BJIS Brazilian Journal of Information Science

A INDEXAÇÃO NA INTERNET

Isidoro Gil-Leiva
Facultad de Comunicación y Documentación
Universidad de Murcia
isgil@um.es

RESUMO

Apresenta-se uma análise da presença da indexação na Internet sustentada na idéia de que boa parte dos pilares básicos nos quais se assentam esta rede de computadores está impregnada em maior ou menor medida pela indexação. Os pilares são os metadados, os buscadores, os usuários e o posicionamento *web*. Cada um destes elementos é revisado para comprovar a abordagem que realizam ao qual denominamos **Universo da Indexação *Web***.

Palavras-Chave: Indexação; Internet; Metadados; Posicionamento *Web*; Usuários; Buscadores; Motores de Busca; Universo da Indexação *Web*.

INTRODUÇÃO

Antes da aparição da Internet, na maioria das ocasiões, os dados, a informação e os trâmites estavam distantes e descentralizados. Para realizar um trâmite administrativo era necessário ir à administração, para ler ou comprar um livro deslocar-se até a biblioteca ou livraria, para contemplar determinadas obras de museu dirigir-se a um, ou para comprar um carro encaminhar-se até uma concessionária e, assim, sucessivamente. Com a generalização da Internet quase todos os ramos de atividades do ser humano foram florescendo na rede e, consequentemente, a competência do mundo físico nos serviços, comércio ou cultura foi transladado ao âmbito digital. No entanto, conforme se incrementou os conteúdos na *Web* foi necessário o desenvolvimento de pontes, para conectar as pessoas com a massa de informação disponível na rede, com a finalidade de acessar de maneira rápida e eficaz este novo ambiente. Essas passarelas são os denominados motores de busca ou pesquisa.

Nessa nova realidade criou-se um espaço que denominamos *Universo da Indexação Web ou Ambiente de Indexação Web* que está conformado por quatro

âmbitos distintos, mas tremendamente interrelacionados. Por um lado, as linguagens de marcações e codificações normalizadas que facilitam a organização e difusão da informação pela *Web*; por outro, os proprietários da *Web* (empresas, instituições ou particulares) que desejam que seus conteúdos tenham a máxima visibilidade, isto é, uma boa posição nas pesquisas, por se tratar de melhoria de serviços, prestígio ou rentabilidade, entre outras variáveis; no terceiro espaço, situam-se os motores de busca que utilizam algoritmos complexos para oferecer um *ranking* da informação encontrada para satisfazer aos clientes; e no quarto âmbito, os usuários dos motores de busca, também, empregam táticas para maximizar o esforço e tempo empregado no uso dos motores.

Este quádruplo que se constitui no *Universo da Indexação* está impregnado pela indexação, e chegou-se a esse *Universo da Indexação* pela extensão progressiva, tanto de conhecimentos e práticas próprios dos indexadores como dos profissionais da informação e da documentação em geral, com vistas à popularização da Internet. A generalização desses conceitos e práticas é uma realidade e facilmente identificável em numerosos exemplos, como o aparecido em um jornal espanhol de caráter geral e tiragem nacional, no qual em cinco colunas e com uma extensão de meia página se lia[1]:

> […] Nossa língua envolve muitos países, mas na *Web* constitui-se um espaço único acessível aos pesquisadores. O Google, por exemplo, não indexa todas as páginas da *Web* em espanhol [...] Nenhum pesquisador reúne todas as páginas em uma língua: indexam somente a parte mais importante da sua *Web* [...] O segundo requisito é que as páginas facilitam o trabalho dos motores de busca, permitindo que seus conteúdos sejam acessíveis sem barreiras para qualquer pesquisador (em vez de afogar seus dados em animações). Assim, hoje poderá indexá-los no Yahoo ou no Google e, amanhã, quem sabe?

Nesse mesmo jornal, alguns meses depois, em outra seção leu-se:

> Para executar este trabalho se utilizam diferentes processos e metodologias baseadas em indexação, relevância e popularidade. A indexação é o conjunto de ações a realizar sobre a página *web*, para que o pesquisador possa acessar toda a informação que contém.

Nesse artigo da imprensa, sem ir além, está se falando do quádruplo *Universo da Indexação* que ressaltamos anteriormente (conceito de indexação, posicionamento *web*, pesquisadores e metadados).

