PR, Brasil.

Recebido em 14/9/2012, e aceito para publicação em 30/11/2012.

Assim, essas hipersintaxes também refletem o momento atual - designado pós-moderno (para alguns), contemporâneo (para outros) -, do qual não temos uma visão esvaziadora, pois as Tecnologias da Informação e Comunicação (TIC) são mais que ferramentas, são peças heterogêneas que, conjugadas ou amalgamadas com os homens, formam determinadas máquinas: máquina abstrata, máquina social, máquina de guerra etc., para usar expressões deleuzianas.

Ademais, as máquinas também formam a *dobra*, conceito com espessura epistêmica complexa, criado por Leibniz e ressignificado por Deleuze (1991), que explica os agenciamentos "maquínicos" e a visão pragmática dos aspectos técnicos e materiais das semióticas. A dobra é uma prega que, em latim, significa *plica,* implicar, e quer dizer: dobrar, unir; já explicar é desdobrar. Seu efeito é que:

> A dobra, portanto, cria uma nova relação dentro-fora; uma nova *topologia*: quando o contato se realiza, isso equivale ao estabelecimento de ligações até então não concretizadas, apenas potenciais, entre os componentes dispersos originais (Oliveira, 2003, p.152, grifos do autor).

O próprio signo é uma dobra, pois pode dobrar-se, desdobrar-se e redobrar-se em vários tipos e semióticas. A dobra do signo instaura mais que o desdobramento do significante/significado, pois pensar a significação como ato (filosofia pragmática) implica pensar o signo como agenciamento maquínico.

Pode-se considerar, também, a *Web* Visível e Invisível como partes de uma dobra, com fronteiras difusas, às vezes ambíguas, mas intrinsecamente unidas, sendo a *Web* Visível o (des)dobramento da interioridade da *Web* Invisível. Assim, pode-se inferir ainda que, nos agenciamentos maquínicos, nas conexões com as TIC, as dobras estejam sempre presentes, configurando novas dimensões, novas topologias e novas possibilidades.

O movimento das semióticas nas TIC produz novas "dobras", tanto dos signos quanto do sentido, uma vez que não há delimitação entre a estrutura física e lógica, lembrando que a dobra é a continuidade do avesso e do direito, e o sentido se distribui dos dois lados, ao mesmo tempo, pois o signo im-*plica* o sentido e o sentido ex-*plica* o signo (Machado, 2009).

Implicação-explicação e envolvimento-desenvolvimento são atributos do signo, pois "O próprio sentido se confunde com esse desenvolvimento do signo, como o signo se confunde com o enrolamento do sentido" (Deleuze, 2010, p.84), assim o é também em relação à *Web* Visível/Invisível no ciberespaço.

É justamente nessa dobra semiótica dos mecanismos de busca que surgem as hibridizações dos mecanismos, das linguagens e da indexação; surgem também as hipersintaxes e as intersemioses. Além disso, ao categorizar os mecanismos de busca, em especial aqueles especializados em *Web* Invisível, descobriu-se uma dobra, uma *Web* maior, oculta, também designada "continente escuro", na qual esses mecanismos fazem a dobra com a *Web* Visível e mostram apenas uma pequena parte, mas insinuam a grande extensão que é o ciberespaço.

Se por um lado é fácil definir a *Web* Visível como aquela composta de páginas da *Web* em *HyperText Markup Language* (HTML), cujos motores de busca optaram por incluí-las em seus índices, a *Web* Invisível é muito mais difícil de se definir e de se classificar por várias razões, sejam elas tecnológicas, políticas ou operacionais.

De acordo com os autores que escrevem sobre a *Web* Invisível, existe a inquietante profusão conceitual sobre a *Web*: Invisível, Profunda, Oculta e Escura. Pode-se considerar todos esses conceitos, de acordo com as dobras de (in)visibilidade do ciberespaço? Segundo Bergman (2001), mais adequado seria a *Web* Profunda (para a *Web* Invisível), uma vez que o termo invisível não seria correto, pois a invisibilidade é apenas uma questão tecnológica ou mesmo política de indexação dos mecanismos de busca.

Já Sherman e Price (2001), na descrição das várias camadas da *Web*, deixam perceber que o termo invisível não é exatamente o par dicotômico da *Web* Visível, mas apenas a existência de planos de invisibilidade, como as desdobras ou texturas do ciberespaço.

Relacionou-se, em algum momento, a *Web* Visível com a indexável, pois explicita bem o olhar e interesse sobre esse objeto. No tocante a seu par, ou suas dobras, o artigo intenta desvelar seu campo semântico com mais vagar. Que nome dar a esse (des)território escuro? *Web* Profunda, *Web* Invisível ou Oculta? Ou todos os nomes?

Dessa forma, para continuar a estudar a *Web* Visível, há a necessidade de desenvolver estudos sobre a *Web*

Profunda, como uma dobra semiótica (ou várias) que compõe o ciberespaço e, especialmente, os buscadores específicos nesse setor.

Como nada é tão simples nos objetos contemporâneos, outra *Web* emerge, considerada *Dark Web* (*the dark side of the cyberspace*) ou a invisível, de fato, posto que servidores e a navegação feita sob o anonimato fazem a dobra *underground* do ciberespaço.

Mais que uma questão terminológica, esses agenciamentos para a Ciência da Informação implicarão pensar a máquina resultante da conjunção de determinado corpo social e suas semióticas e a organização em espaços digitais e, quiçá, explicarão, em parte, o ciberespaço.

A *web* visível

A preocupação com a indexação e a localização de recursos na *Web* é tão antiga quanto o surgimento da própria. Nos seus primeiros anos, a informação na *Web* era, basicamente, recuperada apenas mediante a memorização da *Universal Resource Locator* (URL).

