UNIVERSIDADE FEDERAL DE SANTA MARIA
CENTRO DE CIÊNCIAS SOCIAIS E HUMANAS
PROGRAMA DE PÓS-GRADUAÇÃO EM COMUNICAÇÃO

ORGANIZAÇÃO E GERENCIAMENTO DE CONTEÚDOS JORNALÍSTICOS NA WEB SEMÂNTICA

DISSERTAÇÃO DE MESTRADO

Iuri Lammel

Santa Maria, RS, Brasil

2011

ORGANIZAÇÃO E GERENCIAMENTO DE CONTEÚDOS JORNALÍSTICOS NA WEB SEMÂNTICA

Iuri Lammel

Dissertação apresentada ao Curso de Mestrado do
Programa de Pós-Graduação em Comunicação, área de concentração em
Comunicação Midiática, da Universidade Federal de Santa Maria (UFSM, RS),
como requisito parcial para obtenção do grau de
Mestre em Comunicação Midiática

Orientadora: Profa. Dr. Luciana Mielniczuk

Santa Maria, RS, Brasil

2011

Universidade Federal de Santa Maria
Centro de Ciências Sociais e Humanas
Programa de Pós-Graduação em Comunicação

A Comissão Examinadora, abaixo assinada, aprova a proposta de qualificação da dissertação de Mestrado

ORGANIZAÇÃO E GERENCIAMENTO DE CONTEÚDOS JORNALÍSTICOS NA WEB SEMÂNTICA

elaborada por

Iuri Lammel

Como requisito parcial para obtenção do grau de

Mestre em Comunicação

COMISSÃO EXAMINADORA:

_____

Dr.ª Luciana Mielniczuk (UFSM)
(Presidente / Orientadora)

_____

Dr.ª Suzana Barbosa (UFBA)

_____

Dr. Giovani Rubert Librelotto (UFSM)

Santa Maria, dezembro de 2011.

AGRADECIMENTOS

Agradeço, em primeiro lugar, aos meus pais, que além de me apoiarem com carinho, sustentaram meus estudos até o final da graduação.

Aos meus colegas de mestrado, que sempre apoiaram uns aos outros em momentos de dúvidas de e aflição e que nunca deixaram o ânimo cair nestes dois anos de pesquisa.

À Universidade Federal de Santa Maria (UFSM), instituição pública de ensino superior que me formou gratuitamente e com qualidade em diversos níveis de educação: desde o curso técnico até a pós-graduação.

A dois grupos de pesquisa em jornalismo: o Grupo Jornalismo Digital (JORDI), da UFSM, em que participei desde o período de graduação e que me auxiliou no enriquecimento acadêmico e intelectual; e o Grupo de Pesquisa em Jornalismo On-line (GJOL), da UFBA, que, embora eu nunca tenha participado, foi fundamental na minha formação como pesquisador em jornalismo digital, devido a sua rica produção científica na área.

Ao Centro Universitário Franciscano (UNIFRA), que me acolheu como profissional, acreditou e apostou em meu potencial como professor e há mais e três anos me proporciona uma realização profissional ao me oportunizar o exercício da docência com plena liberdade e confiança.

A minha professora orientadora, Luciana Mielniczuk, que não apenas orientou minha dissertação, como também foi a principal responsável por me direcionar ao caminho da pesquisa em jornalismo digital. Além, claro, de me proporcionar uma grande amizade. Obrigado pelas orientações que recebo desde 2004 e pela compreensão (e paciência!) em relação às minhas limitações neste processo de gerar uma dissertação.

Aos visionários que contribuíram, cada um com sua valiosa parte, para o desenvolvimento do hipertexto, da internet e da web: Vannevar Bush, Ted Nelson, Douglas Engelbart, Bob Kahn, Vinton Cerf, Robert Cailliau e Tim Berners-Lee, entre outros que são, para mim, modelos que adoto como exemplo profissional, acadêmico e científico.

Por fim, em especial, agradeço a minha companheira, Laura Cortes, que suportou períodos de ausências e desânimos de um estudante de mestrado que também enfrenta uma rotina de trabalho diário. Agradeço pela compreensão, pelo apoio e pelo amor demonstrados nestes anos.

RESUMO

Dissertação de Mestrado
Programa de Pós-Graduação em Comunicação
Universidade Federal de Santa Maria

ORGANIZAÇÃO E GERENCIAMENTO DE CONTEÚDOS JORNALÍSTICOS NA WEB SEMÂNTICA

Autor: Iuri Lammel
Orientadora: Luciana Mielniczuk

Entre as tecnologias que transformaram o jornalismo digital desde o seu surgimento, destacam-se duas: a World Wide Web (web), rede de documentos digitais que serviu como plataforma à prática jornalística na internet e determinou as três fases evolutivas do jornalismo digital; e as bases de dados, que, agregadas à web, se tornaram a principal tecnologia estruturante dos produtos jornalísticos na fase de transição entre a terceira e a quarta geração do jornalismo digital. No ano de 2001, o cientista Tim Berners-Lee, inventor da web, publicou um artigo com a proposta de uma expansão para esta rede, a qual foi denominada Web Semântica. O artigo propunha uma mudança no conceito da web: da tradicional rede de documentos para uma rede de dados, com capacidade para representar conceitos reais, como pessoas, lugares e objetos. Um grande diferencial desta proposta é que os computadores teriam capacidade para interpretar tais dados e identificar seus significados. Em uma rede semântica, as informações poderiam ser organizadas e gerenciadas de forma mais eficiente e automatizada, e as conexões entre dados seriam mais ricas do que através dos atuais links entre documentos. O conceito de Web Semântica ainda está em fase de amadurecimento, mas já é possível encontrar em funcionamento produtos digitais que aplicam tal conceito. A proposta desta pesquisa é analisar dois casos que aplicam o conceito da Web Semântica no jornalismo digital, mais especificamente na organização e no gerenciamento das informações jornalísticas. Para o embasamento teórico da investigação, foi realizada uma revisão bibliográfica sobre o jornalismo digital, sobre o paradigma do Jornalismo Digital em Base de Dados (JDBD) e sobre o funcionamento das tecnologias empregadas na Web Semântica, tais como o RDF e as ontologias. A pesquisa apresenta caráter exploratório e emprega como estratégia de investigação o estudo de caso, especificamente dos *sites* BBC World Cup 2010 e BBC Wildlife. A análise foi realizada a partir de oito categorias aplicáveis ao estudo do JDBD. Entre os resultados, é constatado que a Web Semântica potencializa algumas das características do JDBD, principalmente devido à automatização. Além disso, foi identificado nos casos estudados que a interoperabilidade automatizada foi o benefício mais vantajoso da Web Semântica em relação às tecnologias até então utilizadas no jornalismo digital, e que pode se tornar uma ruptura caso o projeto de Web Semântica obtenha êxito.

Palavras-chave: Web Semântica, jornalismo digital, Jornalismo Digital em Base de Dados, BBC.

ABSTRACT

Dissertação de Mestrado
Programa de Pós-Graduação em Comunicação
Universidade Federal de Santa Maria

ORGANIZATION AND MANAGEMENT OF JOURNALISTIC CONTENT ON THE SEMANTIC WEB

Author: Iuri Lammel
Adviser: Luciana Mielniczuk

Among the technologies that have modified the digital journalism since its inception, there are two that can be highlighted: 1) the World Wide Web (Web), a network of digital documents that has being used as a platform to the practice of journalism on the Internet and that determined the three generations of digital journalism; and 2) the databases aggregate to the Web, that have become the main technology behind the structuring of journalistic products in the transition between the third and fourth generation of digital journalism. In 2001, the scientist Tim Berners-Lee, inventor of the web, published a paper with a proposal of an extension to this network, which was called the Semantic Web. The paper proposed a change in the concept of the current web: from the traditional network made of documents to a network made of data, plus the technical ability to represent real concepts, such as people, places and objects. A great advantage of this proposal is that computers would be able to understand the data and identify their meanings. With a semantic network, the information could be organized and managed more efficiently and in an automated way, and the connections between the data would be richer than the current hyperlinks between documents. The concept of the Semantic Web is still maturing, but it is currently possible to find digital products that implement this concept. This research aims to analyze two real cases that apply the concept of the Semantic Web in digital journalism, specifically in the organization and management of the newspaper reports. For the theoretical background of research, we conducted a literature review on digital journalism, paradigm of the Digital Journalism on Databases (JDBD) and how the standard technologies of the Semantic Web work, such as RDF and ontologies. This is an exploratory research and it uses the case study as a method. The cases are the site 'World Cup 2010 BBC' and the site 'BBC Wildlife'. The analysis was performed using eight categories applicable to the study of JDBD. Among the results, it is found that the Semantic Web improve some of the characteristics of JDBD, mainly due to the automation on management tasks. Moreover, it identified that automated interoperability was the more advantageous benefit of Semantic Web to both digital journalism cases, and that it can become a potential rupture if the Semantic Web project come to succeed.

Key-words: Semantic Web, on-line journalism, Digital Journalism on Databases, data journalism, BBC.

LISTA DE FIGURAS

LISTA DE APÊNDICES

LISTA DE ANEXOS

SUMÁRIO

INTRODUÇÃO

A presente pesquisa se caracteriza como um estudo interdisciplinar que envolve conhecimentos dos campos do jornalismo e da ciência da computação. Em diversos momentos deste trabalho, as fronteiras entre os conhecimentos específicos de cada área se entrelaçam. Afinal, o jornalismo, tomado aqui como uma atividade que trabalha essencialmente com dados e informações, não teria como escapar dos efeitos transformadores das tecnologias digitais.

Não queremos defender o pensamento reducionista (e tentador) de que novas tecnologias tendem a melhorar a prática jornalística, pois, como afirma o pesquisador Marcos Palacios (2003, p. 16), corremos o perigo de instaurar "um pensamento guiado por uma lógica evolucionista de caráter simplista". Por outro lado, ao considerarmos que a contemporaneidade é marcada, entre outros fenômenos, pelo surgimento de um ciberespaço que redefine práticas sociais e profissionais (LEMOS; LÉVY, 2010), sabemos que as mudanças tecnológicas têm potencial para transformações nos modos de produzir e consumir a informação jornalística. Se o jornalismo digital é uma atividade baseada em plataformas tecnológicas que passam por constantes mutações, então a produção jornalística praticada neste ambiente também passa por redefinições (PAVLIK, 2000).

No decorrer das décadas de 1990 e 2000, o rápido desenvolvimento e popularização dos computadores e das redes fizeram emergir diferentes plataformas digitais para a distribuição da informação, tais como o disco ótico, o correio eletrônico, a web e os *software* aplicativos em dispositivos móveis. Destes, podemos afirmar que a *World Wide Web* ("WWW" ou simplesmente "web") foi uma das tecnologias que mais influenciaram os estudos brasileiros sobre o jornalismo digital das duas últimas décadas, devido a uma diversidade de fatores, tais como o seu alcance global, a sua facilidade na produção e distribuição de conteúdos e a sua lógica de interconexão de documentos (LEÃO, 1999). Foi principalmente a partir das potencialidades técnicas da web que surgiram estudos sobre as características que diferenciam o jornalismo digital das outras modalidades de jornalismo (PALACIOS, 2003), que delimitou os estudos sobre o desenvolvimento do jornalismo digital em diferentes gerações (MIELNICZUK, 2003), e que abriu espaço para os estudos sobre a produção jornalística em sistemas automatizados (SCHWINGEL, 2004) e sobre o jornalismo digital estruturado em bases de dados (MACHADO, 2006; BARBOSA, 2007, 2008a). Por isso, consideramos que uma mudança na forma como a web funciona apresenta potencial para

influenciar nos modos de produção, circulação e consumo da informação (jornalística ou não) no ciberespaço. E, de fato, uma proposta de mudança na web está em curso.