BJIS Brazilian Journal of Information Science

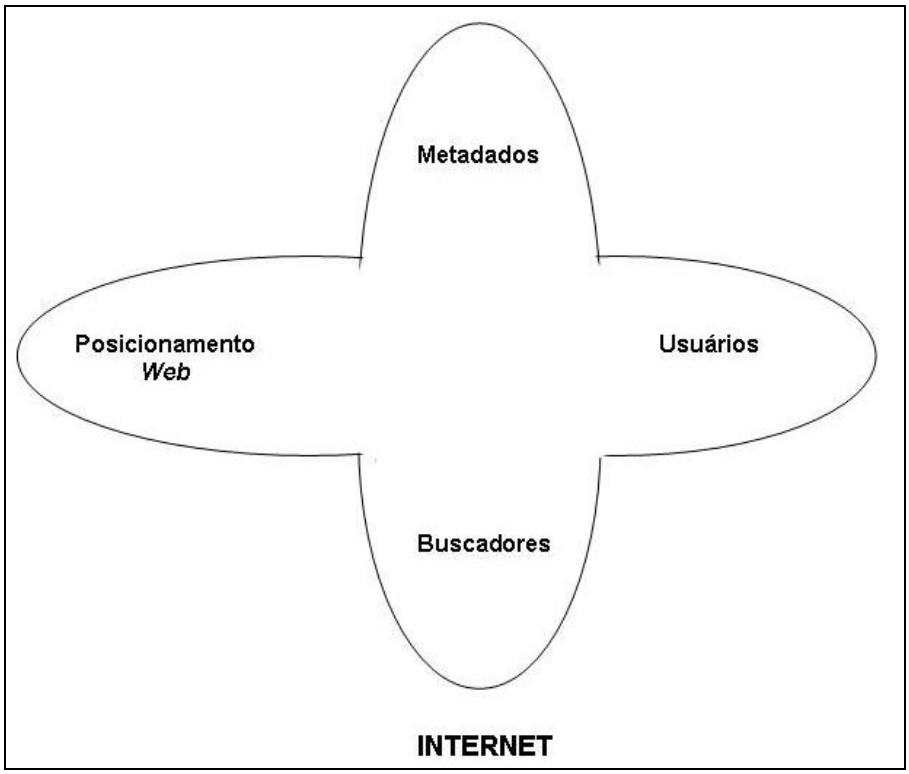

Figura 1: Universo da Indexação *Web*.
Fonte: Elaborado pelo autor.

2 UNIVERSO DA INDEXAÇÃO EM INTERNET

Dando continuidade, nos deteremos em cada um dos componentes do *Universo da Indexação Web* que acabamos de mencionar.

2.1 Metadados

Os metadados estão destinados a ordenar e descrever a informação contida em um documento entendido como objeto, de tal forma que se erigem como reveladores, tanto da descrição formal, quanto da análise de conteúdo, visando melhorar o acesso a esses objetos de informação da rede. Não são mais que estruturas de organização da informação, legíveis por máquina, cuja finalidade é tornar úteis os dados, de diferentes formas, segundo as necessidades concretas de cada serviço de informação digital e segundo a aplicação que lhes são outorgadas[2].

Existe um amplo catálogo de tipologias de metadados propostas por vários autores, mas uma que se ajusta perfeitamente aos nossos interesses é a seguinte:

> Metadados, independentes do conteúdo que reúnem informação sobre a localização do documento, sua data de criação, modificação etc. [...] Metadados dependentes do conteúdo, que aglutinam dados sobre a representação e estrutura da porção de informação que descrevem. Por sua vez, esses metadados se dividem em metadados baseados no conteúdo direto, como, por exemplo, os índices de um documento em texto completo ou a cor e a forma de uma imagem digital, e metadados descritivos do conteúdo como, por exemplo, descritores e identificadores, isto é, os metadados que contêm a descrição de um documento sem utilizar expressamente seu conteúdo[3].

Os conjuntos de metadados empregam marcas ou etiquetas que são pares iguais de palavras ou acrônimos com um alto valor semântico e nemotécnico circuladas por ângulos, entre as que se localiza a informação estruturada. A primeira etiqueta indica que ai começa uma porção de informação determinada e a segunda etiqueta que possui uma barra oblíqua, indica o fim. Essas etiquetas são legíveis facilmente, tanto por programas informáticos como por humanos e têm a facilidade de apresentar, estruturar e intercambiar informação entre computadores.

```
<nome> Antonio Gil Cuenca </nome>
<lugar de nascimento> Águilas </lugar de nascimento>
<cidade> Murcia </cidade>
<endereço> Rua San Vicence, 7 </endereço>
<parentesco> bisavô paterno </parentesco>
```

A partir do conjunto de regras Standard Generalizad Markup Language, mais conhecido como SGML e convertidas em norma ISO, em 1986, surgiram as linguagens de marcação que servem para codificar um documento mediante um conjunto de etiquetas. Após a SGML surgiu a linguagem eXtensible Markup Language (XML), que é mais extenso e convertido quase em um padrão. Destes dois modelos proliferaram linguagens de marcação, a maioria para âmbitos específicos, que servem tanto para esquematizar e distribuir informação de qualquer tipo (linguagem HTML o XML, por exemplo) como para disciplinas ou áreas específicas (EAD e EAC para a Arquivística; CIMI para Museus; MPEG-& para

conteúdo multimídia; OWL para compartilhar ontologias na *Web*; ID3 para fichários de áudio MP3; MDL para conteúdo audiovisual; MCM para medicina etc.). A estas linguagens de marcação, há de se acrescentar outras codificações normalizadas criadas em algum caso anteriormente, mas que compartilham a filosofia do intercâmbio de informação como, por exemplo, MARC, ISAD(g) ou MoReq.

Repassamos agora, algumas linguagens de marcação e codificações normalizadas, enfatizando os metadados que podem abrigar a indexação.

2.1.1 HTML

A linguagem de marcação HyperText Markup Language (HTML) utiliza a seção de cabeçalho para transmitir ao servidor *Web* a informação sobre o documento. Toda a informação que se proporciona no cabeçalho está compreendida entre a etiqueta <head> e a etiqueta </head>. Existe uma série de etiquetas reservadas especificamente para o cabeçalho como a de <title> e </title> ou a etiqueta META *keywords*, que serve para inscrever aí palavras-chave ou frases significativas (para a indexação) e, indicar assim, aos motores de busca o conteúdo exato da página *Web* (para a recuperação).

Código Fonte de HTML com as Etiquetas META de Palavras-Chave
<head> <meta name="generator" content="HTML Tidy, see www.w3.org" /> <title>UKOLN</title> **<meta name="keywords" content=** "national centre, digital information management, cultural heritage, library, awareness, research, information services, public library networking, bibliographic management, distributed systems, metadata, resource discovery, conferences, lectures, workshops" **/>** … …

A etiqueta <keywords> foi pensada para armazenar palavras ou frases relevantes visando à recuperação. Sua utilização por parte dos *webmaster* é muito baixa e desigual ao teor de alguns estudos realizados[4], talvez por desconhecimento, descuido mesmo intencional devido ao abuso excessivo que se fez delas em muitas ocasiões.