Como método pioneiro de indexar e facilitar a busca na *Web*, destacam-se as ferramentas de procura em repositórios *File Transfer Protocol* (FTP) e os armazenados nos *Gophers,* como o *Archie* (Cendón, 2001).

A evolução incipiente da quantidade de conteúdo fez novas formas de organização ser construídas, aparecendo então os diretórios, hoje quase extintos no seu modelo clássico, que consistiam em índices de *sites* indexados manualmente, nos quais novas páginas podiam ser submetidas, na maioria das vezes, pelos próprios usuários por meio de critérios específicos.

A indexação manual, contudo, mostrava-se cada vez mais ineficaz, em face do volume de informação na *Web*, e sua consequente necessidade de indexação fez surgirem os mecanismos/motores de busca ou, simplesmente, buscadores (os pioneiros, *HotBot*, Altavista e *Northern Light*, entre outros), cujos índices eram e ainda são feitos por intermédio da indexação mecanizada (robôs), com a utilização de algoritmos matemáticos próprios para a localização e a indexação do conteúdo disperso no ciberespaço.

A evolução dos diretórios e mecanismos de busca é significativa, em 2000, o *Google* (motor de busca) e o *Web Top* (híbrido, motor de busca e diretório) indexavam, respectivamente, 56% e 50% do conteúdo total do ciberespaço (Cendón, 2001). De toda forma, boa parcela desse volume continua não indexado, ou seja, permanece invisível para a indexação mecanizada.

Nesse sentido, Rajaraman (criador do Kosmix) confessa que os mecanismos de busca indexam uma fração muito pequena do ciberespaço. 'Eu não sei, para ser honesto, que fração. Ninguém tem uma estimativa muito boa de como é grande a *Web* Profunda. De cinco a cem vezes maior do que a *Web* de superfície é a única estimativa que conheço' (Beckett, 2009, p.2).

A informação na *Web* pode ser categorizada, para fins de indexação, em suas diretrizes: a parte visível, ou seja, páginas que podem ser somadas ao banco de dados dos buscadores, e a parte invisível, cujo conteúdo, por razões expostas, não pode ser indexado pelos buscadores tradicionais.

Em 1994, Jill Ellsworth utilizou, pela primeira vez, o termo "*Invisible Web*" para designar o conteúdo que não era indexado pelos buscadores (Bergman, 2001). A pesquisa de Bergman (que prefere o termo *Web* Profunda à *Web* Invisível) também detectou que o conteúdo invisível na *Web*, ou a *Web* Invisível, era de 400 a 550 vezes maior do que estava até então na *Web* indexável, com 7 500 *terabytes* de informação comparado com 19 da *Web* indexável (visível).

É interessante observar que os mecanismos de busca investem grande soma de recursos para incrementar seus algoritmos de indexação e busca, atuando em uma plataforma de múltiplas sintaxes e semióticas, trazendo à superfície arquivos que antes eram considerados não indexáveis, bem como gerando padrões semânticos de busca a partir da pragmática dos leitores.

Vale ressaltar ainda que a *Web* Visível vem se especializando em nomenclaturas e práticas distintas, como *Web* 2.0 ou Social, *Web* Semântica, *Web* Pragmática, entre outras em devir no ciberespaço.

Com o crescimento da *Web*, seus limites estão tornando-se turvos. Beckett (2009) pergunta-se "Agora a *Web* é tudo?" Responder a esse questionamento im-*plica* em explicar alguns conceitos, além os pertinentes à *Web* Invisível.

Se em certo sentido sim, conceitualmente não, a *Web* não é tudo, mas é o principal constructo (dobra) do ciberespaço e vem crescendo a passos largos.

Ela pode ser definida como a interface de convergência entre as linguagens e a interoperabilidade necessária para efetuação das trocas simbólicas. Já o ciberespaço é um espaço semântico/semiótico, onde o signo se dá em várias semióticas, desterritorializado, nômade, em escrita espacializada e com a memória em constante modificação. Se a *Internet* é a rede mundial de computadores, base técnica do ciberespaço, este é a rede de signos e pessoas.

A *web* invisível

A *Web* Invisível nasce juntamente com a tecnologia de banco de dados no ciberespaço, posteriormente com a inclusão do *e-commerce* e, por último, com a adaptação dos servidores para permitir a visualização de informações por meio da geração de páginas dinâmicas.

Buscando uma definição de partida, Sherman e Price (2001, p.57, tradução nossa[3]) a definem como:

> Páginas de textos, arquivos, muitas vezes de alta qualidade e com autoridade informacional disponíveis na *World Wide Web* cujos motores de buscas gerais não podem, devido a limitações técnicas, ou não querem, por escolha deliberada, adicionar aos seus índices de páginas *Web*. Às vezes também é referida como '*Web* Profunda' ou 'material escuro'.

Bergman, em um relatório de 2001, afirma que a *Web* Profunda é imensurável e, no seu estudo, realizado entre 13 a 30 de março de 2000, apresentou alguns resultados interessantes, a saber:

- a *Web* Profunda é a maior categoria crescente de informações no ciberespaço;
- existem mais de 200 000 *sites* profundos;
- o conteúdo da *Web* Profunda é de alta qualidade;
- a qualidade do conteúdo total da *Web* Profunda é de 1 000 a 2 000 vezes maior que a *Web* de superfície;
- mais da metade do conteúdo da *Web* Profunda reside em base de dados especializadas;
- 95% da *Web* Profunda é gratuita, acessível ao público mediante assinaturas (Bergman, 2001, p.2).