No ano de 2011, a web completou 20 anos desde seu lançamento público e, no decorrer deste período, apresentou atualizações em especificações técnicas importantes (como as atualizações do HTML publicadas pela W3C[1]), além de ter sido enriquecida com o surgimento de tecnologias paralelas, como *plug-ins* para multimídia e linguagens de *script*. Porém, na essência, a organização da web continuou funcionando com base no mesmo conceito de sua origem: como uma rede de documentos conectados. No início da década de 2000, o cientista britânico Tim Berners-Lee, idealizador da própria *World Wide Web*, apresentou um artigo em que propunha um conceito mais avançado para esta rede. A esta proposta, ele denominou "Web Semântica": uma rede que funcionaria não apenas como um sistema de associações de documentos criados prioritariamente para a leitura humana, mas como uma rede de dados, em que os computadores também seriam capazes de identificar os significados dos conteúdos publicados nas páginas (BERNERS-LEE et al., 2002).

Em outras palavras, na Web Semântica as informações publicadas na rede são preparadas para serem compreendidas tanto por humanos quanto por máquinas, o que resultaria em uma web mais eficiente e autônoma na busca e na associação de informações. Para Berners-Lee et al. (2002), passaríamos do paradigma de **web de documentos** para a de **web de dados**, estruturados e adaptados para a interpretação das máquinas. As vantagens de um sistema semântico global alcançariam diversas áreas que trabalham com a organização e o compartilhamento de dados, além da automação em operações que envolvem o gerenciamento dos mesmos, tais como na ciência da computação (BERNERS-LEE et al, 2002; SHADBOLT et al, 2006; KASHYAP et al, 2008; SEGARAN et al, 2009), na ciência da informação (CODINA, 2011; SOUZA E ALVARENGA, 2004) e também no jornalismo (BERTOCCHI, 2010), devido à natureza informativa dessa prática profissional. Tal cenário abre caminho para o desenvolvimento de produtos jornalísticos mais complexos e integrados aos conteúdos publicados na rede, pois na Web Semântica a estruturação dos dados é universalmente padronizada, o que permite o seu compartilhamento.

Quando tratamos de "vantagens" desta tecnologia, não as consideramos exatamente como novidades, mas como o melhoramento, em algum aspecto, das funções até então

[1] A W3C é um grupo de especialistas e de empresas que desenvolvem as principais tecnologias e padrões da Web. Segundo o site da W3C: "*The World Wide Web Consortium (W3C) is an international community that develops standards to ensure the long-term growth of the Web*". Em tradução livre: "A World Wide Web Consortium (W3C) é uma comunidade internacional que desenvolve padrões que asseguram o crescimento da Web em longo prazo". Disponível em: <http://www.w3.org>. Acesso em: 23 jun 2010.

desempenhadas por outras tecnologias, como, por exemplo, no encurtamento do tempo, na maximização da eficácia ou na automatização de operações de publicação, distribuição, recuperação e gerenciamento de dados. Para evitarmos o determinismo presente na ideia de um "processo evolucionário linear de superação de suportes anteriores por suportes novos" (PALACIOS, 2003, p. 22), tratamos estes vantagens como **continuidades e potencializações** de características já exploradas pelo jornalismo digital. Por isso, antes de se analisar as potencialidades trazidas pela Web Semântica, é necessário que tenhamos claro quais as características já exploradas pelos produtos jornalísticos digitais na atualidade.

A prática do jornalismo digital está inserida em um cenário bastante diversificado em termos tecnológicos: além dos computadores e da web, temos a proliferação dos dispositivos móveis conectados em rede, como os *smartphones* e os *tablets*. Com o crescimento vertiginoso na produção e no consumo de dados, uma tecnologia específica se destaca: a base de dados (BD). Mais do que uma mera ferramenta de armazenamento, a BD passa a ser a tecnologia fundamental na organização, estruturação e apresentação das informações, e, por isso, define as funcionalidades e a estética dos produtos informacionais, e passa a ser considerada um formato cultural de nossa época (MACHADO, 2006; MANOVICH, 2001). Atualmente, os produtos jornalísticos tomam as bases de dados como o recurso estruturante em suas diferentes fases produtivas: apuração, composição e circulação (MACHADO, 2006). Por isso, a atual geração do jornalismo digital pode ser caracterizada como a de um Jornalismo Digital em Base de Dados (JDBD) (BARBOSA, 2007).

Logicamente, não basta que uma prática profissional adote uma tecnologia para que seja decretado o início de uma nova geração. Barbosa (2007) lista uma série de indícios que demonstrariam transformações nas práticas jornalísticas e que comprovariam um movimento de transição de paradigma no jornalismo digital, tais como: o desenvolvimento de sistemas de gestão de conteúdos mais complexos, ampla adoção de recursos da Web 2.0, uso crescente de aplicações *mash-ups*, entre outros. Destes indícios, destacamos três que serviram de mote para a presente pesquisa: "[o surgimento de] novos elementos conceituais para a organização da informação; maior integração do material de arquivo na oferta informativa; produtos experimentais que incorporam o conceito de web semântica" (BARBOSA, 2007, p. 9).

Os três indícios citados por Barbosa surgem como iniciativas necessárias em um cenário de saturação na massiva oferta de informação, gerada pelas facilidades oferecidas pelas tecnologias digitais na reprodução de conteúdos. Se por um lado temos uma grande quantidade de informações disponibilizadas, por outro temos como consequência problemas relacionados à busca, localização, acesso e recuperação dessas informações. A Web

Semântica se propõe a ser uma solução para essa situação, pois, com a capacidade das máquinas em compreender o significado das informações, temos como consequência um processo de busca e recuperação de dados mais eficiente. Se, segundo autores e entusiastas da Web Semântica, esta tecnologia oferece vantagens às ciências da informação ao aproveitar o potencial dos computadores para organizar e gerenciar as informações (ou "o conhecimento") de uma forma mais eficiente (BERNERS-LEE et al., 2002; SHADBOLT et al., 2006;), então questionamos neste trabalho: quais seriam as potencialidades que a Web Semântica ofereceria para a organização e o gerenciamento dos conteúdos jornalísticos?

A Web Semântica é um projeto ainda em desenvolvimento. Segundo Kashyap et al. (2008), na engenharia da computação, existe uma ideia conhecida como "regra 5-5-5", de que uma nova tecnologia demora aproximadamente 15 anos entre o período de sua concepção até sua disseminação no mercado de massa. Os primeiros cinco anos são reservados para a pesquisa, os próximos cinco anos para refinamento dos produtos baseados nestas pesquisas, e por fim, os últimos cinco anos são para a saturação do conceito no mercado. Ao considerarmos que a Web Semântica tem como início deste período o ano de 2001, quando Berners-Lee, Hendler e Lassila publicaram o artigo em que apresentam sua proposta, podemos considerar que o momento atual (2011) é de transição entre a experimentação do conceito e o início da aplicação efetiva do mesmo.

Buscamos nesta investigação estudar as contribuições da Web Semântica na organização do conteúdo jornalístico a partir da análise de casos que aplicaram com sucesso este conceito no jornalismo digital. Por tratarmos de produtos digitais pioneiros, devido à incipiente fase da Web Semântica, definimos que o processo metodológico deve adotar a estratégia de estudo de caso, a fim de apresentar e analisar os resultados de tais produtos dentro do contexto do jornalismo. Ao refletirmos sobre a dimensão do conceito de Web Semântica, percebemos que os produtos podem vir a explorar determinados benefícios e deixar de explorar outros. Por isso, consideramos que a melhor metodologia para este trabalho é aquela que analisa mais de um caso, para abrangermos uma quantidade maior de funções das tecnologias semânticas na nossa observação. Adotamos, então, como objetivo principal, identificar contribuições do uso das tecnologias semânticas na organização e gerenciamento dos produtos jornalísticos digitais. Para que isso seja possível, precisamos alcançar resultados nos seguintes objetivos específicos: 1) identificar quais são as tecnologias semânticas utilizadas nos produtos jornalísticos selecionados, 2) compreender como elas são aplicadas, 3) identificar quais as razões do uso destas tecnologias, e, por fim, 4) relacionar os dados obtidos na investigação dos casos selecionados ao atual paradigma do Jornalismo Digital em Base de

Dados, a fim de se compreender as possíveis contribuições da proposta da Web Semântica à prática do jornalismo digital. Para isso, os resultados da pesquisa foram analisados à luz das categorias levantadas por Barbosa (2007, 2008a) em estudos sobre o JDBD, a fim de se descobrir se há indícios de potencializações destas características.

Os objetivos citados caracterizam a atual pesquisa como exploratória devido ao trabalho de identificação das tecnologias empregadas, do seu *modus operandi* no produto em análise e, também, devido à busca de esclarecimentos sobre como um determinado fenômeno funciona em um contexto, no caso a Web Semântica no jornalismo. Segundo Gil, as pesquisas exploratórias são "desenvolvidas com o objetivo de proporcionar visão geral, de tipo aproximativo, acerca de determinado fato. Este tipo de pesquisa é realizado especialmente quando o tema escolhido é pouco explorado [...]" (1989, p. 45). Para o autor, geralmente este tipo de pesquisa é realizado através de levantamento bibliográfico, entrevistas não padronizadas e estudos de caso.

Para delimitarmos o universo da análise, determinamos que os casos selecionados deveriam ser produtos desenvolvidos por iniciativas oriundas do *mainstream* jornalístico, ou seja, de organizações consolidadas no mercado. Após pesquisas bibliográficas e documentais e observações diretas de produtos da web, a organização escolhida[2] foi a British Broadcasting Corporation (BBC), emissora pública de rádio e televisão do Reino Unido. A emissora possui uma equipe de profissionais especializados em arquitetura da informação e desenvolvimento web, e já demonstrou o uso de tecnologias semânticas em mais de um produto. Para realizarmos a nossa investigação, selecionamos dois produtos da BBC, cada um deles como um caso a ser estudado: o site **BBC World Cup 2010** (um site que abriga todo o conteúdo jornalístico da BBC relacionado à Copa do Mundo de 2010) e o site **BBC Wildlife** (um site que reúne uma grande produção de conteúdo multimídia sobre a vida natural).

Para cada caso analisado, tanto a coleta quanto a análise dos dados foram realizadas com o apoio de um protocolo (APÊNDICE A) que divide o processo em duas etapas: uma para a descrição do produto estudado e outra para a análise do emprego das tecnologias semânticas. Na primeira etapa, foi realizada uma observação direta semiestruturada dos produtos digitais selecionados, para que fossem registradas a identificação do produto e a descrição de suas funcionalidades. Na segunda etapa, alimentada pela coleta de dados secundários, buscamos:

[2] O processo de seleção do *corpus* é detalhado no 3º capítulo.

1º) identificar o contexto que justificasse o uso das tecnologias semânticas,
2º) identificar as principais tecnologias semânticas empregadas pelo produto estudado,
3º) descrever o funcionamento das tecnologias semânticas identificadas,
4º) analisar qualitativamente as vantagens encontradas pelas respectivas organizações ao utilizarem tecnologias semânticas, sob a luz das características do JDBD apresentadas por Barbosa: dinamicidade, automatização, flexibilidade, inter-relacionamento/hiperlinkagem, densidade informativa, diversidade temática, visualização (BARBOSA, 2007) e convergência (idem, 2008).

A análise foi realizada pela confrontação de dados obtidos em diferentes fontes, como artigos, documentos, entrevistas, debates e apresentações disponibilizados pelos técnicos desenvolvedores dos produtos estudados.