2.1.2 Dublin Core

A *Dublin Core Metatadata Iniciative* é um grupo de trabalho constituído por bibliotecários, pesquisadores de bibliotecas digitais e provedores de informação que começou a funcionar em Dublin (Ohio), em 1995, com o propósito de proporcionar recomendações sobre a descrição de recursos de informação e de seu intercâmbio. *Dublin Core* proporcionou quinze metadados para a descrição de um recurso informacional. Para o conteúdo (Título, Assunto, Descrição, Fonte, Língua, Relação e Cobertura), para a propriedade intelectual (Autor, Editor, Colaborador, Direitos) e para o formato (Data, Tipo, Formato, Identificação). A versão 1.1 do conjunto de elementos de metadados *Dublin Core* passou, em 2003, a ser uma norma internacional sob o número ISO 15836:2003 (UNE-ISO 15836:2007).

As etiquetas relacionadas diretamente com a indexação são: Assunto, Palavras-Chave, Produtor e Data. O *Dublin Core* define essas etiquetas da seguinte maneira:

> Name: subject
> Label: **Subject and Keywords**
> Definition: The topic of the content of the resource.
> Comment: Typically, a Subject will be expressed as keywords, key phrases or classification codes that describe a topic of the resource. Recommended best practice is to select a value from a controlled vocabulary or formal classification scheme.
>
> Name: creator
> Label: **Creator**
> Definition: An entity primarily responsible for making the content of the resource.
> Comment: Examples of Creator include a person, an organization, or a service. Typically, the name of a Creator should be used to indicate the entity.
>
> Name: date
> Label: **Date**
> Definition: A date of an event in the lifecycle of the resource.
> Comment: Typically, Date will be associated with the creation or availability of the resource. Recommended best practice for encoding the date value is defined in a profile of ISO 8601 and includes (among others) dates of the form YYYY-MM-DD.

Etiqueta Subject de Dublin Core para XML
<?xml version="1.0"?> <metadata <dc:title>Universitat Politécnica de Valencia</dc:title> <dc:creator>Universidad Politecnica de Valencia</dc:creator> <**dc:subject**> UNIVERSIDAD INVESTIGACION DOCENCIA CENTROS DEPARTAMENTOS ESTUDIOS ALUMNADO PROFESORES ASIGNATURAS PROYECTOS <**/dc:subject**> <dc:type> Text </dc:type> <dc:format> text/html </dc:format> <dc:format> 1447 bytes </dc:format> <dc:identifier> http://www.upv.es </dc:identifier> … …

2.1.3 Encoded Archival Description (EAD)

Iniciou-se o trabalho com esta linguagem de marcação em princípios da década de 1990, na Universidade de Califórnia, Berkeley, para criar uma estrutura normalizada de dados que propiciasse o intercâmbio e acesso aos instrumentos de descrição usados pelos arquivos. A Descrição Arquivística Codificada (EAD) se compõe de três elementos principais: 'Cabeçalho EAD' com a etiqueta <eadheader>, 'Preliminares' e a etiqueta <frontmatter> e, em terceiro lugar, o elemento 'Descrição de Arquivo' com a etiqueta <archdesc>. Desta etiqueta <archdesc> se desdobram as demais etiquetas que permitem representar os instrumentos de descrição dos documentos de um arquivo.

Um dos subelementos <archdesc> se denomina 'Encabeçamento de Acesso Autorizado' que tem por etiqueta <controlaccess>. Pois bem, neste subelemento repousa o resultado da indexação agrupado em várias etiquetas como:

<subject>	assunto
<corpname>	nome da instituição
<persname>	nome da pessoa
<unitdate>	data que aparece na unidade que se descreve

BJIS, v.1, n.2, p.47-68, jul./dez. 2007. Disponível em: <http://www.bjis.unesp.br/pt/>. ISSN: 1981-1640

2.1.4 Encoded Archival Context (EAC)

A iniciativa de criar um conjunto de etiquetas para representar os contextos arquivísticos surgiu em 2001 entre o Projeto LEAF (Linking and Exploring Authority Files) e o Programa de Tecnologias para a Sociedade da Informação da Comissão Européia.

O Contexto Arquivístico Codificado (EAC) é um conjunto de etiquetas que permite que todo documento contenha dois elementos obrigatórios, o cabeçalho <eacheader> e a descrição do contexto <condesc>. A etiqueta <eacheader> contém dados sobre o controle da descrição do autor e o contexto da descrição. E a descrição do contexto <condesc> acolhe a descrição do autor. Por outro lado, tanto <eacheader> como <condesc> contêm elementos específicos nos quais se abriga o resultado da indexação.

<identity> Identidade, elemento obrigatório que agrupa elementos para proporcionar encabeçamentos de nomes, tanto autorizados como catálogos alternativos para identificar a entidade.

<corphead> Encabeçamento de órgão corporativo, que envolve, por exemplo, nomes de associações, instituições, empresas, exposições, expedições, férias etc.

<date> Data, identifica quaisquer datas que mereçam codificações.

<existdate> Data de existência, para as datas de existência da entidade que se está descrevendo, como estabelecimento e dissolução de órgãos e data de nascimento ou morte de pessoas.

<auth> Autoridade que especifica um vocabulário controlado ou autorizado, utilizados para compor o registro de autoridade.

<famname> Nome de família, destinado a um grupo de pessoas intimamente relacionadas por laços de sangue ou pessoas que constituem uma casa.