Sherman e Price (2001), referindo-se ao relatório *Bright Planet*, afirmam que Bergman incluiu, em seu estudo, informações efêmeras, como *sites* de informações sobre o tempo, entre outros. Excluindo essas bases de dados e outras do gênero, estimam que a *Web* Invisível seja entre 2 e 50 vezes maior que a *Web* Visível.

Para ilustrar a indexação realizada pelos mecanismos gerais e específicos em *Web* Profunda, Bergman (2001) apresenta uma ilustração clássica (Figura 1).

Antes de se considerar as camadas de invisibilidade ou as dobras da *Web* Visível, elencará-se alguns motivos pelos quais o conteúdo do ciberespaço não é plenamente indexável.

O primeiro motivo é por questões técnicas ou deliberadas; o segundo, por políticas de exclusão ou impossibilidade tecnológica. Algumas considerações a respeito de cada diretriz são tecidas baseadas no exposto por Sherman e Price (2001) e Branski (2004).

Questões técnicas deliberadas

Os motores de busca alimentam seus índices através dos *Spiders*, *Crawlers*, ou *Robots*, termos cujo

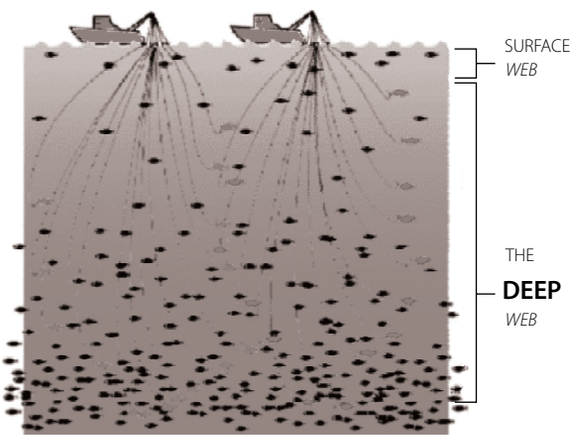

Figura 1. *Harvesting the Deep and Surface Web with a Directed Query Engine*.

Fonte: Bergman (2001, p.6).

[3] *Text pages, files, or other often high-quality authoritative information available via the World Wide Web that generalpurpose search engines cannot, due to technical limitations, or will not, due to deliberate choice, add to their indices of Web pages. Sometimes also referred to as the "Deep Web" or "dark matter".*

significado refere-se a robôs que efetuam uma varredura à procura de novas páginas na *Web*. Tais robôs trabalham com lógicas próprias que, por motivos comerciais, nem sempre estão acessíveis, embora sua função básica seja pesquisar, relacionar, adentrar diretórios e subdiretórios na *Web* e somá-los aos índices dos buscadores para os quais operam. Leem linguagens e instruções, as quais podem ser escritas exclusivamente para eles nos *sites*, no momento de sua construção.

As instruções estão nos arquivos *robots.txt* em operação com o *Robots Exclusion Protocol,* arquivo preparado pelo mantenedor de determinado *site* especialmente para informar os robôs sobre a não indexação de determinada página/recurso aos índices dos mecanismos.

Essa informação lógica é interpretada pelo robô como a atribuição de instruções específicas mediante o que é programado após as linhas *User-agent* e *Disallow* dentro do arquivo *robots.txt*. A lógica é simples: após *User-agent*, se houver asterisco, a instrução serve para qualquer mecanismo de busca, contudo, se houver alguma especificação, por exemplo, *googlebot*, o *robots.txt* será aplicado apenas para o *Google*. Já a linha *Disallow* instrui o robô de que tudo o que estiver depois da barra inclinada não poderá ser indexado: o *webmaster*, então, poderá tornar parcial ou totalmente invisível determinado *site* <http://robotstxt.org/robotstxt.html>. Essa situação pode ser observada no Quadro 1.

Há também a possibilidade de restringir um *site* aos robôs com a *metatag noindex*, colocada no cabeçalho de páginas HTML. Seu funcionamento é bastante simples:

```
<html>
<head>
<title>...</title>
<META NAME="ROBOTS" CONTENT="NOINDEX, NOFOLLOW">
</head>
```

O campo *meta name* indica para quem destina a instrução (nesse caso, para os robôs), enquanto *content* é a *tag* com a instrução específica, seguida, após a vírgula, com *nofollow*, que instrui os robôs a não analisar o *site* em questão.

Há algumas complicações com a utilização da *metatag noindex*. A "*The Web Robots Page*", página dedicada aos robôs desde 1995, aponta duas complicações principais:

1) os robôs podem ignorar a *metatag noindex*, principalmente aqueles que trabalham como *malware*, além de *spams* que "varrem" a rede procurando endereços de *e-mails*;

2) a instrução *nofollow* só se aplica à página em questão. É possível que um robô encontre algum *link* de entrada para a página instrução de impedimento.

Outra forma de exclusão de determinada página do campo da visibilidade tem relação com a forma de disponibilidade de informação. *Sites* cujos conteúdos são acessados por meio de senhas enquadram-se nessa situação, como também as páginas cuja natureza do conteúdo exige privacidade.

Exclusão por política ou por limitação tecnológica

Essas questões têm grande conexão com o formato de apresentação da informação, que impossibilita a leitura do conteúdo pelos robôs e, consequentemente, a indexação. Os mecanismos de busca têm uma séria dificuldade em indexar materiais não verbais ou que não

Quadro 1. O arquivo robots.txt.

www.site.com.br.html	www.site.com.br/videos.hmtl	www.site.com.br/noticias.html	www.site.com.br/esportes.html
Arquivo robots.txt (sempre na raiz do servidor)			
User-agent:	Instrução cumulativa	Instrução cumulativa	Instrução cumulativa
Disallow:/			
Conteúdo invisível	Conteúdo invisível	Conteúdo invisível	Conteúdo invisível

estão em *Hypertext Markup Language* (HTML), e a maioria deles não consegue indexar os seguintes tipos (Sherman; Price, 2001, p.58):

- PDF ou *Postscript* (exceto o *Google*);
- *Flash*;
- *Shockwave*;
- Programas executáveis;
- Material comprimido.