O desenvolvimento do presente texto está organizado em três capítulos. No primeiro, intitulado "Jornalismo de Dados", é realizada uma retomada dos estudos sobre jornalismo digital nos últimos anos no Brasil: apresentamos alguns aspectos importantes sobre o jornalismo digital, como as suas características e as suas três gerações iniciais. Seguimos para o referencial teórico sobre a tecnologia das bases de dados (BDs), as BDs como formato cultural e como estética de nosso tempo (*database aesthetic*), até chegarmos ao uso das bases de dados no jornalismo. Tratamos, então, do paradigma do Jornalismo Digital em Base de Dados (JDBD), importante conceito para a nossa análise. É neste trecho que apresentamos as categorias para estudo sobre JDBD propostas por Barbosa (2007, 2008a) e que aplicamos em parte da análise dos dados. Para finalizar o capítulo, apresentamos brevemente alguns dos termos e conceitos empregados em outras partes do mundo para a prática do jornalismo em uma era marcada pelas quantidades massivas de dados que circulam globalmente, além de práticas emergentes no jornalismo que surgem em decorrência deste cenário, como as infografias interativas e os aplicativos jornalísticos. Por termos tratado também destes conceitos, julgamos mais apropriado generalizar o título do capítulo como Jornalismo de Dados (livre tradução do termo amplamente utilizado *data journalism*), pois acreditamos que o termo englobe também as práticas do JDBD.

No segundo capítulo, passamos para o referencial teórico relacionado ao campo da Computação. Aqui, tratamos de apresentar e explicar o conceito de Web Semântica, de acordo com a proposta de Berners-Lee et al (2002). O referencial aborda as principais tecnologias semânticas recomendadas pela W3C (triplos em RDF e ontologias em OWL), além de tópicos

derivados desta combinação de tecnologias, como a linguagem de *query*[3] SPARQL, os repositórios de triplos, a técnica de *tagging*, entre outros. Por fim, apresentamos o projeto Linked Data, que é uma série de práticas padronizadas para se publicar dados abertos na web, apropriados para o compartilhamento entre diferentes sites na lógica da Web Semântica. Cabe ressaltar que a Web Semântica é um conceito de uma rede semântica de dados, e que a W3C não é a única que propõe soluções tecnológicas para a realização desta proposta (AKERKAR, 2009). Nossas escolhas sobre as soluções abordadas foram determinadas pelas tecnologias semânticas empregadas pelos casos estudados.

É importante esclarecer que buscamos explicar o que é, como funciona e para que serve a Web Semântica de uma forma didática. Acreditamos que o conceito da Web Semântica ainda não ocupa um lugar destacado nos debates acadêmicos sobre o jornalismo digital; ao menos não no Brasil. No decorrer dos nossos estudos, encontramos um número reduzido de bibliografias da área que tratam de explicar esta proposta sob o ângulo de um jornalista. Acreditamos que o tema Web Semântica deverá ser mais explorado pela comunidade acadêmica do campo da comunicação em trabalhos futuros, e, por isso, esperamos que o capítulo sobre a Web Semântica possa vir a auxiliar no entendimento desta tecnologia àqueles que não estão habituados com os estudos da área tecnológica.

No último capítulo, apresentamos a análise dos dois casos selecionados. Cada um dos casos foi identificado e teve seu funcionamento descrito. Também trazemos para cada caso uma apresentação dos autores que nos baseamos para coletar os dados. Por fim, relatamos para cada caso as tecnologias semânticas empregadas, o funcionamento das mesmas e a análise comparativa com as características do JDBD. O resultado da análise demonstra que a Web Semântica potencializa algumas características do JDBD, principalmente devido à combinação das mesmas com a capacidade apurada de automação, e aponta para uma provável ruptura em relação às atuais características do jornalismo digital, que só será viável caso se consolide de fato uma rede de dados semântica na web.

[3] O termo *query* significa um comando de pesquisa por determinados dados em um banco de dados.

1 JORNALISMO DE DADOS

Com o surgimento da web, seguido de sua popularização, as práticas profissionais baseadas na produção e distribuição de conteúdo informativo e midiático sofreram transformações, algumas bastante evidentes. No jornalismo, a web também teve um impacto significativo nas rotinas de produção e no consumo. As potencialidades do suporte digital em rede criaram possibilidades na construção de narrativas e na apresentação das mesmas, pois, além de herdar a multimidialidade dos diferentes suportes tradicionais, a interface da web é interativa e hipertextual (CANAVILHAS, 2001). Como consequência, a prática jornalística na web, denominada neste texto como jornalismo digital[4], desenvolveu certas características que a destacam de outras modalidades de jornalismo, como o impresso, o radiojornalismo e o telejornalismo.

1.1 Fases e características do Jornalismo Digital

Entre a metade da década de 1990 e o início da década de 2000, alguns estudiosos apresentaram propostas de caracterizações da prática jornalística em suportes digitais em rede. Palacios (2003) realizou uma compilação dessas características do jornalismo digital e também sugeriu outras, resultando assim em um total de seis:

- **Multimidialidade/convergência**: é a convergência das mídias tradicionais (imagem, som, texto) na narrativa. Isso é possível devido ao formato digital dos dados, que permite integrá-los no suporte. Também é possível acrescentar à narrativa outros recursos multimídia, como as animações 2D ou 3D.
- **Interatividade:** é a relação estabelecida entre o usuário e o site e/ou o jornalista. Nesta relação, o leitor sente-se parte integrante do processo jornalístico, pois pode influenciar a narrativa com suas ações. Esta interatividade também pode ocorrer entre os usuários do site, com recursos como chats e fóruns de discussões; ou entre o usuário e os produtores do conteúdo, como os jornalistas, via e-mail. Mielniczuk

[4] Na literatura, há diferentes propostas de nomeação da prática do jornalismo na internet. Mielniczuk (2003) cita algumas das propostas apontadas por autores, como "jornalismo eletrônico", "jornalismo digital", "jornalismo multimídia", "ciberjornalismo", "jornalismo *online*" e "webjornalismo". Cada termo implica em relações da prática jornalística com outros suportes que não apenas a web, por isso, na época, a autora acompanhou Canavilhas (2001) ao apontar o termo "webjornalismo" como o mais apropriado, pois segue a mesma lógica de nomeação de outras modalidades do jornalismo, como o radiojornalismo e o telejornalismo, em que o nome do suporte é colocado antes do termo "jornalismo". Entretanto, neste trabalho tratamos a prática com o termo mais abrangente "jornalismo digital" devido à recente emersão de novas tecnologias digitais que não dependem da web, como no caso dos aplicativos para *smartphones*.

(2001) aponta ainda a interatividade entre usuário e máquina; e entre usuário e a própria publicação, através do hipertexto. O pesquisador Alex Primo sugere a substituição do termo "usuário" pelo termo "interagente", pois tal termo "emana a idéia de interação, ou seja, a ação (ou relação) que acontece entre os participantes. Interagente, pois, é aquele que age com outro" (PRIMO, 2003, p. 7).

- **Hipertextualidade**: é a possibilita de interconexão entre textos a partir de links. Leão (2001) define que os blocos de informações interconectados pelos links podem ser denominados de lexias, que podem ser texto, imagem, som, vídeo etc; ou uma composição com vários destes elementos.
- **Customização do conteúdo/personalização**: é a possibilidade de o interagente configurar o site jornalístico de acordo com seus interesses. Estas configurações podem ser visuais (cores, tamanho dos caracteres etc.), editoriais (pré-seleção dos assuntos, hierarquização de editorias etc.) entre outras.
- **Instantaneidade/atualização contínua:** é a extrema agilidade na atualização do conteúdo disponibilizado para o usuário. Ao contrário da periodicidade do jornalismo impresso, no jornalismo digital as notícias são publicadas instantaneamente e em fluxo contínuo. A televisão e o rádio também são instantâneos, porém a disponibilidade da informação é limitada no tempo, ou seja, o telespectador/ouvinte precisam estar a consumir a informação no exato momento em que ela é veiculada, ao contrário do jornalismo digital, em que o fluxo contínuo é armazenado para acesso a qualquer momento.
- **Memória**: é a capacidade de armazenar os produtos jornalísticos já produzidos anteriormente. Segundo Palacios (2002), o armazenamento de informações é mais viável técnica e economicamente na web do que em outras mídias. Esta memória pode ser disponibilizada tanto aos produtores quanto aos interagentes do conteúdo.

Para o autor, essas não são características novas, pois, de certa forma, também podem estar presentes em suportes anteriores. Segundo Palacios, "[...] as características do Jornalismo na web aparecem, majoritariamente, como Continuidades e Potencializações e não, necessariamente, como Rupturas com relação ao jornalismo praticado em suportes anteriores" (2003, p. 22). Contudo, para o autor, é possível apontar algumas rupturas e a principal delas é a memória, pois, pela primeira vez na história, o jornalismo pode se aproveitar de um espaço praticamente ilimitado, disponível tanto ao produtor quanto ao

consumidor da informação; e, ainda, tal quantidade potencialmente ilimitada de informações é combinada às outras características do jornalismo digital, como a interatividade e a instantaneidade. Logo, a especificidade do jornalismo na web se encontra "não apenas pela Potencialização das características já descritas, mas principalmente pela combinação dessas características potencializadas, gerando novos efeitos" (PALACIOS, 2003, p. 24).

As características que diferenciam o jornalismo digital não apareceram de uma hora para a outra. As potencialidades foram descobertas e postas em prática de forma gradual, de acordo com a evolução da web. Enquanto tais características ainda não eram exploradas, os profissionais jornalistas tendiam a repetir na web os formatos e linguagens dos suportes tradicionais a que eram costumados a produzir. Segundo Canavilhas,

> Marshall McLuhan afirmava que o conteúdo de qualquer medium é sempre o antigo medium que foi substituído. A internet não foi excepção. Devido a questões técnicas, (baixa velocidade na rede e interfaces textuais), a internet começou por distribuir os conteúdos do meio substituído - o jornal. Só mais tarde a rádio e a televisão aderiram ao novo meio, mas também nestes casos se limitaram a transpor para a internet os conteúdos já disponibilizados no seu suporte natural (2001, *online*).

Os sites jornalísticos não passaram a explorar as características do jornalismo digital de forma uniforme. A iniciativa de se explorar as características ocorreu de forma gradual e dispersa. Mesmo assim, é possível definir alguns períodos na recente história desta prática, para fins de estudo sobre o desenvolvimento do jornalismo nos ambientes digitais em rede. Mielniczuk (2003) propõe uma classificação dividida em três momentos: o webjornalismo[5] de primeira geração (ou fase da transposição), o webjornalismo de segunda geração (ou fase da metáfora) e o webjornalismo de terceira geração.

Na primeira geração, os conteúdos das páginas jornalísticas são apenas reproduções de partes de grandes jornais impressos; ou seja, o jornalismo digital era uma transposição de algumas das matérias do jornal impresso para um formato digital, sem adaptação de linguagem e de formato. A atualização era feita a cada 24 horas, pois dependia do fechamento da edição do jornal impresso para que fosse realizada a substituição das matérias nos *sites*.

Na segunda geração, que começou aproximadamente no final dos anos 1990, começa a existir a preocupação em explorar alguns dos recursos da web, como a atualização de notícias durante o decorrer do dia, geralmente em seções chamadas "últimas notícias"; também há maior exploração do hipertexto e do *e-mail* (entre o leitor e o jornal/jornalista). Mesmo assim,

[5] No texto em questão, a autora decidiu por adotar o termo webjornalismo, que tratamos aqui como sinônimo de jornalismo digital.

o modelo do suporte impresso continua como uma referência para o formato dos produtos jornalísticos na web. No webjornalismo de terceira geração, toma força o pensamento de que essa é uma prática diferente do jornalismo impresso, com um potencial de linguagem e formato próprios. Os *sites* jornalísticos passam a utilizar recursos mais específicos da web como os de multimídia (som, imagem), *chats*, enquetes, fóruns de discussões, opções de configuração do *site* de acordo com os interesses do usuário, e o emprego do hipertexto não só na organização da informação, como também dentro da narrativa jornalística. A autora cita como exemplo desta geração o *site* jornalístico MSNBC (www.msnbc.com), que não surgiu de um jornal impresso tradicional, mas da fusão entre uma empresa de *software* (Microsoft) e outra de telejornalismo (NBC). Embora a classificação de Mielniczuk identifique repetições de tendências em *sites* jornalísticos no decorrer dos últimos anos, não significa que todos os produtos jornalísticos da atualidade façam parte da terceira geração; ainda existem sites que se enquadrariam dentro da primeira, da segunda ou até em mais de uma geração.