<pershead> Encabeçamento de nome pessoal, para identificar uma pessoa com certeza. Consta de elementos tais como sobrenome, nome próprio, patronímico o topônimo.

<place> Lugar, de caráter natural, uma jurisdição política como Montes Apalaches; Baltimore, Md.; Chinatown, São Francisco; ou Kew Gardens, Inglaterra.

<subject> Matéria, que identifica uma matéria associada ou protegida pelos materiais descritos em uma instância EAD. Assim mesmo, para esta etiqueta se

recomenda o uso de vocabulários controlados para facilitar o acesso às matérias dentro e entre sistemas de instrumentos de descrição. <persname> Nome de pessoa, para os nomes pessoais. <corpname> Nome corporativo, para os nomes corporativos. <geogname> Nome geográfico, para os nomes geográficos.

2.1.5 Consortium for the Interchange of Museum Information (CIMI)

Trata-se de um consórcio de instituições e organizações para conservar e difundir a herança cultural e museus principalmente. Nesse conjunto de metadados, a indexação se ordena nas etiquetas listadas abaixo:

<date>	data.
<dateRange>	duas datas para delimitar um período.
<geo>	nome comum geográfico como "vale", "montanha", etc.
<geogName>	característica geográfica ("Valle de Leiva", "Monte Sinai" etc.)
<keywords>	lista de palavras-chave ou frases que identificam o tema ou natureza do texto.
<orgName>	nome de uma organização.
<persName>	nome próprio de pessoas.
<placeName>	nome de um lugar.
<region>	região geopolítica.
<country>	nome de um país, nação, colônia etc.
<textClass>	natureza ou tema de um texto em torno de uma classificação ou um catálogo.

2.1.6 Text Encoding Initiative (TEI)

TEI é uma norma interdisciplinar e internacional que ajuda bibliotecas, museus, editores etc., a representar toda classe de textos humanísticos para a pesquisa e ensino, utilizando para isso uma estrutura de etiquetas. Algumas etiquetas pensadas para conter o produto da indexação são as seguintes:

<keywords>	Lista de palavras ou frases que identifiquem o tema do texto
<person>	Nomes de indivíduos dentro dos manuscritos
<institución>	Nomes de instituições
<origDate>	Datas que aparecem dentro dos manuscritos

<origPlace>	Lugares denominados nos manuscritos
<country>	Países
<region>	Regiões

Uma vez analisadas as etiquetas utilizadas por algumas linguagens de marcação para conter o resultado da indexação, revisemos agora conhecidas codificações normalizadas que perseguem o mesmo propósito.

2.1.7 Machine Readable Cataloging (MARC)

A Biblioteca do Congresso dos Estados Unidos desenvolveu o formato LC MARC na década de 1960, como um conjunto de sinalizadores que combinam números, letras e símbolos para acrescentá-los aos registros catalográficos. Dessa maneira, cada porção de informação bibliográfica precedida pelos sinalizadores (como por exemplo, 300 ; 1# ; $a ; $c) pode ser lida pelos computadores. Abaixo, apresentam-se as etiquetas MARC que contêm epígrafes ou cabeçalhos de assunto, de nomes e de datas:

Assuntos	**Nomes**	**Datas**
600 - Nomes pessoais 610 - Entidades 611 - Títulos de conferencias 648 - Termos cronológicos 650 - Assuntos 651 - Locais geográficos 653 - Termos não controlados 654 - Termos temáticos facetados	100 e 700 - Nome pessoal 110 e 710 - Nome institucional	005 033 260$c 362

033 2	1951-2007
600	Coudenhove-Kalergy, Richard
600	Briand, Aristide
610 24	$aUnião Européia$x-Instituições
610 24	$aUnião Européia$x-História
650	$aDireito comunitário
700 1	$aMiguel, Mario de

2.1.8 ISAD(g)

A Norma Internacional Geral de Descrição Arquivística (International Standard Archival Description(g)) foi elaborada pelo Comitê de Normas de Descrição, do Conselho Internacional de Arquivos, composto por arquivistas de diferentes nacionalidades. Como se indica na introdução da própria norma, constitui-se em um guia geral para a elaboração de descrições arquivísticas que identifiquem e expliquem o conteúdo e o contexto dos documentos de arquivo, com o objetivo de torná-los acessíveis e intercambiáveis.

A norma define vinte e seis elementos que se dividem entre as seguintes sete áreas: Identificação, Contexto, Conteúdo e Estrutura, Condições de Acesso e Uso, Documentação Associada, Notas e, por último, Controle da Descrição. Como se observa não se dedicou nenhum elemento especificamente voltado para a indexação. A falta de um elemento voltado à indexação, obriga os profissionais de arquivo, quando estão descrevendo documentos, a decidir onde colocar os dados relativos à indexação. Observa-se que alguns profissionais os colocam no elemento *6.1 Notas*. Se observarmos a norma, verificamos que esse elemento (6.1) "serve para consignar informação especial ou qualquer outra informação significativa não incluída em nenhum outro elemento da descrição." Outras instituições incluem a informação relacionada à indexação no elemento *3.1 Alcance e conteúdo*, destinado segundo a norma, a "dar uma visão de conjunto (por exemplo, períodos de tempo, âmbito geográfico) e realizar um resumo de conteúdo (por exemplo, tipos documentais, matéria principal, procedimentos administrativos) da descrição, apropriados ao nível de descrição". Embora esse elemento pareça mais adequado para incluir a indexação, não é o ideal incluir a indexação junto ao resumo, tipologias documentais ou procedimentos administrativos. Uma terceira via empregada pelos arquivistas é a inclusão de um aparte denominado *Pontos de Acesso*, e que vem depois do último elemento.