Em todos os casos, a informação encontra-se comprimida dentro de um formato de arquivo ou extensão (respectivamente, de acordo com os itens supracitados, .pdf, .ps, .flv, .swf, .exe/deb/bat etc, .zip/t.ar.gz/.rar etc.).

A dificuldade é como um robô poderá ler e indexar a informação comprimida em um formato não verbal. No caso de PDF, o *Google* é um dos únicos que conseguem estender seus robôs para efetuar a leitura do arquivo formatado, contudo, ler e interpretar multimídia como os arquivos de vídeo (*Flash, Shockwave* e outros) não é tarefa tecnologicamente fácil.

Alguns buscadores da *Web* Invisível dedicam-se a buscar informações desses tipos.

A *web* invisível: alguns apontamentos conceituais

Que nome dar a esse (des)território escuro? *Web* Profunda, *Web* Invisível ou Oculta? Araújo (2001) questiona-se: invisível ou oculta? Para o autor, o termo invisível parece ser inadequado por denotar algo completamente inacessível, fora de alcance, o que não é totalmente verdadeiro, pois basta que se saiba uma ferramenta de busca especializada ou mesmo a URL para ter acesso a esses conteúdos. Nesse sentido, o termo "oculta" seria mais apropriado.

Há, de fato, uma parcela que permanece invisível aos mecanismos de busca. Essa parcela da *Web* é composta por banco de dados aos quais o acesso é possível por meio de pagamento e/ou inscrição, pois "Por serem guardados em diretórios protegidos por senha, eles se encontram fora do alcance dos motores de busca" (Araújo, 2001, *online*).

Já o termo *Web* Profunda, de certa forma, também está relacionado a uma limitação de muitos motores de busca; o fato de eles não varrerem todo o conteúdo de um *site*, pois:

> Como dito anteriormente, os textos da *Web* costumam estar armazenados em diretórios de modo bastante semelhante à forma como guardamos textos em pastas em nossos PC. Uma pasta (diretório) pode conter outras pastas e assim por diante em uma relação de inclusão que pode alcançar vários níveis de profundidade. O fato relevante é que os motores de busca nem sempre são programados para fazer uma pesquisa em profundidade nos servidores da *Web* e param em determinado nível. O que estiver além dele não será encontrado nem indexado e, portanto, estará fora de alcance para o usuário (Araújo, 2001, *online*).

Pode-se deduzir, então, de acordo com Araújo (2001), que essas três realidades coexistem: invisível, oculta e profunda. Esta última seria a *Web* Opaca, de acordo com Sherman e Price (2001). Ainda, segundo Sherman e Price (2001), esse é o paradoxo da *Web* Invisível, pois é fácil compreender sua existência, mas é difícil defini-la concretamente com termos específicos.

A literatura sobre o assunto, via de regra, é internacional, ademais há uma discussão sobre a terminologia mais adequada. Em respeito às traduções, usou-se a terminologia empregada por seus respectivos autores, em algumas citações, mesmo sendo estas parafraseadas.

Traçar uma linha entre a *Web* Visível e a Invisível não é tão simples assim e, mais uma vez, o conceito da dobra reaparece, posto que os buscadores podem trazer à superfície alguns conteúdos.

Para Sherman e Price (2001) não existe uma classificação dicotômica entre visível e invisível, mas camadas, gradações de invisibilidade e acesso aos conteúdos no ciberespaço. Nesse sentido, apresentam quatro tipos de invisibilidade, começam com a opaca, relativamente acessível aos mecanismos, até chegarem à verdadeiramente invisível (Figura 2). Dito de outro modo, os motivos pelos quais os mecanismos não podem ver o conteúdo profundo, que são: a *Web* Opaca; *Web* Privada; a *Web* Proprietária e; a *Web* realmente Invisível.

Sherman e Price (2001) afirmam que essa classificação diz menos respeito às distinções rápidas e complexas e mais ao limite amorfo da *Web* que, de todo modo, torna sua definição difícil, a não ser, para nós, pela aproximação de conceito de dobra semiótica.