É necessário ressaltar que a proposta de classificação das fases do webjornalismo de Mielniczuk, publicada em 2003, surgiu em um contexto de plena evolução tecnológica dos computadores, das redes e dos *software* aplicativos. Desde então, as potencialidades da web foram incrementadas com o surgimento, popularização ou intensificação no uso de outras tecnologias que se integraram à rede, além da maturação daquelas já exploradas. Como exemplo, podemos citar as bases de dados (BDs) que, embora já fossem utilizadas na web em meados da década de 1990, começaram a ser exploradas de forma mais complexa e diversificada nos anos 2000, como no caso dos *blogs*. Naturalmente, as empresas jornalísticas passaram a experimentar a aplicação desses recursos em seus produtos. Segundo Ribas, "a utilização dos Bancos de Dados aparece em um momento de avanços do terceiro estágio do webjornalismo" (2004, p. 9). Dentro deste contexto, autores como Schwingel (2005), Barbosa (2007) e Larrondo, Mielniczuk e Barbosa (2008) propõem o surgimento de uma quarta geração do jornalismo digital, caracterizada pelo uso sistemático das BDs.

1.2 Jornalismo Digital em Base de Dados (JDBD)

Desde o início da década de 1990, quando a web surgiu, a conexão de novos servidores na internet passou a crescer em um ritmo exponencial, e a publicação de novas páginas acompanhou esse ritmo, já que os servidores também são utilizados para a

hospedagem de sites[6]. O crescimento da publicação de conteúdos em páginas HTML estáticas passou a ser um problema quando empresas e usuários começaram a utilizar a rede para atividades que exigiam operações de gerenciamento de dados. Tal situação resultou no desenvolvimento de soluções mais flexíveis para o gerenciamento de dados na internet, através das bases de dados.

1.2.1 Bases de dados

Uma base de dados (BDs), ou banco de dados, é um "mecanismo capaz de manipular, armazenar e organizar informações de modo que possam ser recuperadas rapidamente e a qualquer momento" (OLIVIERO, 2002, p. 26). Logo, as BDs não são apenas estruturas para armazenamento de dados, elas também servem para gerenciá-los de forma mais eficiente. Segundo Barbosa (2007), alguns autores da literatura especializada diferenciam os termos "banco de dados" de "base de dados": "banco" é geralmente utilizado para se referir ao conteúdo, enquanto que "base" é utilizada para se referir à estrutura lógico-matemática. Entretanto, por não termos como foco o debate sobre padrões técnicos da tecnologia e por considerarmos a terminologia nas pesquisas em países que se destacam nessa área de estudo, como EUA (*database*), Espanha e Portugal, decidimos adotar o termo "base de dados" para nos referirmos a ambos os conceitos e assim acompanhamos, neste trabalho, a mesma escolha terminológica de Barbosa.

De acordo com Takai (et al, 2005), as possíveis ações de gerenciamento dos conteúdos armazenados nas BDs são definidas e executadas pelo Sistema Gerenciador de Banco de Dados (SGBD). Os SGBDs surgiram na década de 1960 e, desde então, evoluíram em diversos tipos ou modelos, cada qual mais apropriado para determinados contextos. Entre esses modelos, os mais utilizados são: o modelo hierárquico[7], o modelo em redes[8], o modelo relacional[9] e o modelo orientado a objetos[10].

---

[6] É possível observar o aumento do número de servidores em cada ano em uma página da Internet Systems Consortium, que apresenta estas estatísticas em uma tabela atualizada periodicamente. Disponível em: <www.isc.org/solutions/survey/history>. Acesso em: 12 dez. 2010.

[7] Surgiu nos primeiros SGBDs; são estruturados em hierarquias ou árvores, e os registros são associados uns aos outros em sequências hierárquicas, como se fossem "galhos" (TAKAI, et al, 2005).

[8] Surgiu como uma extensão ao modelo hierárquico, quebra a ordem hierárquica ao permitir associação dos registros a vários outros que estejam fora de suas sequências, ou seja, de outros "galhos" (TAKAI, et al, 2005).

[9] Amplamente utilizado nos dias atuais, são modelos baseados em tabelas, em que cada tabela possui dados estruturados em colunas e linhas, que podem ser relacionados a outras tabelas da base de dado.

[10] Surgiu para sanar algumas limitações do modelo relacional em determinados casos específicos e mais complexos (TAKAI, et al, 2005).

Ainda segundo Takai (et al, 2005), os sistemas de bases de dados podem ser estruturados em diferentes arquiteturas. Uma arquitetura muito utilizada é a do cliente-servidor, apropriada para redes de computadores. Nesta arquitetura, os servidores (computadores principais) armazenam os dados, que são então solicitados pelas máquinas clientes (como PCs e impressoras), conectados aos servidores em um ambiente em rede. Desta mesma forma funcionam os sistemas gerenciadores de bancos de dados. Segundo Oliviero:

> - As informações pertencentes ao banco de dados ficam concentradas em um ou mais servidores que têm por objetivo "servir" as demandas de consultas, alterações, inclusões, etc. requisitadas pelos seus "clientes".
> - Todo processo é realizado no servidor (ou servidores) pelo gerenciamento de banco de dados. Os clientes (usuários finais) apenas recebem em suas estações as informações já processadas e organizadas, diminuindo drasticamente o tráfego na rede e conseqüentemente aumentando o desempenho do sistema com respostas mais rápidas e eficientes (OLIVIEIRO, 2002, p. 28-29).

A web utiliza a lógica cliente-servidor, pois as páginas em HTML são armazenadas em servidores conectados à internet, enquanto os computadores (clientes) fazem a requisição destes arquivos, que são enviados, armazenados localmente e então interpretados pelos navegadores. Então, quando um site utiliza um sistema de armazenamento de conteúdo em bases de dados, significa que um SGBD gerencia os dados em um servidor que, por sua vez, alimenta a página HTML enviada para os clientes da web (os *software* navegadores instalados nos computadores pessoais) (REESE, 2000).

Além da arquitetura cliente-servidor, a web também utiliza a BD relacional. Este modelo é baseado em organização por tabelas, em que cada tabela possui dados estruturados em colunas e linhas, que podem ser relacionados a outras tabelas da base de dado. Então, quando um computador faz uma requisição de dados armazenados em uma base de dados, o SGBD instalado no servidor executa as ações necessárias nas tabelas que formam a base de dados alocada neste servidor em questão. Entre as ações possíveis, podemos citar: a inclusão de novos dados, a alteração ou exclusão de dados armazenados e a recuperação (busca) de determinados dados.

A web começou como sistema de documentos digitais estáticos, ou seja, sem o uso de bases de dados para o armazenamento dos conteúdos. Com o tempo, as BDs se consolidaram como uma forma mais eficiente de armazenamento de dados na web, e uma das razões para essa consolidação foi a disseminação das ferramentas de publicação e dos sistemas gerenciadores de conteúdos (*Content Managment System* ou CMS), que são sistemas

direcionados "à administração e gerenciamento do conteúdo, voltado para publicação, para os processos de seleção, aprovação e edição" dos mesmos (SCHWINGEL, 2009, p. 2).

As páginas que publicam informações diretamente no código HTML são chamadas **estáticas**, enquanto as que publicam a partir de bases de dados são chamadas de **dinâmicas**, pois têm seus conteúdos modificados mais facilmente e muitas vezes de forma automatizada. Nesta mesma linha de pensamento, Kashyap divide o conteúdo da web em dois grupos: o primeiro, chamado de **web superficial**, é um grupo de páginas estáticas publicamente disponíveis na rede. O outro grupo, denominado **web profunda**, consiste em bases de dados acessíveis à web e também de páginas dinâmicas, que não são "largamente conhecidas pelo usuário 'comum', mesmo que a informação disponível na web 'profunda' seja 400 a 550 vezes maior que a informação na 'superfície'" [11] (2008, p. 23, tradução nossa[12]). Com base na sistematização das gerações do jornalismo digital (MIELNICZUK, 2004), podemos associar a web superficial aos produtos encontrados nas primeiras gerações e a web profunda aos produtos da terceira geração e também aos produtos da quarta geração do jornalismo digital, que seria a fase caracterizada pelo uso sistemático das bases de dados (BARBOSA, 2007).

1.2.2 Bases de dados como forma cultural

As funcionalidades das BDs em modelos relacionais e estruturadas na arquitetura cliente-servidor têm um poder potencial de criação bastante significativo nos meios digitais, justamente devido às possibilidades de associações e combinações de dados digitais, mesmo nos casos em que os dados se encontram em formatos diferentes, pois apresentam natureza bastante flexível nas combinações. Para se compreender essa natureza dos dados, Manovich (2001) lista em seu o texto *The Language of New Media* cinco princípios das "novas mídias" que se aplicam aos conteúdos digitais:

- **representação numérica**: qualquer mídia digital, independente de ser originalmente criada no computador ou convertida de um suporte analógico, é composta por códigos digitais que são representados numericamente; logo, todas as mídias digitais podem ser manipuladas matematicamente;
- **modularidade**: todas as mídias digitais são formadas pelas mesmas estruturas modulares, independente da escala em que se encontram: a foto é formada por

[11] *[…] wich are not widely known by "average" surfers, even though the information available on the "deep" Web is 400 to 550 times larger than the information on the "surface".*
[12] As traduções realizadas neste trabalho foram realizadas pelo autor do presente trabalho. Para cada trecho traduzido, apresentamos também a citação na língua original, em nota de rodapé.

pixels, o vetor é formado por curvas e linhas etc. Tais mídias podem ser combinadas, mas podem manter suas estruturas modulares independentes umas das outras, como no caso de uma animação em Flash, que combina áudio, imagens, textos e vídeos. Mesmo que exista combinação, cada mídia mantém sua estrutura mínima. Nas páginas HTML, ocorre o mesmo;

- **automação**: os princípios da representação numérica e da modularidade permitem que certas operações sejam automatizadas na criação, na manipulação e no acesso das mídias, removendo em parte a participação humana no processo de criação;
- **variabilidade**: como consequência dos princípios da representação numérica e da modularidade, as novas mídias podem existir em diferentes (potencialmente infinitas) versões. Ao invés de variabilidade, seria possível utilizar também os termos "mutável" ou "líquido";
- **transcodificação**: considerada por Manovich a consequência mais substancial da computadorização da mídia, o princípio diz que as novas mídias, quando digitalizadas, passam a ser codificadas tanto em um formato com organização estrutural compreensível por humanos (como o significado simbólico de uma imagem a partir das linhas, curvas etc), quanto em uma organização estrutural "compreensível" pelas máquinas a partir de convenções estabelecidas (como a cor RGB dos pixels, a dimensão da foto, o tamanho do arquivo). Com a transcodificações, os computadores podem relacionar diferentes tipos de arquivos (textos, áudios, vídeos etc) a partir destas convenções.

Ao identificar a organização estrutural reconhecida por humanos como "camada cultural" (*cultural layer*) e as convenções dos computadores como "camada computacional" (*computer layer*), e ao considerar que as novas mídias são criadas, distribuídas, armazenadas e arquivadas em computadores, Manovich (2001) acredita que a camada computacional deverá começar a influenciar de forma significativa na lógica cultural tradicional da mídia; ou seja, a camada computacional deverá influenciar a camada cultural.

Para ilustrar como as mídias podem ser estruturadas por BDs e como podem explorar os princípios propostos por Manovich, citamos como exemplo os vídeos do site YouTube[13]: enquanto os suportes tradicionais de vídeo (cinema e televisão) apresentam basicamente uma sucessão de imagens sincronizadas com uma ou mais trilhas de áudio, o YouTube tem a

[13] http://www.youtube.com

capacidade de apresentar o mesmo recurso (imagens em movimento com áudio), mas também permite a combinação desse produto audiovisual com conteúdos que estejam em outros formatos e armazenadas em BDs, como, por exemplo, comentários (em textos) ou *links* (através de figuras clicáveis em forma de caixas). Para ilustração, apresentamos um caso específico de um vídeo do YouTube: nele, é apresentado um comentário do próprio autor (tela à direita da Figura 1), que aparece em determinado local do plano do vídeo e em limitado período de tempo (circulado na Figura 1). O comentário foi inserido de forma dinâmica no vídeo, pois estava armazenado em uma tabela de base de dado.