Evidentemente, esta indefinição da norma pode provocar que de um lado, os elementos *3.1 Alcance e Conteúdo* e *6.1 Notas* possam acabar abrigando informação muito heterogênea. Por outro lado, incluir a indexação sob *Pontos de Acesso* não parece o mais adequado, porque entre outras coisas, como é sabido, os

elementos *1.2 Título*, *2.1 Nome do Produtor* ou *4.3 Língua* também são pontos de acesso ao documento descrito. Portanto, seria desejável que na próxima revisão e atualização da norma ISAD(g) se incluísse um elemento específico para abrigar a indexação do documento descrito, posto que um dos objetivos principais desta norma seja o de tornar acessível os documentos de Arquivo e, sem dúvida, o resultado da indexação (assuntos, nomes próprios de pessoas, de coisas ou de lugares e datas ou períodos) é um meio insubstituível para facilitar as pesquisas e, finalmente, o acesso aos documentos.

Apesar da melhoria da norma, os arquivos estão efetuando um grande trabalho de descrição e indexação de documentos como se observa nos seguintes exemplos. Inclusive alguns arquivos estão sendo pioneiros no uso de catálogos para a indexação e a posterior recuperação dos documentos através da Internet, como é o caso do *Archivo Municipal de Arganda del Rey* (Madrid) o do Archivo del Reino de Valencia (Valencia).

Figura 2 - Indexação Associada a uma Fotografia Descrita Seguindo a ISAG(g) no Arquivo Municipal de Logroño.
Fonte: Arquivo Municipal. Fotografias. ESP AML FO nº 779.

1.2 Título: Inauguração das obras da nova estação de trem
…
…
6.1 Nota: Há margens brancas. A série que a formam são os nºs. de inventário de 717 a 786

Descritores Geográficos

Via o Termo:

Localidade: Logroño
Província: La Rioja
País: Espanha

Organismos e Entidades:
Onomásticos: Foto Payá (fotógrafo). Fernández Ladreda e Menéndez Valdés, José Maria (general e ministro de Obras Públicas). González Gallarza, Eduardo (general e ministro del Aire). Pernas Heredia, Julio (Prefeito de Logroño)
Matérias: Atos públicos e oficiais. Trem de Ferro
Outros: Estação de trem

Arquivo Municipal de Arganda del Rey:

Referência: ES012800148.AMAR/DD0000200006
…
…
Título: Carta executória de Felipe III a petição de Juan Batista de Granados contra Juna de Higuera pelo não pagamento das rendas de uma Loja de pescados e azeite.
…
…
Área de Conteúdo

Descritores:

Matérias:
Executória
Renda
Pescado
Azeite
Abundância

Pessoas:
Bautista de Granados, Juan
Muñoz, Sebastián
Higuera, Juan de la

Lugares:
Valladolid (MU)
…
…

BJIS, v.1, n.2, p.47-68, jul./dez. 2007. Disponível em: <http://www.bjis.unesp.br/pt/>. ISSN: 1981-1640

2.1.9 Moreq

No seio da União Européia e no contexto do intercâmbio de dados entre as administrações européias confeccionou-se em 2001 um Modelo de Requisitos (MoReq) para implantar-se um Sistema de Gestão de Documentos Eletrônicos de Arquivo (SGDEA). Neste modelo de requisitos se destinam metadados com o objetivo de abrigar a indexação:

> *12.4.3 Palavras-Chave Descritas*
> O SGDEA deve admitir a associação de termos incluídos em um vocabulário controlado como termos descritivos referentes ao assunto.
>
> *12.4.22 Nome Baseado em Palavras-Chave*
> Convém que as denominações de expedientes estejam baseadas em termos incluídos em um vocabulário controlado e em relações extraídas de um catálogo. Assim mesmo, é conveniente que permita a vinculação do catálogo ao quadro de classificação.
>
> *12.7.2 Assunto*

2.2 Posicionamento *Web*

A Search Engine Optimization (SEO), quer dizer 'Otimização de Motores de Busca' é um conjunto de técnicas aplicadas a uma página *web,* visando obter uma melhor posição nas listas oferecidas pelos motores de busca, quando uma determinada pesquisa é realizada. Nesse sentido, ressalta-se que se utiliza mais comumente como "posicionamento *web*". Desde o seu aparecimento em meados de 1990, se desenvolveu um mercado específico para trabalhar com isso, tanto empresas e profissionais SEO, quanto programas informáticos.

O posicionamento *web* logo despertou um grande interesse traduzido em numerosa literatura impressa, mas, sobretudo, profusão na própria Internet[5].

Para tentar conseguir um bom posicionamento *web*, embora não é claro que se consiga, porque cada motor de busca utiliza critérios diferentes na hora de oferecer resultados, é necessário um numeroso conjunto de técnicas. Essas táticas

são conhecidas como "fatores SEO" e agrupam tantos os fatores endógenos que atuam dentro da página *web* (otimização do título, do conteúdo, das etiquetas etc.), quanto os fatores exógenos (PageRank, texto nos *links*, *links* externos etc.). Na Tabela 1 se mostram várias classificações de fatores *SEO* ordenados por relevância[6]:

Tabela 1 - Fatores SEO Ordenados por Relevância

Top 10	Top 15
1. Etiqueta *title* 2. Texto nos *links* 3. Uso de palavras-chave no documento 4. O acesso do documento 5. *Links* internos 6. O principal assunto da página *web* 7. *Links* a páginas *web* externas 8. Popularidade na comunidade específica 9. Popularidade geral 10. Repetição excessiva de palavras-chave.	1. Etiqueta *title* 2. Palavras-chave usadas no documento 3. Estrutura de *links* internos 4. Conteúdo único 5. *Links* para páginas *web* externas *6.* Antigüidade do *WebSite* 7. A meta etiqueta Descrição 8. Palavras-chave na URL *9.* Etiquetas *title* e *Alt* 10. Etiquetas negrito, sublinhadas, H1 11. Densidade da página 12. Língua da página 13. Atualização do conteúdo 14. Tamanho da página 15. Validação W3C

Apresenta-se (Anexo 1) uma compilação de várias recomendações para se obter um bom posicionamento *web*. Em uma rápida análise dos fatores *SEO* verifica-se a presença de elementos muito próximos da teoria e da prática de indexação.