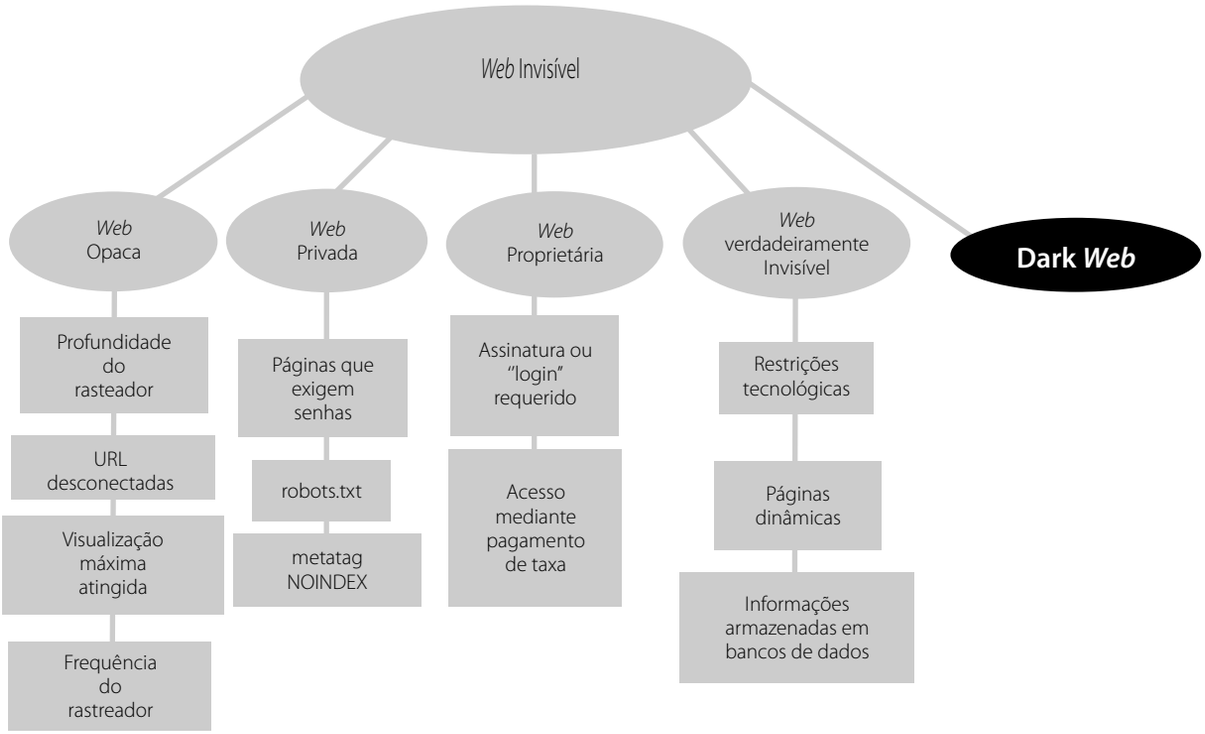

Figura 2. As várias *Web*.

Fonte: Adaptado de Ford e Mansourian (2006, p.585).

***Web* opaca**

A *Web* Opaca compõe-se de *sites* que misturam arquivos e mídias, dentre os quais alguns são facilmente indexáveis e outros são incompreensíveis aos rastreadores. Por isso mesmo, pela dificuldade em classificar esses *sites* em *Web* Visível ou Invisível, são designados como *Web* Opaca. Além disso, segundo Sherman e Price (2001), há outros motivos de cunho tecnológico para a existência dela, ou seja, arquivos que podem ser, mas não são incluídos nos índices dos mecanismos de busca, por várias razões, a saber:

a) profundidade do rastreador (*crawler*): reduzir a profundidade ajuda a reduzir os custos de indexação. No passado, era comum trazer apenas páginas "exemplares" de um *site* como citação de (boa) representação de sua existência. Apesar de os mecanismos não revelarem sua profundidade de rastreamento, há uma tendência para rastrear mais profundamente e indexar mais páginas;

b) número máximo de resultados visíveis: quando o número máximo de páginas visualizáveis for atingido, em resposta a uma pergunta, o mecanismo de busca retorna um número limite de resultados visíveis. As páginas que os algoritmos não incluíram, em ordem de relevância, tornam-se irrecuperáveis para aquela *query* em especial. Esse tipo de limitação é cada vez menos comum. Na maioria das vezes, os mecanismos mostram a quantidade de páginas recuperadas e, de toda forma, algumas cifras mostram a impossibilidade de percorrer até a última delas. Uma rápida pesquisa por "USA" no *Google* tem uma revocação próxima de 5 bilhões de resultados;

c) frequência do rastreador: pode ocultar páginas da *Web* Visível por algum tempo. Por isso é importante que a frequência seja eficiente, especialmente em *sites* que já foram indexados, devido a sua idade média; Sherman e Price (2001) explicam: depois de dois anos, um *site* até pode ter o mesmo número de URL, mas apenas a metade das páginas originais permanecem, as demais são novas;

d) URL desconectadas ou páginas que não têm *links*: isso ocorre porque existem duas formas básicas

para indexar o conteúdo da *Web*: ou o autor envia um pedido de submissão a um mecanismo ou o robô descobre por si próprio. Para que o segundo seja possível, é necessário que outras páginas, já indexadas, apontem para a nova e, dessa forma, quando o robô visitar uma página indexada verificará a existência de um novo *link* e, consequen-temente, a acrescentará em seus índices (Sherman; Price, 2001).

A *web* privada

A *Web* Privada consiste em páginas que são deliberadamente excluídas dos mecanismos, ou seja, o conteúdo possui restrição deliberada pelos mantenedores, por três motivos:

1) páginas protegidas por *password*: o conteúdo só é acessível para associados ou pessoas que tenham algum tipo de senha. A maioria dos fóruns de discussão se inclui nesse quesito e, mais recentemente, as redes sociais;

2) o uso de *no index*: impede que o robô indexe a página;

3) o uso de arquivos *robots.txt* para impedir o acesso de buscadores na página.

A diferença entre *no index* e *robots.txt* é basicamente a abrangência do limite de proibição da indexação. Enquanto o primeiro restringe o rastreamento de páginas, o segundo pode proibir a visita de um buscador no *site* inteiro, mediante uma lista de arquivos ou partes chamada *robots.txt*.

A *web* proprietária

Trata-se de conteúdo indexável, entretanto, restrito por ser propriedade de seus mantenedores (instituições e órgãos, entre outros), acessível mediante registro, em muitos casos gratuitos, assinatura e/ou pagamento de taxas.

Portais de conteúdo cuja visualização é realizada mediante assinatura enquadram-se nessa parcela da *Web* Invisível. A visualização é geralmente feita por meio de um nome de usuário e senha fornecidos para o assinante, o que lhe garante o direito de ter acesso à informação proprietária.