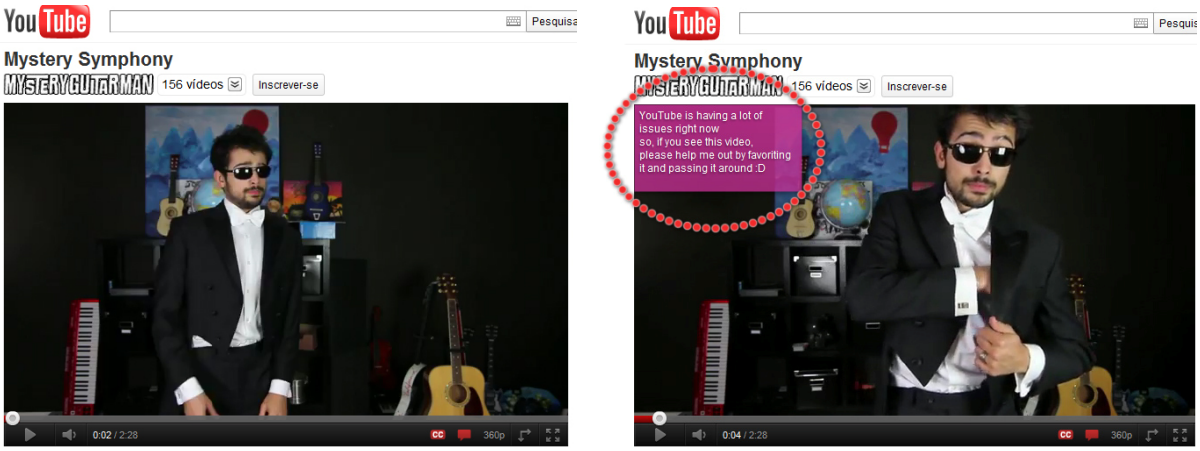


Figura 1 – Vídeo do YouTube com inserção dinâmica de comentário sobreposto ao vídeo[14]

Ao final do vídeo, são mostradas duas caixas em determinadas áreas que funcionam como links (marcadas na Figura 2), para remeter o usuário a outros vídeos do mesmo autor. A localização e o tamanho das caixas, assim como o período de tempo e o link a qual remetem, são informações fornecidas por uma base de dado.

[14] Mystery Symphony. Disponível em: < http://www.youtube.com/watch?v=Ul95hTnO3h4>. Acesso em: 25 jan 2011.

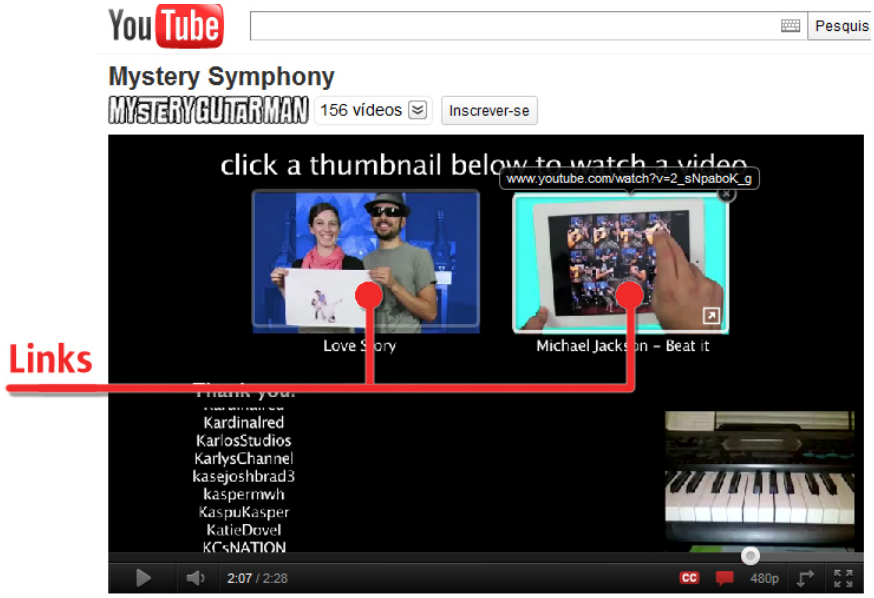


Figura 2 – Vídeo do YouTube com inserção dinâmica de links sobrepostos ao vídeo[15]

O inter-relacionamento de diferentes formatos de mídias ocorreu porque as BDs podem relacionar dados que estão em formatos diferentes, mas codificados com o mesmo código binário (princípio da representação numérica). Para isso, o site relacionou tabelas[16] de base de dados diferentes (ex.: relacionou a base de dados do audiovisual com as tabelas de comentários armazenadas em outras tabelas), cruzou tais dados de forma automatizada (princípio da automação) para formar um novo produto resultado de várias combinações, embora os elementos que formam esse novo produto ainda mantenham as suas características próprias (princípio da modularidade). Esse produto ainda poderia ser apresentado de outras maneiras, através da agregação de elementos surgidos posteriormente à publicação, como sobreposição de novos comentários oriundos de redes sociais ou combinações com outras mídias relacionadas que seriam publicadas no futuro (princípio da variabilidade).

Além do produto audiovisual com sobreposição dinâmica de dados armazenados em BDs, o YouTube também apresenta uma página HTML que combina o resultado de outros cruzamentos de dados. Seguindo no exemplo do vídeo anterior, podemos perceber que a página dedicada ao vídeo em questão também apresenta resultados de buscas em BDs

[15] Mystery Symphony. Disponível em: < http://www.youtube.com/watch?v=Ul95hTnO3h4>. Acesso em: 25 jan 2011

[16] O exemplo do vídeo no YouTube ilustra o funcionamento de uma base de dado relacional em um produto que envolve composição de uma página com elementos multimídia. Porém, cabe ressaltar que a empresa Google desenvolveu um modelo de base de dado próprio, denominado *Bigtable*, que tecnicamente não é considerado relacional, mas distribuído, embora utilize tabelas, linhas e colunas (CHANG et al, 2006). Segundo desenvolvedores da empresa, embora o *Bigtable* não seja tecnicamente considerado modelo relacional, ele se assemelha a esse modelo no seu funcionamento, porém com algumas especificidades que o grande volume de dados gerado em seus serviços exige e o modelo relacional não comporta. Logo, tomamos o exemplo do YouTube como uma possibilidade viável em uma base de dados relacional.

diversas; é o caso da seleção e apresentação de informações relativas ao vídeo (ver marcação A na Figura 3), tais como descrição, número de visitas, avaliações, lista de vídeos sugeridos por outros usuários como resposta ao vídeo apresentado na página (marcação B na Figura 3), comentários de usuários (marcação C na Figura 3) e vídeos relacionados ao apresentado na página (marcação D na Figura 3).

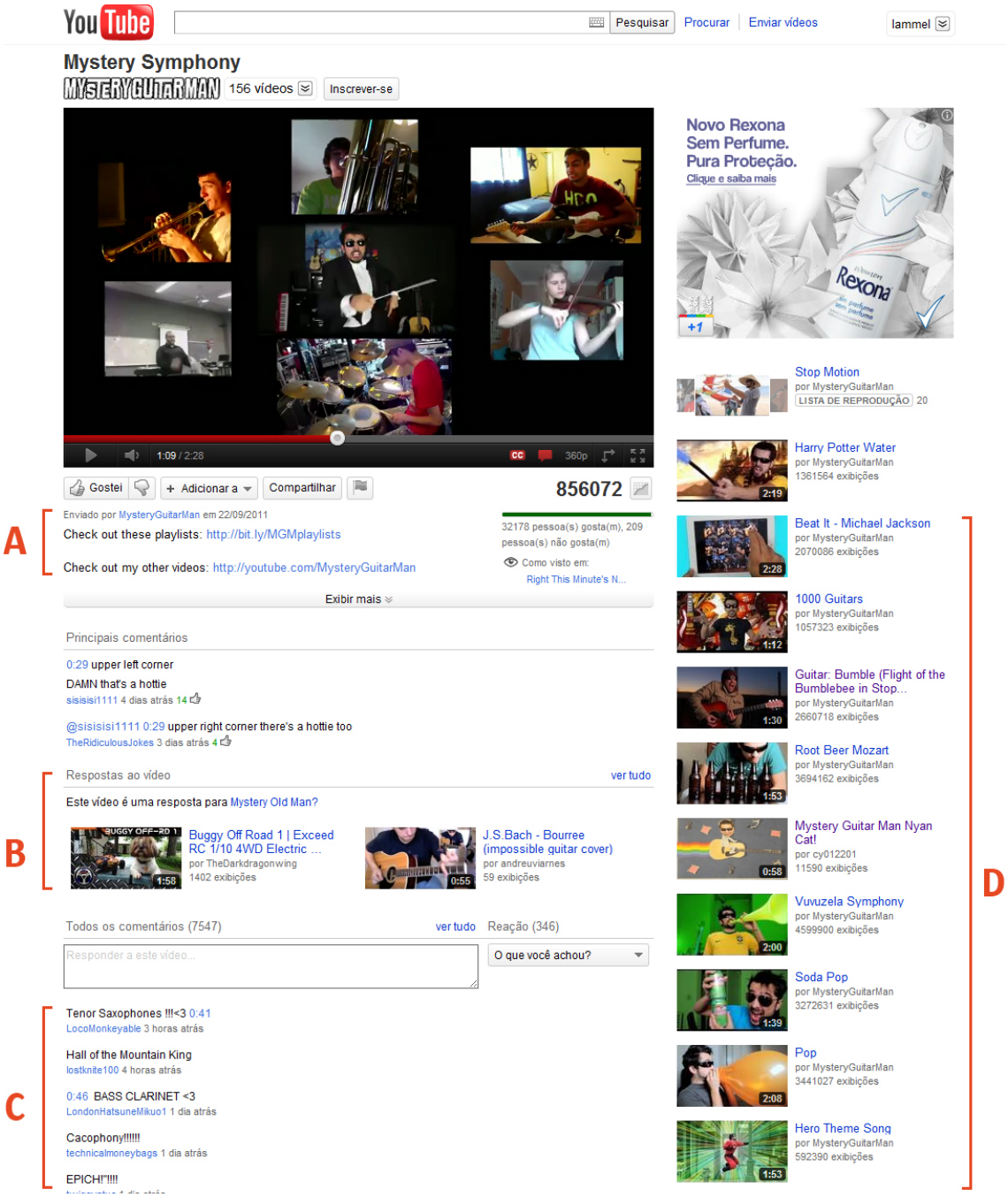


Figura 3 – Página de vídeo do YouTube com inserção dinâmica de dados[17]

[17] Mystery Symphony. Disponível em: < http://www.youtube.com/watch?v=Ul95hTnO3h4>. Acesso em: 25 jan 2011

Provavelmente, o autor[18] do vídeo apresentado neste caso tem na criação de seus audiovisuais a influência da camada computacional. Percebemos que, em alguns de seus vídeos, a personagem costuma realizar gesticulações com as mãos e apontar com os dedos para as caixas de link criadas pelo editor do vídeo, demonstrando que a produção de imagens pode ser planejada de acordo com os elementos gerados a partir de BDs. Além disso, seus vídeos podem apresentar uma linha de criação que privilegia a participação de seu público, já que em muitos produtos há a incorporação de materiais produzidos pelos usuários, como o caso ilustrativo da Figura 3, em que a edição reúne gravações enviadas pelos usuários para, então, formar um concerto musical de maneira colaborativa. A proposta de envio de materiais é articulada nos comentários da página do YouTube, ou seja, através de um elemento gerado pela BD.