2.3 Buscadores

A recuperação da informação na Internet é possível por meio dos buscadores. Estes podem ser classificados em índices temáticos ou diretórios e em motores de busca. A Tabela 2 contém uma comparação dos diretórios e motores de busca.

Tabela 2 – Comparação Diretórios *Versus* Motores de Busca

	Recursos Utilizados	Representação do Conteúdo	Representação da Pesquisa	Apresentação dos Resultados
Diretórios	Realizam-no pessoas	Classificação manual	Implícita (navegação por categorias)	Páginas criadas antes da consulta. Pouco exaustivos, muito precisos.
Motores de Busca	Principalmente de forma automática por meio de robôs	Indexação automática	Explícita (palavras-chave, operadores etc.)	Páginas criadas dinamicamente em cada consulta. Muito exaustivos, pouco precisos.

Fonte: Martínez Méndez – 2000 - p.27.

Cada um dos motores de busca utiliza algoritmos[7] secretos para ordenar do mais relevante ao menos relevante os resultados devolvidos aos usuários. O algoritmo que gerou mais literatura nos últimos anos foi o desenhado e patenteado pelos criadores do Google, em 1998, denominado de 'PageRank'. Trata-se de um sistema complexo baseado nas redes de conexões existentes entre as páginas *web*. Embora a totalidade dos critérios que o sistema utiliza para calcular este dado seja reservada, parece ser considerável a freqüência de aparição das palavras, sua posição no texto, o número de conexões que se encaminham até uma página ou a importância da página que recebe e emite seu voto. Assim uma página *web* apontada vinte conexões, aparentemente possui menos interesse do que as que são apontadas mil conexões. Quando se faz uma pesquisa no Google ocuparão os primeiros lugares as páginas que têm um PageRank alto e, que também, coincidam com a temática da pesquisa.

2.4 Usuários

O uso da Internet está convertendo cada usuário em um *paradocumentalista* em potencial. Inconscientemente, os usuários dos buscadores assimilam a terminologia, os conceitos e as práticas que, até meados de 1990, eram quase exclusivas de profissionais da informação e documentação.

Os usuários que recorrem assiduamente à Internet para localizar informação (o último disco de seu cantor preferido, um livro da biblioteca municipal, um seguro

para a casa, a compra mensal no supermercado etc.) estão familiarizados com conhecimentos como:

- Para efetuar uma pesquisa é necessário escolher cuidadosamente as palavras-chave. E estas, quanto mais específicas melhor.
- Também é possível servir-se de frases e não somente de palavras simples.
- Alguns buscadores permitem colocar entre aspas o texto, de forma a se obter uma concordância exata de palavras compostas ou frases.
- Para restringir uma pesquisa há os operadores booleanos (and, or, not ou no).
- É possível procurar informação ou um dado em um campo ou etiqueta específica (Ex.: Palavra-Chave: vôos de baixo custo).
- Especificar uma data ou intervalos reduz o número de documentos recuperados.
- A navegação por lista de assuntos, uma classificação temática ou inclusive um catálogo para selecionar o assunto ou o termo desejado.

Por fim, popularizaram-se conceitos como *recuperação da informação, documentos relevantes, palavras-chave, assuntos e descritores, termo específico e termo geral, operadores booleanos, pesquisa em campos e etiquetas, campos e etiquetas índice, listas de palavras-chave, catálogos etc.*

BJIS, v.1, n.2, p.47-68, jul./dez. 2007. Disponível em: <http://www.bjis.unesp.br/pt/>. ISSN: 1981-1640

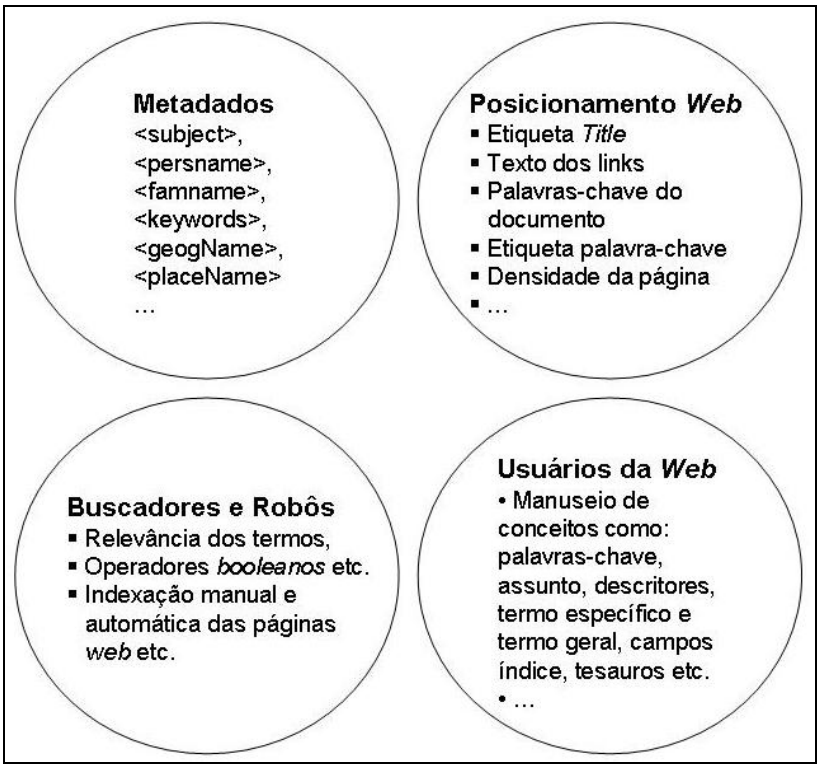

Figura 3: Universo da Indexação na Internet.
Fonte: Elaborado pelo autor.