A *web* verdadeiramente invisível

Pode ser caracterizada por quatro motivos, de acordo com Sherman e Price (2001), embora admitam que os mecanismos sempre estão desenvolvendo seus algoritmos e adaptando métodos para indexar novos tipos de formatos, o que torna essa caracterização fluida. Seguem os quatro motivos que caracterizam a *Web* Invisível.

1) formatos de arquivos como o PDF, *Postscript, Flash, Shocwave*, programas executáveis e arquivos comprimidos;

2) política de exclusão dos mecanismos, uma vez que alguns arquivos podem ser indexados, mas não o são, como os formatos PDF*;*

3) páginas dinâmicas que são geradas mediante solicitação ou consultas;

4) informações armazenadas em banco de dados.

Como o livro "*The Invisible Web*" foi publicado em 2001, os arquivos em formato PDF não eram indexados pela falta de estrutura de metadados nos documentos armazenados nas *Intranets*, embora, à época, o *Google* já o fizesse.

Especialmente as imagens e vídeos com pouco ou nenhum texto constituem outro tipo de linguagem para a *Web* Invisível. Eles podem ser incluídos (uma década depois já são), entretanto, por fornecerem pouca pista sobre o seu assunto, os mecanismos híbridos trazem à superfície resultados com problemas intersemióticos, isto é, de tradução, embora os desenvolvedores estejam trabalhando para superar essas limitações.

Dark web: o continente (verdadeiramente) escuro do ciberespaço

Outra forma de invisibilidade foi criada por um projeto ambicioso, como tese, em 2000, de autoria de Ian Clarke, então estudante da *Edinburgh University*, cujo resultado foi a criação do programa *FreeNet* (Becket, 2009).

O *FreeNet* foi criado pensado na liberdade de expressão e de conteúdo, como o protótipo perfeito de informação livre e sem restrições - principalmente judi-

ciais -, para seus usuários. Um usuário do *FreeNet* compartilha, ao participar da rede, uma parcela do seu disco rígido para armazenar informações criptografadas que ele mesmo jamais saberá do que se trata. Basicamente, o *FreeNet* é uma *Internet* paralela dentro da própria *Internet*, para usuários que querem privacidade sem rastreabilidade.

Também chamada *Dark Net, Web* Invisível e espaço de endereço escuro (embora não sejam exatamente sinônimos), essas metáforas servem para ilustrar e reforçar o caráter realmente invisível dessa modalidade da *Web* e significam, de certo modo, "[...] para além dos limites da vida da maioria das pessoas *online* [...] ignorada pela mídia e bem compreendida por apenas alguns cientistas da computação" (Beckett, 2009, p.3).

Iniciativa semelhante ocorreu com a criação do programa *The Onion Router* (Thor), um projeto voluntário para aqueles que procuram tráfego de informação anônima na *Internet* (Beckett, 2009). O desenvolvimento inicial do Thor era para o Laboratório de Pesquisa Naval Americano, para proteger a comunicação governamental. Hoje, o Thor pode ser utilizado por qualquer pessoa, embora essa liberdade tenha causado problemas legais, como aponta Beckett (2009).

Na prática, como funciona o Thor? Com seu uso, o roteamento de pacotes é randômico e a informação é encriptografada, ou seja, "perde-se" a identidade do solicitante.

Através do Thor, surgiu uma iniciativa de construção de *sites* utilizando o sufixo *onion*. Todo *site* que possui tal sufixo é inacessível e ilegível a qualquer navegador *Web* normal, sendo exclusivo dos usuários da rede Thor. Os motivos de permanecerem praticamente na total invisibilidade, na maioria das vezes, referem-se ao fato de seu conteúdo ser judicialmente ilegal.

Quadro 2. As dobras semânticas da *Web* Visível/Invisível.

Significado	Conceito	Conceito
Parte da *Web* Visível, ou seja, páginas que podem ser somadas ao banco de dados dos buscadores.	*Web* Visível Sherman e Price (2001)	*Web* de superfície Bergman (2001)
Páginas de textos, arquivos (muitas vezes de alta qualidade e com autoridade informacional) disponíveis na *Web*, os quais os motores de buscas não podem, devido a limitações técnicas, ou não querem, por escolha deliberada, adicionar aos índices de páginas *Web*.	*Web* Invisível Sherman e Price (2001)	*Web* Profunda Araújo (2001) Bergman (2001)
A *Web* Opaca consiste em *sites* que misturam arquivos e mídias, dentre os quais alguns são facilmente indexáveis e outros são incompreensíveis aos rastreadores. A profundidade, a frequência do rastreador e as páginas desconectadas (URL) podem ser motivos da opacidade de páginas na *Web*.	*Web* Opaca Sherman e Price (2001)	*Web* Oculta Araújo (2001)
A *Web* Privada consiste em páginas deliberadamente excluídas dos mecanismos pelo mantenedor (protegidas por *password*, *noindex* ou *robots.txt)*.	*Web* Privada Sherman e Price (2001)	
A *Web* Proprietária diz respeito ao conteúdo indexável, mas restrito por ser propriedade de seus mantenedores (instituições e órgãos, entre outros), acessível mediante registro, em muitos casos gratuitos, assinatura e/ou pagamento de taxas.	*Web* Proprietária Sherman e Price (2001)	
Algo que aparentemente está completamente inacessível, mas, mediante o uso de uma ferramenta, é possível localizar. Melhor seria, portanto, dizer que existe significativa parte da *Web* Oculta para os motores de busca mais populares.	*Web* Oculta Araújo (2001)	*Web* Profunda Bergman (2001) *Web* Opaca Sherman e Price (2001)
Rede global de usuários e computadores que operam à margem da visibilidade e das agências fiscalizadoras, com conteúdos intencionalmente escondidos e protocolos de comunicação inacessíveis para um sistema sem configuração correta.	*Dark Web* (Sem autoria determinada)	*Web* Invisível, espaço de endereço escuro, espaço de endereço sujo Beckett (2009)
A *Dark Net* é o conjunto de redes e tecnologias utilizadas para compartilhar conteúdo digital, como *peer-to-peer* de compartilhamento de arquivos, CD e DVD. A *Dark Net* não é uma rede independente, mas uma camada de aplicação e protocolo montados em redes físicas já existentes.	*Dark Net* Biddle *et al.* (2002)	