Assim como o vídeo do YouTube apresentado no exemplo anterior, outros produtos culturais de nossa era são planejados, desenvolvidos, estruturados e apresentados com uma estética característica que os diferenciam dos produtos tradicionais. Nessa estética, as diferentes mídias são combinadas com elementos gerados a partir das BDs, tais como os comentários e avaliações de usuários, os *links* e as sugestões automáticas de conteúdos relacionados. "Do mesmo modo que a narrativa literária ou cinematográfica é um plano arquitetônico na Modernidade, a Base de Dados emerge como uma forma cultural típica para estruturar as informações sobre o mundo/realidade na cultura dos computadores" (MACHADO, 2006, p. 17). Portanto, mais do que uma mera ferramenta de armazenamento, as BDs passam a ser a tecnologia fundamental na organização, estruturação e apresentação de conteúdos diversos, tanto os culturais e artísticos quanto os próprios produtos informativos e midiáticos (como os jornalísticos).

A importância das BDs emerge não apenas pela função facilitadora na inserção, edição, seleção e combinação de dados, mas também por ser a estrutura elementar de uma estética típica da era dos computadores; uma era marcada pelo crescimento exponencial dos dados e acostumada com a estrutura do hipertexto e com a conveniência da interação homem-máquina, características essas que distanciam os atuais conteúdos digitais dos formatos tradicionais, limitados no espaço/tempo e com possibilidades hipertextuais e interativas restritas. Para este formato típico dos computadores, Farbiaz e Barbosa (2009) apresentam o termo **estética base de dados** (*database aesthetic*), termo que na área da arte digital significa

[18] O MysteryGuitarMan é um produtor assíduo do site YouTube, com produção mensal de vídeos e com mais de 300 milhões de exibições em janeiro de 2012. Estatísticas disponíveis na página do usuário no YouTube. Disponível em: <http://www.youtube.com/user/MysteryGuitarMan>. Acesso em: 25 jan. 2012.

"os princípios estéticos aplicados na imposição da lógica das bases de dados a qualquer tipo de informação, filtro de coleções de dados e visualização dos dados" [19] (PAUL, *online*, p. 1), princípios esses presentes nos produtos das BDs que caracterizam a produção cultural de nossa era. Por isso, Manovich defende que as BDs são **formas culturais** típicas das sociedades em redes, pois estruturam todo o processo criativo quando o objeto consiste de uma ou mais interfaces vinculadas às BDs (MACHADO, 2006). No contexto do jornalismo digital, a estética base de dados é uma metáfora com um "modo particular para a apresentação das informações jornalísticas já desvinculado da metáfora do impresso - *broadsheet metaphor* - e que procede diretamente do emprego das BDs" (FARBIAZ E BARBOSA, 2009, p. 1).

Segundo Machado, é evidente que há uma migração do conhecimento produzido pelas organizações jornalísticas para as BDs, e por isso que "a plena incorporação destas organizações à lógica do ciberespaço pressupõe uma adequação de suas estruturas ao formato das Bases de Dados" (2006, p. 7). Para o autor, a modalidade jornalística que usa as BDs utiliza esta tecnologia para todos os processos de produção jornalística: apuração, composição e circulação.

1.2.3 Bases de dados no jornalismo

Embora o jornalismo digital tenha passado a adotar as BDs em seus produtos na terceira geração do webjornalismo (RIBAS, 2004), não foi a primeira vez que elas foram incorporadas a essa prática profissional. Ainda na década de 1970, segundo Barbosa (2007), as BDs já eram utilizadas nas redações, porém não como forma de organização ou apresentação da narrativa jornalística; elas eram utilizadas como ferramentas de arquivamento e, em seguida, como auxílio ao processo de apuração dentro das redações, contribuindo para o desenvolvimento da Reportagem Assistida por Computador (*CAR*). Como exemplo de sistema de armazenamento, Machado (2006) cita o caso do The New York Times que na metade dos anos 1980 já possuía uma base de dados com três milhões de documentos. Entretanto, poucas empresas jornalísticas são estruturadas em BDs, por mais que estas ofereçam vantagens à pesquisa e apuração jornalísticas. O autor tenta buscar uma resposta a esse enigma e, para isso, evoca os conceitos de *mnémè* e *anámnèsis* do filósofo grego Aristóteles. A *mnémè* significa a simples conservação do passado; já a *anámnèsis* consiste na ativação desse passado no presente. As redações geralmente seguem a linha do primeiro

[19] "*[…] aesthetic principles applied in imposing the logic of the database to any type of information, filtering data collections, and visualizing data […]".*

conceito ao utilizarem as BDs apenas como sistema de armazenamento da memória em redes de dados internas e, assim, deixam de aproveitar a potencialidade de se construir narrativas com a exploração dos dados armazenados de forma estruturada.

As BDs já eram utilizadas em redações antes do surgimento da web, mas podemos identificar potencialidades de sua incorporação especificamente no jornalismo digital. Segundo Machado, a lógica arquivista no conceito de *mnémè* "contraria as características da memória no ciberespaço porque mantém um processo individual e centralizado da produção" (2006, p. 26). O autor afirma ainda que para haver a incorporação da lógica das bases de dados às empresas jornalísticas, deverá ocorrer a "utilização casada das funções de modelo de estruturação da informação, espaço para a criação de narrativas e lugar para a ativação da memória" (2006, p. 27). Nesse sentido, a base de dados se constituiria como "espaço para a criação de narrativas" porque mais do que um sistema matemático-lógico de armazenamento, as bases de dados assumem três funções na sociedade: "1) de formato para a estruturação da informação; 2) de suporte para modelos de narrativa multimídia e 3) de memória dos conteúdos publicados" (MACHADO, 2006, p. 16); e por essa razão o autor concorda com Manovich na afirmação de que a base de dados é uma forma cultural típica das sociedades das redes, assim como a tradicional narrativa linear também é uma forma cultural, construída em suportes lineares como voz, impresso, TV e rádio. No entanto, ao invés de contrapor as duas formas culturais (BDs x narrativas), Manovich afirma que é necessário reconsiderar o conceito de narrativa, pois se no conceito tradicional uma narrativa é um objeto cultural que possui um narrador, um ator (ou mais) e uma história com uma sequência de eventos (MACHADO, 2006), hoje, com as interfaces interativas, as narrativas nas "novas mídias" giram em torno de um espaço não necessariamente linear, navegável (através dos hiperlinks), ativado por um usuário que detém o controle da navegação.

Além da produção de narrativas, as BDs na web também potencializam o consumo da informação, pois, diferentemente das redes internas e privadas de arquivamento de dados, no jornalismo digital os usuários têm acesso às BDs de forma instantânea, através de sistemas de busca presentes nos sites ou a partir do próprio produto jornalístico, já que os conteúdos armazenados em BDs são apresentados ao usuário em interfaces hipertextuais. Estas interfaces são apresentadas em forma de narrativa e a potencialidade está justamente na possibilidade de desenvolver diferentes modelos de narrativas a partir das BDs. É por isso que a base de dados não é em si um novo tipo de narrativa ou uma concorrente da narrativa linear tradicional, mas sim um "suporte para o desenvolvimento de diferentes modelos de narrativa multimídia" (MACHADO, 2006, p. 24).

No jornalismo digital, a tecnologia da base de dados oferece alguns recursos que, combinados, enriquecem as formas de se organizar, gerenciar e apresentar as informações. Segundo Barbosa, no jornalismo, as BDs:

> [...] desempenham um conjunto de funções percebidas tanto quanto à gestão interna dos produtos, quanto aos processos de apuração e contextualização, à estruturação das informações, à composição das peças informativas, assim como à recuperação das informações e à apresentação dos conteúdos (BARBOSA, 2007, p. 27).

Para que as funções citadas possam ser aplicadas aos produtos jornalísticos, é necessário que os conteúdos de tais produtos sejam formatados e inseridos nas BDs de forma prática, ágil e acessível ao jornalista, já que nem sempre esses profissionais apresentam conhecimentos técnicos apurados de informática. Para a publicação de conteúdos formatados à lógica das BDs, utilizam-se sistemas de publicação que são "ferramentas ou sistemas que facilitam a inclusão de informações em produtos ou serviços internet com vistas a deixar o conteúdo na página ou no mecanismo para ser acessado a *posteriori*" (Schwingel, 2008, p. 5). Estes sistemas de publicação são constituídos basicamente por formulários digitais que permitem a inserção de dados textuais e multimídia em uma base de dados. Geralmente, os sistemas exigem uma identificação, com senha do usuário que publica os dados, e permitem o acesso de múltiplos usuários que podem portar permissão para a edição de um mesmo conteúdo, resultando em sistemas de produção colaborativa.

As ferramentas de publicação, além de alimentarem as BDs em uma estrutura apropriada, passaram a ter a capacidade de gerenciar os conteúdos armazenados, tanto de forma manual, através da edição dos conteúdos pelos jornalistas, como de forma automática, ao realizarem operações massivas ou especializadas sem a intervenção humana, como no caso de se reordenar ou filtrar milhares de registros armazenados, ou de se inter-relacionar dados diferentes a fim de se obter novos dados. Esta ferramenta de publicação mais complexa foi denominada de *Content Management Systems* (CMS) ou simplesmente Sistemas Gerenciadores de Conteúdo (SGC) que, além da publicação, edição e automatização de operações, também oferecem ferramentas para seleção, aprovação e edição dos conteúdos (SCHWINGEL, 2009), aproximando ainda mais essa ferramenta dos processos produtivos jornalísticos. Schwingel esclarece que um sistema publicador para jornalismo digital é mais complexo que as ferramentas de publicação utilizadas em *blogs*, "pois visam incorporar efetivamente as características do Jornalismo Digital tanto na concepção do site (na

arquitetura da informação do produto) quanto na estrutura da notícia (na arquitetura da informação de cada matéria)" (SCHWINGEL, 2004, p. 5).

Portanto, os possíveis novos modelos de narrativas na web não dependem apenas da estrutura das BDs, mas também dos CMS, já que são eles que determinam a entrada e o gerenciamento dos conteúdos nas BDs. Para Machado,

> [...] mais do que definir o sistema de gestão de conteúdos como requisito tecnológico essencial para a composição de narrativas multimídia em Bases de Dados, existe a necessidade de perceber que, no caso jornalístico, este sistema deve apresentar determinadas características particulares. A diversidade de etapas do processo de produção de conteúdos jornalísticos – apuração, composição, circulação – demanda a existência de um sistema complexo de produção e gestão, que seja capaz de incluir subsistemas específicos (MACHADO, 2006, p. 62).

Os CMS não são apenas ferramentas de entrada e gerenciamento de conteúdos; os CMS são, geralmente, plataformas que englobam toda a estrutura do site (desde a entrada dos dados até a apresentação da interface) e podem realizar operações automatizadas de seleção, filtro e categorização dos dados armazenados para apresentá-los ao usuário. Além da automatização na apresentação, alguns CMS mais complexos também podem automatizar a inserção de dados nas BDs, como no caso de sites que automaticamente armazenam o número de vezes que uma notícia foi acessada, compartilhada ou avaliada pelos usuários.