CONCLUSÕES

A revisão de elementos importantes que alicerçam a rede Internet como, por exemplo, os metadados, os buscadores, os usuários e o posicionamento *web* nos permitiram constatar nossa hipótese de partida. Parte-se da idéia de que estes suportes básicos estão impregnados, em maior ou menor medida, pela indexação, o que conduz a criar um *Universo da Indexação na Web*, propiciado pela extensão progressiva de conceitos e práticas próprias dos indexadores como conseqüência da popularização e extensão de Internet.

REFERÊNCIAS

ALIMOHAMMADI, D. Meta-tag: a means to control the process of web indexing. **Online Information Review**, v.27, n.4, p.238-242, 2003.

CRAVEN, T. Variations in use of meta tag keywords by web pages in different languages. **Journal of Information Science**, v.30, n.3, p.268-279, 2004.

CRAVEN, T. Web authoring tools and meta tagging of page descriptions and keywords. **Online Information Review**, v.29, n.2, p.129-138, 2005.

GEORGE, D. The ABC of SEO: search engine optimization strategies. Morrisville, NC: Lulu Press, 2005.

GRAPPONE, J. Search engine optimization: an hour a day. San Francisco (CA): Sybex; Chichester: Wiley, 2006.

KENT, P. Search engine optimization for dummies. Indianapolis: Wiley, 2006.

MARCOS, M. C. et al. Evaluación del posicionamiento web en sistemas de información terminológicos online [en linea]. **Hipertext.net**, n.4, 2006. Disponible em: <http://www.hipertext.net>. Acesso em: 25 set. 2007.

MARTÍNEZ MÉNDEZ, F. J. **Propuesta y desarrollo de un modelo para la evaluación de la recuperación en Internet**. Murcia: Universidad de Murcia, 2000. (Tese de doutorado).

MÉNDEZ RODRÍGUEZ, E. **Metadatos y recuperación de información**: estándares, problemas y aplicabilidad en bibliotecas digitales. Gijón: Trea, 2002.

MERLO VEGA, J. A.; SORLI ROJO, A. El uso de metainformación en los webs de las bibliotecas españolas. In: FESABID. **Actas…** [S.l.p.]: Fesabid, 2000. p.155-164

SIROVICH, J.; DARIE, C. **Professional search engine optimization with PHP**: a developer's guide to SEO. Indianapolis: Wiley, 2007.

Prof. Dr. Isidoro Gil-Leiva
Facultad de Comunicación y Documentación
Universidad de Murcia
isgil@um.es

Artigo recebido: Maio, 2008

Artigo Aceito: Junho, 2008

Anexo 1: Recomendações para um Bom Posicionamento *Web*

Recomendações A	Recomendações B
1. Pesquisa as palavras-chave mais utilizadas para buscar páginas *webs* que ofereçam o mesmo que a própria página. Utiliza a ferramenta Wordtracker ou sugestão de palavras-chave do Google Adwords.	**1.** **Não há atalhos**: Não há uma forma rápida e fácil de obter bons resultados. Mas ao contrário, exige muito trabalho. Também é necessário ter paciência. Os resultados não chegam da noite para o dia.
2. Escolhe as palavras-chave para otimizar as páginas do *site*; não só para a página principal, mas também para as páginas internas. Tenta fazer com que cada página cubra de 2 a 4 palavras-chave específicas para o conteúdo da página.	**2.** **Escreve bem o conteúdo**: Está é provavelmente a coisa mais importante que deve ser feita, se quiser ser encontrado na *Web*. Gramaticamente correto, o suficientemente específico e sempre atualizado.
3. Inclui as palavras-chave de cada página nos títulos, cabeçalhos (H1, H2...) e o texto.	**3.** **Pensa em escrever corretamente (boa ortografia)**: Se escrever em inglês (no nosso caso português), observar as diferenças entre escrever em inglês americano e britânico (como as diferenças entre o português do Brasil e de Portugal).
4. Inclui as palavras-chave no texto dos *links* de navegação interna, de forma que as palavras-chave, para que otimizem uma página concretamente se encontrem em todos os *links* que apontam a referida página. Se utilizar imagens como botões de navegação, inclui as palavras-chave na etiqueta *ALT*.	**4.** **Escreve os títulos das páginas descritivos**: Fazer os títulos das páginas simples, mas descritivos e relevantes, de forma que seja mais fácil para os buscadores saber do que se trata cada página, e que a pessoa ao observar o resultado de busca, possa determinar rapidamente se a página *web* contém o que ela está buscando. Alguns argumentam que é um dos elementos mais importantes da página *web*. Não use o mesmo título para todos os documentos.
5. Assegurar que toda a página *web* é navegável por um navegador que tenha desativado '*javascript*' e '*flash*'. Caso contrário, implementar um sistema de navegação alternativo com etiquetas <noembed> ou <noscript> ou um mapa da página *web*.	**5.** **Usa cabeçalhos [headings] reais**: Usar os elementos [HTML] h1 a h6 para os cabeçalhos.
6. Obter todos os *links* possíveis da página *web* com outras páginas *web* e, se possível, assegurar que no texto dos *links* sejam incluídas as palavras-chave mais importantes. Esse é o conselho mais importante para se obter um bom posicionamento.	**6.** **Usa URLs amigáveis para os buscadores**: Evitar URLs gerados automaticamente e que usem o "query string" para que o servidor saiba que informação traz a base de dados. Os robôs de busca possivelmente tenham dificuldades com este tipo de URLs. Utilizar URLs amigáveis para os buscadores e legíveis para as pessoas.
7. Pesquise com as palavras-chaves estabelecidas na página *web* em diretórios como DMOZ, todos os diretórios generalistas que encontrar e todos os diretórios especializados na temática da página *web*.	
8. Identificar outras páginas *webs* similares e solicitar um *link* para a	