A *Dark Web* ilustra bem a tensão entre a privacidade e a publicidade; a liberdade de expressão e até valores maniqueísta do bem e do mal, arquétipos humanos ressignificados ou virtualizados no ciberespaço. Embora o *Freenet* tenha sido pensado para uma *Dark Net*, ou seja, rede para compartilhamento de conteúdos e arquivos livres na *Web* (Biddle *et al*., 2002) seu uso tem sido feito, em grande parte, por criminosos, para a pedofilia, tráfico e satanismos.

Para efeito de síntese, o Quadro 2 apresenta uma comparação entre os conceitos da *Web* Visível/Invisível, de acordo com os autores (indicados no quadro), para estabelecer as relações entre eles.

Descobrindo a *web* invisível: mecanismos de busca especializados

Se a *Web* é a dobra semiótica do ciberespaço, este, por sua vez, apresenta máquinas dentro de máquinas. Assim, os mecanismos de busca são as redobras, trazendo à visibilidade a *Web* Invisível.

Esses buscadores da *Web* Invisível são específicos e acessam uma variedade de interfaces (Sherman; Price, 2001). A Figura 3 apresenta alguns mecanismos de busca da *Web* Invisível e suas principais características de funcionamento.

Infomine: é um buscador desenvolvido por bibliotecários da Universidade da Califórnia. Indexa livros e periódicos eletrônicos, boletins, listas de discussões, catálogos de bibliotecas e diretórios de pesquisadores, entre outros tipos de informações similares. A página inicial do buscador aceita pesquisas livres e a possibilidade de percorrer por áreas do conhecimento (ou diretórios) <http://infomine.ucr.edu/>.

Internet Archive: seu propósito é ser uma grande biblioteca do ciberespaço para acesso de pesquisadores, historiadores e o público interessado em seu conteúdo. O *Internet Archive* faz a indexação de páginas antigas de *sites* que não mais existem ou foram atualizados. O projeto teve início em 1996, em São Francisco, e ilustra o interessante estatuto da memória no ciberespaço, em constante modificação <http://www.archive.org/>.

Hakia: os mantenedores do *Hakia* o denominam semântico, por ser um buscador que procura resolver os problemas morfológicos da língua. Oferece para a compra dois outros mecanismos "otimizados" para negócios e informações aeroespaciais, ao que tudo indica, relacionando termos polissêmicos às áreas especializadas em questão. O interessante é que seus resultados de busca são separados por *Deep Web*, *Surface Web* e Regular *Web* <http://www.hakia.com/>.

DeepDyve: a busca nos índices do *DeepDyve* é livre, contudo o acesso aos documentos recuperados faz-se mediante pagamento. A associação a esse buscador tem um período bastante limitado, até 14 dias. Indexa qualquer tipo de informação textual, inclusive grande quantidade de informação também visível para os mecanismos tradicionais <http://www.deepdyve.com/>.

Complete Planet: desenvolvido pelos mantenedores *site BrightPlanet* <www.brightplanet.com>, de Michael Bergman, permite a busca de arquivos invisíveis de todos os tipos, seja por busca simples, avançada ou diretórios.

Biznar: a empresa criadora do *BizNar* (*Deep WebTechnologies*) também possui mais buscadores com o intuito de indexar outros tipos de informação não ligadas ao mundo dos negócios. O que faz do *BizNar* muito útil são suas relações semânticas que eliminam boa parte da polissemia dos buscadores tradicionais, ligando qualquer palavra-chave aos negócios. Além disso, quando o mecanismo finaliza a busca, ele mostra categorias dentro do mundo dos negócios em que o tópico pesquisado mais apareceu, os *Result Topics*, como, por exemplo,

Figura 3. Mecanismos de busca da *Web* Invisível.
Fonte: Elaborada pelos autores.

Marketing, Publicidade & Propaganda etc. <http://biznar.com/biznar/search.html>.

Family Search: é uma das maiores e mais completas bases genealógicas disponíveis no ciberespaço. O *Family Search* forma seus índices com censos de várias épocas, listas telefônicas e até mesmo listas de obituários de todo o mundo. Traz dados de nascimento, morte, residência, telefone, data de casamento, filiação e até mesmo o nome do navio em que a pessoa imigrou no caso de não ser nativo de determinada região/país <https://www.familysearch.org/>.

Metabuscadores da *Web* Invisível: (ou metamotores) são mecanismos que utilizam os índices de vários buscadores para responder uma *query*. Essas ferramentas não possuem nenhuma base de dados, utilizando exclusivamente dados de outras ferramentas de busca (Cendón, 2001). No caso de metabuscadores da *Web* Invisível, o funcionamento é basicamente o mesmo: uma interface tratará de buscar nos índices de buscadores de conteúdo invisível.