1.2.4 JDBD: paradigma para a quarta geração do jornalismo digital

Ao associarmos os produtos jornalísticos da terceira geração com a emersão de diversas tecnologias e práticas sociais na web em um contexto da estética base de dados, podemos apontar para indícios de uma nova geração de produtos jornalísticos, em que os jornalistas não apenas inserem as características do jornalismo digital em seus produtos, mas também experimentam novas narrativas e diferentes suportes além da web. Barbosa cita diversos destes indícios que caracterizam um movimento para a quarta geração:

> O cenário no qual emerge a quarta geração do ciberjornalismo é marcado pela consolidação das bases de dados como estruturantes da atividade jornalística e como agentes singulares no processo de convergência jornalística; equipes mais especializadas; desenvolvimento de sistemas de gestão de conteúdos (SGC) mais complexos e baseados preponderantemente em softwares e linguagens de programação com padrão *open source*, formato XML (eXtensible Markup Language), algoritmos; acesso expandido por meio de conexões banda larga; proliferação de plataformas móveis; consolidação do uso de blogs; ampla adoção de recursos da Web 2.0; incorporação de sistemas que habilitam a participação efetiva do usuário na produção de peças informativas; produtos diferenciados criados e

> mantidos de modo automatizado; sites dinâmicos; narrativas multimídia; utilização de recursos como RSS (Really Simple Syndication) para recolher, difundir e compartilhar conteúdos; aplicação da técnica do tagging na documentação e na publicação das informações; uso crescente de aplicações mash-ups; do conceito de geolocalização de notícias ou geocoding news; uso do podcasting para distribuição de conteúdos em áudio; ampla adoção do vídeo em streaming; novos elementos conceituais para a organização da informação; maior integração do material de arquivo na oferta informativa; produtos experimentais que incorporam o conceito de web semântica; emprego de metadados e data mining para categorização e extração de conhecimento; aplicação de novas técnicas e métodos para gerar visualizações diferenciadas para os conteúdos jornalísticos que auxiliam a sobrepujar a metáfora do impresso (*broadsheet metaphor*) como padrão (BARBOSA, 2008a, p. 9).

Como observado acima, os produtos jornalísticos passam a incorporar novas técnicas em suas fases de apuração, composição e circulação, que potencializam as características do jornalismo digital de terceira geração tais como a interatividade (ex.: "incorporação de sistemas que habilitam a participação efetiva do usuário na produção de peças informativas"), a multimidialidade (ex.: "uso do podcasting para distribuição de conteúdos em áudio; ampla adoção do vídeo em streaming;"), customização (ex.: "conceito de geolocalização de notícias ou geocoding news"), atualização contínua (ex.: "utilização de recursos como RSS para recolher, difundir e compartilhar conteúdos"), hipertextualidade (ex.: "aplicação de novas técnicas e métodos para gerar visualizações diferenciadas para os conteúdos jornalísticos") e memória (pelo uso intensivo do próprio banco de dados).

Barbosa (2007) sugere que nessa transição entre a terceira e a quarta geração, desponta um paradigma que passa a definir as características dos produtos jornalísticos da quarta geração. A esse paradigma, a autora denominou **Jornalismo Digital em Base de Dados (JDBD)**, que, em suas palavras, é:

> [...] o modelo que tem as bases de dados como definidoras da estrutura e organização, bem como da apresentação dos conteúdos de natureza jornalística, de acordo com funcionalidades e categorias específicas, que vão permitir a criação, a manutenção, a atualização, a disponibilização e a circulação de produtos jornalísticos digitais dinâmicos (BARBOSA, 2007, p. 218).

As funcionalidades citadas pela autora em sua conceituação de JDBD foram identificadas através da leitura de outros autores que estudam o tema. No total, Barbosa elencou 18 funcionalidades das BDs no jornalismo digital, que são:

- Indexar e classificar as peças informativas e os objetos multimídia;
- Integrar os processos de apuração, composição e edição dos conteúdos;
- Conformar padrões novos para a construção das peças informativas;
- Agilizar a produção de conteúdos, em particular os de tipo multimídia;
- Propiciar categorias diferenciadas para a classificação externa dos conteúdos;

- Estocar o material produzido e preservar os arquivos (memória), assegurando o processo de recuperação das informações;
- Permitir usos e concepções diferenciadas para o material de arquivo;
- Garantir a flexibilidade combinatória e o relacionamento entre os conteúdos;
- Gerar resumos de notícias estruturados e/ou matérias de modo automatizado;
- Armazenar anotações semânticas sobre os conteúdos inseridos;
- Habilitar o uso de metadados para análise de informações e extração de conhecimento, seja por meio de técnicas estatísticas ou métodos de visualização e exploração, como o *data mining*;
- Ordenar e qualificar os colaboradores e "repórteres cidadãos";
- Orientar e apoiar o processo de apuração, coleta e contextualização dos conteúdos;
- Regular o sistema de categorização de fontes jornalísticas;
- Sistematizar a identificação dos profissionais da redação;
- Cartografar o perfil dos usuários;
- Transmitir e gerar informações para dispositivos móveis (celulares, computadores de mão, *iPods*, entre outros);
- Implementar publicidade dirigida (BARBOSA, 2007, p. 220).

As funcionalidades citadas não são regras: são possibilidades. Nem sempre os produtos jornalísticos em BDs exploram tais funcionalidades, mas é possível perceber que eles compartilham algumas características que os definem. Em uma investigação (doutoral), Barbosa (2007) analisou diversos destes produtos e elencou sete categorias que demarcam e complementam as particularidades do JDBD. São elas: dinamicidade, automatização, inter-relacionamento/hiperlinkagem, flexibilidade, densidade informativa, diversidade temática e visualização. A seguir, apresentamos uma breve explanação sobre cada uma delas:

a) **dinamicidade:** é a característica básica das BDs de dinamizar os conteúdos apresentados em produtos da web. Ao contrário do conteúdo estático dos sites produzidos apenas em HTML, os conteúdos oriundos das BDs são dinâmicos porque podem mudar seu estado sem a intervenção direta de um programador no código-fonte do site em que tal conteúdo é apresentado. É a dinamicidade que possibilita a característica da automatização. Ela também vai permitir a legitimação das outras categorias;

b) **automatização:** ocorre quando os dados são manipulados de forma automática pela máquina, ou seja, quando não há a necessidade da intervenção humana direta para que ocorra uma mudança de estado. Há três tipos básicos de automatização: a parcial (aplicada apenas a algumas etapas do processo de produção jornalística), a procedimental (quando mais etapas do processo jornalístico ocorrem de forma automatizada) e a total (quando o produto jornalístico funciona de forma totalmente automatizada). A automatização permite que os jornalistas poupem tempo em atividades repetitivas e se dediquem à produção intelectual e analítica;

c) flexibilidade: a tecnologia das BDs traz certas facilidades à produção jornalística, pois assegura maior agilidade, qualidade e flexibilidade à produção. Com elas, os sistemas de

apuração se tornam menos hierarquizados, os conteúdos são mais facilmente recuperados e o trabalho dos jornalistas se torna mais autônomo e descentralizado, já que podem produzir e publicar de qualquer lugar com acesso à rede;

d) inter-relacionamento/Hiperlinkagem: considerado pela autora como um dos grandes potenciais das BDs, é a "capacidade de identificar padrões combinatórios e inter-relacionamentos diversos entre as informações" (BARBOSA, 2007, p. 238). A tecnologia tem o poder de vasculhar rapidamente grandes quantidades de dados e identificar quais deles podem ser inter-relacionados, de acordo com o contexto;

e) densidade informativa: é a quantidade de informações presente em um conteúdo. Geralmente, uma notícia é inicialmente apresentada com uma baixa densidade, mas com o decorrer do tempo a densidade é elevada com a inserção de novas informações, na medida em que a notícia é complementada, alterada, corrigida, contextualizada ou aprofundada. Essa característica é baseada no conceito de **resolução semântica**, apresentado por Fidalgo (2004), que usa como metáfora o conceito de resolução já utilizado para se referir a imagens digitais formadas por mais *pixels* por polegada (maior resolução) ou menos *pixels* por polegada (menor resolução). Um produto jornalístico que obtém dados de diversas fontes terá uma densidade informativa maior;

f) diversidade temática: também relacionada ao conceito de resolução semântica, a categoria representa a diversidade de tematizações além das mais comuns (como política, economia, esportes, cultura, ciência, saúde e tecnologia);

g) visualização: são as diferentes maneiras de se representar na tela as informações jornalísticas armazenadas nas BDs. Nesta categoria, deve-se considerar as noções de metadados, de *data mining* e da *tree map*, esta a responsável pela geração de visualizações típicas da estética de base de dados, como o *Squarified*, um tipo de interface que apresenta manchetes em retângulos com dimensões que se alteram de acordo com a popularidade das notícias (ver exemplo na Figura 4);

Figura 4 – Tela do site OurSignal, que reúne publicações de diversos sites e os apresenta em retângulos[20]

Um ano após a publicação da pesquisa, a autora apontou mais quatro funcionalidades das BDs no jornalismo digital:

> - Sustentar a produção e a distribuição dos conteúdos;
> - Gerenciar o fluxo de informação e o conhecimento nas redações;
> - Integrar distintas plataformas;
> - Suportar ações de interação que envolvam usuários e profissionais através do conteúdo informativo e de entretimento (reportagens investigativas associadas a informações de serviço, ou até mesmo vinculados a algum game, por exemplo) (BARBOSA, 2008a, p. 12).

Da mesma maneira que ocorreu com as funcionalidades, Barbosa integrou uma nova categoria em pesquisa posterior:

h) convergência: é tomar as bases de dados como um agente central no processo de convergência. A este processo, Barbosa deixa claro que é muito mais do que apenas a união de diversos formatos de mídia em um único produto. A convergência ocorre nos processos de produção e de distribuição, nas plataformas, no comportamento de produtores e consumidores. Segundo a autora:

> Para o jornalismo, a convergência significa integração entre meios distintos, produção de conteúdos combinando multi-plataformas para publicação e distribuição, convergência estrutural com a reorganização das redações e a introdução de novas funções para os jornalistas, uso associado de tecnologias da informação, softwares, sistemas inteligentes, audiência ativa, exploração do potencial interativo, hipertextual e multimídia da internet, e também a construção de narrativas jornalísticas em conformidade com tais recursos (BARBOSA, 2008b, p. 2)

[20] Disponível em: <http://oursignal.com>. Acesso em: 13 dez. 2011.

As categorias do JDBD abrangem diversos aspectos da produção jornalística, o que demonstra o quanto a estrutura da informação é determinada pelas bases de dados. Porém, nem sempre esta prática é referenciada na academia e no mercado como "Jornalismo Digital em Base de Dados". Em alguns países, como EUA e Inglaterra, há jornalistas, empresas jornalísticas e pesquisadores que tratam deste tema e utilizam outros termos para se referenciarem a tal prática. Barbosa (2011), em entrevista para um *blog* especializado em jornalismo digital, diz que há outros termos que, para ela, estão no "escopo que abarca o Jornalismo Digital em Base de Dados" (2011, *online*). A autora cita os termos *data driven journalism* e *data journalism*. Em pesquisas livres na web realizadas pelo autor do presente trabalho, outros termos também surgem, como *database journalism* e *data visualization*; este último utilizado não apenas por jornalistas, mas também por *designers* ou cientistas da computação que trabalham com a visualização iconográfica dos dados, com o intuito de facilitar a obtenção de informações relevantes de grandes coleções de dados.

1.3 Jornalismo de dados

Atualmente, é possível encontrar organizações e profissionais da área da comunicação que experimentam articular coleções de dados estruturados a narrativas jornalísticas através da exploração de recursos computacionais. Grande parte destas organizações trata desta prática como "*data journalism*", traduzido neste trabalho como jornalismo de dados.

1.3.1 Conceito de *data journalism*

Dos resultados obtidos na presente investigação sobre *data journalism*, foram encontrados diversos materiais produzidos por organizações jornalísticas que tomaram a dianteira na integração de produtos jornalísticos com tecnologias da informação. Uma destas organizações é o jornal britânico The Guardian, que possui uma equipe com profissionais que se autodenominam *data journalists*; também mantém blogs sobre o tema e ainda apoiou a publicação de um livro sobre o tema. No livro *Facts are sacred: the power of data*, o jornalista Simon Rogers, do The Guardian, apresenta algumas discussões sobre o que é *data journalism*. Entre vários conceitos, Rogers (2011) afirma que a prática é uma forma de se obter histórias interessantes a partir de coleções de dados que, em seu estado bruto, não parecem contar história alguma. Para ele, a prática não é nova, mas a diferença é que agora há o auxílio de computadores e, não menos importante, de dados estruturados em planilhas ou

outros arquivos formatados de uma maneira que as máquinas consigam manipular estes dados. Ainda que os computadores realizem processamentos automatizados, Rogers deixa claro que o bom jornalismo de dados depende das habilidades e competências de um bom profissional jornalista.