própria página. Buscar *links* que apontam para páginas *webs* concorrentes, utilizando a cadeia de busca "link:http://www.direccion.com". Pensar em novos métodos para obter mais *links* é muito importante. **9.** Enviar a página web para todos os buscadores que for possível. **10.** Repassar o que o Google considera *spam* para não correr o risco de que eliminem a base de dados.	**7. Faça com que te vinculem:** Não há maneira mais fácil nem sustentável para resolver isso exceto prover bons conteúdos. Os vínculos próximos são muito importantes para o SEO. **8. Usa uma marcação [HTML] válida, semântica, rápida e acessível**: Validar o HTML e evitar a marcação [HTML] de apresentação. Usar uma marcação [HTML] rápida e limpa o quanto for possível. Incrementar a proporção de conteúdo em HTML [no lugar de gráficos], isso fará com o *site* seja mais rápido e mais atrativo para os buscadores. **9. Agrega com cuidado ao *site* os buscadores**: Ainda que valorizado excessivamente, agregar o *site* aos diretórios e buscadores pode ser útil, especialmente se o *site* é novo e, ainda, não foi captado pelo Google ou outros buscadores. **10. Não engane os buscadores**: Não use camuflagem (*cloaking, link farms*), excesso de palavras-chave (*keyword stuffing*), texto alternativo com *spam* (*alt text spamming*) ou outros métodos enganosos. Funcionaram por um curto período, mas corre o risco de ser penalizado, e o que é pior, proibido nos buscadores. **11. Evita usar marcação (*Frames*)**: Podem causar problemas para a pessoa que acessar o *site* nos buscadores. **12. Tenha cuidado com a detecção de navegadores**: Se necessitar de algum tipo de detecção de navegadores, assegure que funciona quando um robô de um buscador (o qualquer outro agente) chegue, se os robôs não podem entrar o *site* não será encontrado. **13. Não perca tempo com os *meta tags***: A maioria dos buscadores já não valorizam o conteúdo dos *meta tags*, visto que são exaustivamente usados pelos *spammers*. As palavras-chave do <keywords> ajudam muito pouco, por isso não valem a pena.

Fonte: Recomendações A - <http://www.guiabuscadores.com/posicionamiento/posicionamiento-rapido.html>. Acesso em: 19 out. 2007.
Recomendações B - <http://uxespanol.blogspot.com/2006/03/principios-para-la-optimizacin-en.html>. Acesso em: 19 out. 2007.

NOTAS

[1] Millán, José Antonio. *El español y los buscadores*. El País (seção Opinião), Sexta-feira, 22 de Setembro de 2006, p.15. Em troca, o segundo texto reproduzido apareceu na seção 'Cartas ao Diretor' também do jornal El País em março de 2007.

[2] Méndez Rodríguez, 2002, p.47.

[3] Méndez Rodríguez, 2002, p.60-61), que por sua vez se apóia em Kashyap e Sheth, 2000, p.19-21.

[4] Merlo Vega e Sorli Rojo [2000] analisaram 165 bibliotecas espanholas; Craven [2004 e 2005] estudou as etiquetas <keywords> de páginas *web* das 19 línguas mais presentes na *Web*, e no outro trabalho, determinou o efeito das ferramentas de adição de páginas *web* sobre as meta tags keywords respectivamente; Alimohammadi (2004) calculou a presença de etiquetas keywords em 346 *sites web* de Irán; e, por último, Marcos et al. [2006] estudaram o posicionamento de dez páginas *web* que hospedavam bases de dados e um dos critérios foi a presença de <keywords>.

[5] Algumas publicações impressas recentemente são Sirovich e Darie [2007], Kent [2006], Grappone [2006] ou George [2005]. As páginas web existentes sobre este tema produzem verdadeiro vértice devido a sua quantidade pelo que oferecemos somente dois exemplos: um em inglês e outro em espanhol, que tomam como referência a primeira, mas se ampliam aspectos.
http://www.optimi-seo-tion.com [Acesso em: 4 nov. 2007]
http://www.guiabuscadores.com/posicionamiento/ [Acesso em: 4 nov. 2007]

[6] Classificações retiradas de:
http://www.web1marketing.com/blog/index.php/archives/top-10-seo-factors/ [Acesso em: 19 out. 2007]
http://ezinearticles.com/?Top-15-On-Page-Good-SEO-Factors&id=294120 [Acesso em: 19 out. 2007]

[7] Um algoritmo é um conjunto ordenado e finito de operações que permitem achar a solução de um problema. Os motores de busca utilizam algoritmos para localizar, dar valor e posteriormente, oferecer páginas *web* relevantes para a pesquisa efetuada.