Turbo10: agrega uma variedade de fontes e o acesso é exclusivo para assinantes. O Turbo10 prefere usar o termo *deep net* a outros para designar o conteúdo invisível, pois, segundo Hamilton (2003, *online*, tradução nossa[4]):

> [...] o Turbo10, no entanto, prefere usar o termo 'Deep Net' porque algumas dessas fontes de informação não são baseadas na *Web* (por exemplo, redes par-a-par) e os conteúdos dessas bases não estão escondidos ou invisíveis para os metamotores de busca. O desafio para um metamotor de busca comercial são, primeiro, conectar-se a essas fontes da *Deep Net*; segundo, selecionar o que é mais relevante; terceiro, retornar resultados relevantes o mais rápido possível.

Considerações Finais

Muita coisa já mudou desde que os primeiros artigos a respeito da *Web* Invisível foram escritos e parcela de informação invisível tornou-se visível, novos métodos de invisibilidade, como o *FreeNet* e *Onion*, foram criados. Na tona dessas discussões, atualmente, está o *site WikiLeaks* e suas "transgressões éticas", pauta de reflexões a respeito do aspecto público e privado das informações que circulam na *Web* Invisível e as complexidades dos objetos virtuais e simbólicos da sociedade contemporânea.

Boa parte das dificuldades, contudo, ainda é ligada à forma de indexar o conteúdo não verbal e às questões legais. Muitas informações, que na década passada eram invisíveis, já foram implementadas nos buscadores tradicionais, o que demonstra que a evolução tecnológica é uma forma de trazer maior quantidade de conteúdo invisível para o campo da visibilidade. É o caso das informações a respeito de temperatura (fornecidas mediante a estratégia temperatura + local em buscadores como *Yahoo!* e *Google*) e informações geoespaciais ou geopolíticas de locais já cartografados (*Google Maps*, *DuckDuckGo*).

Os recursos da chamada *Web* 2.0 também ajudam na procura de informações invisíveis. Um sujeito pesquisador pode utilizar serviços de perguntas e respostas para buscar questões já formuladas idênticas às suas e encontrar o que procura. Exemplos desses tipos são as redes sociais e o *Yahoo Answers*, este último, em especial, com várias perguntas do tipo "*como eu acho*", "*como encontro*".

Um pouco de familiaridade com os mecanismos tradicionais também pode ser um método de busca, *query* como "site: <www.site_que_quero_encontrar_algo palavra-chave>" buscará nos índices apenas do *site* especificado na maioria dos buscadores, o que é muito útil em *sites* que não fornecem campos de busca.

A evolução desse tipo de estratégia é muito estimulante e muitos mecanismos fornecem uma interface gráfica para esse tipo de busca *custom search*, podendo ser implementados dentro do próprio *site* pelo *webmaster*.

Enfim, os mecanismos de busca são considerados o "ponto" dobra no ciberespaço, máquina dentro de máquina, desdobrando a *Web* Invisível para a Visível, a localização de uma na outra, em um *continuum* semiótico que é o ciberespaço.

[4] *Turbo10, however, prefers to use the term 'Deep Net' because some of these information sources are not web-based (e.g., peer to peer networks) and the contents of these databases are not hidden or invisible to metasearch engines. The challenges for a commercial metasearch engine are, first, to connect to these Deep Net sources, second, to select the most relevant, and third, to return relevant results as fast as possible.*

Referências

ARAÚJO, J.P. *Invisível, oculta ou profunda?*: a web que poucas ferramentas enxergam. 2001. Disponível em: <http://www.comunicar.pro.br/artigos/weboculta.htm>. Acesso em: 21 jun. 2012.

BECKETT, A. *The dark side of the internet*. 2009. Available from: <http://www.guardian.co.uk/technology/2009/nov/26/dark-side-internet-freenet>. Cited: 21 Dec. 2011.

BERGMAN, M.K. White paper: the deep we surfacing hidden value. *Journal of Eletronic Publishing,* v.7, n.1, 2001. Available from: <http://dx.doi.org/10.3998/3336451.0007.104>. Cited: 23 Sept. 2011.

BIDDLE, P. *et al*. *The darknet and the future of content distribution.* 2002. Available from: <http://msl1.mit.edu/ESD10/docs/darknet5.pdf>. Cited: 16 July. 2012.

BRANSKI, R.M. Recuperação da informação na web. *Perspectivas em Ciência da Informação*, v.9, n.1, p.70-87, 2004.

CENDÓN, B.V. Ferramentas de busca na web. *Ciência da Informação,* v.30, n.1, p. 39-49, 2001.

DELEUZE, G. *A dobra*: Leibniz e o barroco. Campinas: Papirus, 1991.

DELEUZE, G.; GUATTARI, F. *Mil platôs*: capitalismo e esquizofrenia. São Paulo: Editora 34, 1995.

DELEUZE, G. *Proust e os signos*. 2.ed. Rio de Janeiro: Forense Universitária, 2010.

FORD, N.; MANSOURIAN, Y. The invisible web: na empirical study of cognitive invisibility. *Journal of Documentation*, v.62, n.5, p.584-596, 2006.

HAMILTON, N. *The mechanics of a deep net metasearch engine.* 2003. Available from: <http://www2003.org/cdrom/papers/poster/p170/poster/poster.html>. Cited: 21 Dec. 2011.

MACHADO, R. *Deleuze, a arte e a filosofia.* Rio de Janeiro: Jorge Zahar, 2009.

OLIVEIRA, L.A. Biontes, bióides e borgues. In: NOVAES, A. *O homem-máquina*: a ciência manipula o corpo. São Paulo: Companhia das Letras, 2003. p.139-174.

SANTAELLA, L. *As matrizes da linguagem e pensamento*: sonora, visual e verbal. São Paulo: FAPESP, 2005.

SHERMAN, C.; PRICE, G. *The invisible web*: uncovering information sources: search engines can't see. Medford: Cyberage Books, 2001.