> Você pode se tornar um programador de gabarito se quiser. Mas o maior trabalho é muito mais pensar sobre os dados como um jornalista do que como um analista. O que há de interessante sobre esses números? O que há de novo? O que acontece se eu mesclar isso tudo com outras coisas? Responder a estas questões é mais importante do que qualquer outra coisa (ROGERS, 2011, edição para Kindle, *location* 82-1637) [21].

Para o The Guardian, o profissional que trabalha com *data journalism* é um *data journalist*. No site *Data Blog*[22], o jornal publica postagens de jornalistas especializados em processar dados "crus" (*raw data*) para obtenção de informações relevantes ou apresentação dos mesmos em formatos mais interessantes visualmente, como infográficos e tabelas. Um desses jornalistas, Paul Bradshaw, publicou no referido *blog* um artigo em que ensina aos leitores como ser um *data journalist*. Para isso, Bradshaw (2010) apresenta um processo de quatro passos básicos:

1º) **Encontrar os dados** (*finding data*): é uma ação que, dependendo da situação, exige desde conhecimentos para a operação de técnicas típicas da técnica CAR (*Computer Assisted Reporting*) até conhecimentos mais específicos, como a mineração de dados com o uso das linguagens MySQL ou Python.

2º) **Interrogar os dados** (*interrogating data*): uma operação que demanda do jornalista um bom conhecimento do contexto em que os dados estão inseridos e, também, de estatísticas, em que planilhas eletrônicas podem auxiliar.

3º) **Visualizar os dados** (*visualising data*): visualizar e combinar dados costuma ser uma operação realizada por designers e programadores, porém muitos jornalistas já começam a explorar essa operação devido à quebra de barreiras técnicas que permitem experimentar tais operações e ao fato dos jornalistas terem consciência das possibilidades que têm em mãos.

---

[21] *You can become a top coder if you want. But the bigger task is to think about data like a journalist, rather than na analyst. What's intersting about these numbers? What's new? What would happen if I mashed it up with something else? Andwering those questions is more importante than something else.*

[22] Disponível em: <http://www.guardian.co.uk/news/datablog>. Acesos em: 13 nov 2011.

4º) **Combinar dados** (*mashing data*): muitas ferramentas para combinação e visualização de dados estão disponíveis hoje na web para estudantes e jornalistas, tais como o Many Eyes e o Yahoo Pipes[23].

Assim como o The Guardian, há outras empresas jornalísticas que também rumam para o desenvolvimento de equipes especializadas em *data journalism* e publicam sites dedicados ao tema. Além do já citado *Data Blog* do The Guardian, podemos citar o *Data Desk*[24], do Los Angeles Times, em que são apresentados produtos jornalísticos baseados em BDs, como infografias interativas em base de dados. Outro jornal influente com iniciativas semelhantes é o The New York Times que publica dois sites especializados: o *blog Open*[25], escrito pela equipe de programadores e desenvolvedores, com debates sobre questões relacionadas a jornalismo e computação, e o site *Linked Open Data*[26], em que são disponibilizadas coleções de dados estruturados para uso em aplicações da Web Semântica (*linked data*) e abertos para o livre uso por parte dos usuários (*open data*).

Até aqui, o jornalismo de dados parece ser um termo aplicado ao processo de apuração jornalística em coleções de dados estruturados. Porém, não é só nas rotinas de produção jornalística que o termo se aplica: também são utilizadas técnicas de gerenciamento dos dados na apresentação dos produtos jornalísticos. Algumas organizações jornalísticas, como o The Guardian e a BBC, costumam integrar jornalistas, programadores e *designers* em operações de busca, combinação e apresentação dos dados em produtos multimídia interativos. A seguir, tratamos sobre a aplicação das bases de dados na estruturação visual dos produtos jornalísticos.

1.3.2 Visualização de dados

As funções das bases de dados alcançam os diversos aspectos do produtos jornalístico. Desde as rotinas produtivas até o consumo. Destes aspectos, a apresentação visual do produto é uma das mais impactadas pelas funções das BDs, justamente porque é nela que são materializadas as experimentações de novas formas de narrativa nos produtos jornalísticos, além de ser também a etapa em que entram em jogo as características da multimidialidade e da interatividade. A visualização ocorre através de uma interface que pode ser construída com

[23] Disponível em: <http://pipes.yahoo.com/pipes/>. Acesso em: 15 nov 2011.
[24] Disponível em: <http://projects.latimes.com/index/>. Acesso em: 15 nov 2011.
[25] Disponível em: <http://open.blogs.nytimes.com/>. Acesso em: 15 nov 2011.
[26] Disponível em: <http://data.nytimes.com/>. Acesso em: 15 nov 2011.

elementos hipertextuais, interativos e multimídia, e ainda alimentada de forma dinâmica e automatizada pelas BDs, caracterizando assim uma interface que porta uma estética de base de dados típica da cultura dos computadores. São interfaces que apresentam elementos típicos como links, listas dinâmicas, *rankings* de mais lidos ou acessados, convergência de formatos de mídia, menus interativos, caixas com colaborações de usuários, entre outros.

Entre as possibilidades de narrativas jornalísticas baseadas em BDs, podemos citar as infografias, que são elementos jornalísticos que unem grafismos (imagens, fotografias, ilustrações, mapas, símbolos etc) e informações textuais, e que geralmente são utilizadas como complemento, contextualização ou auxílio na compreensão de matérias jornalísticas. Embora as infografias já existam no suporte impresso desde muito antes dos computadores, a computação, a internet e as BDs agiram como agentes remediadores[27] na evolução deste recurso. Em estudo sobre esse tipo particular de narrativa, a pesquisadora Adriana Rodrigues desenvolveu em sua dissertação uma investigação sobre diferentes tipos de infografias interativas em bases de dados. Para a autora:

> A infografia interativa em base de dados conduz, entre outros fatores, a uma redefinição do próprio conceito de infografia. Entendemos por infografia em base de dados, como o nome sugere, aquelas produzidas tendo como mola propulsora o cruzamento ou inserção das bases de dados nas suas produções, e cujo nível de complexidade se eleva, pois pode requerer do usuário uma interpretação, uma análise mais aprofundada com níveis de interatividade maior, a depender de cada gráfico, funcionando como um mecanismo de exploração da informação (RODRIGUES, 2009, p. 37).

Antes de se integrarem às BDs, as infografias já tinham sido potencializadas pela web com o uso de recursos multimídia e interativos, como as animações em Flash e os links do hipertexto; no entanto, as BDs possibilitaram novas aplicações a esses recursos, como no caso do processamento e visualização instantâneos de grandes quantidades de dados ou a possibilidade do usuário interferir na visualização, como, por exemplo, ao inserir dados em campos de formulário e, a partir disso, a infografia alterar a visualização de acordo com as coordenadas inseridas pelo mesmo. Segundo Rodrigues (2009), a essas possibilidades Manovich utiliza o termo **visualização dinâmica de dados**.

---

[27] Barbosa apresenta o conceito de *remediation* segundo os autores Bolter & Grusin: "implica o reconhecimento do meio anterior, da sua linguagem e da sua representação social. Significa dizer que todos os meios têm o seu sistema de produção afetado pela chamada nova mídia, que, por outro lado, também possibilita algumas rupturas. [...] De acordo com os autores norte-americanos, as novas mídias remediam, melhoram seus predecessores [...]. A internet, por sua vez, remedia todos os meios, melhorando-os em muitos aspectos e acrescentando recursos novos" (BARBOSA, 2005, p. 1315-1316).

Ainda em relação às infografias, as BDs oferecem recursos necessários para a combinação de diferentes tipos de dados em um mesmo plano visual, seja ele em 2D ou 3D. A estrutura da internet e da web permitem ainda que tais combinações possam ser realizadas a partir de dados oriundos de fontes diferentes, como, por exemplo, de mais de um site ou serviço online. A essa possibilidade de combinação, Manovich denomina **remixabilidade** e caracteriza o momento atual como de "profunda remixabilidade" (RODRIGUES, 2009). Como exemplo, podemos citar as infografias que mostram textos informativos ou sinais visuais combinados com mapas do serviço Google Maps. Na Figura 5, Rodrigues apresenta um exemplo de infografia que identifica de forma georreferenciada as ocorrências de homicídios na cidade de Los Angeles. Na coluna à esquerda, o usuário pode selecionar os filtros desejados e, à direita, são mostrados indicadores visuais e textuais em uma camada acima do mapa gerado pelo serviço Google Maps.

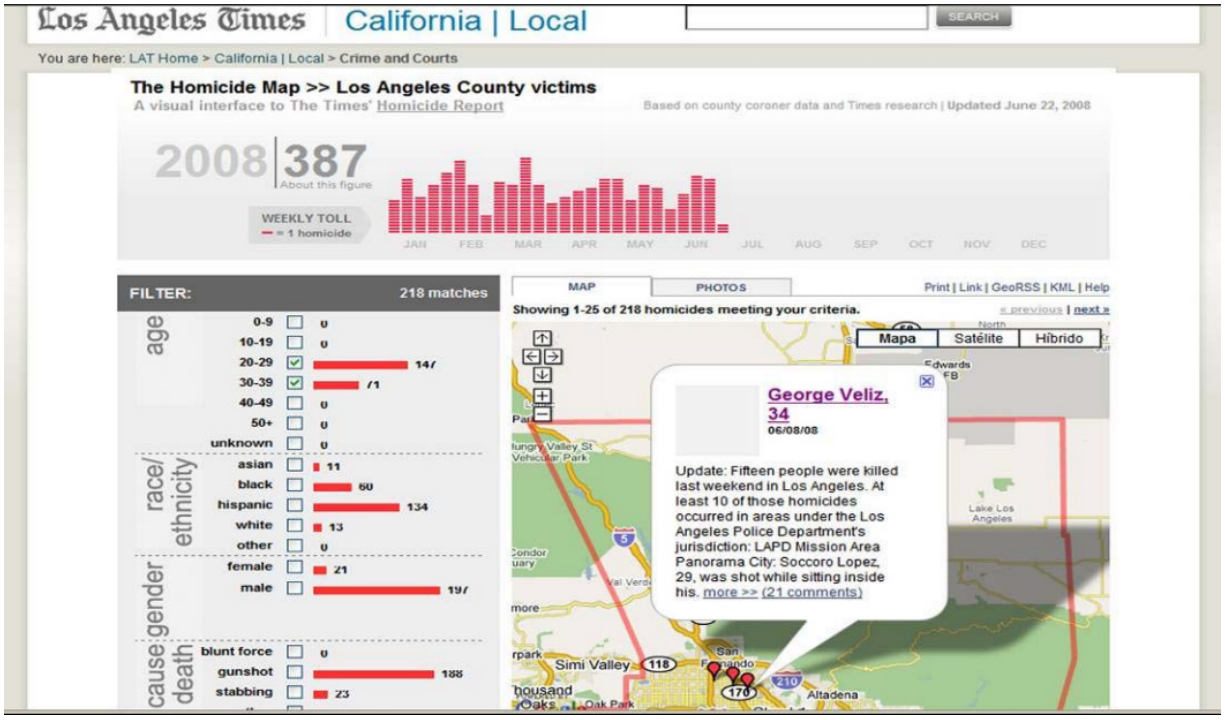


Figura 5 – "Infografia em base de dados do Los Angeles Times sobre a ocorrência dos homicídios" (RODRIGUES, 2009, p. 44)

Embora o impacto visual seja uma característica dos infográficos da web, a autora ressalta que mais importante do que este impacto é a organização e a clareza dos dados ali representados. Por isso, é importante a reflexão sobre a forma como os dados são organizados na BD e como são recuperados na infografia. A infografia deve estar estruturada como um mapa, como um esquema de navegação ao usuário, mas mantendo possibilidades de

navegação (não-linearidade), isto é, o cruzamento entre os dados. Após a análise de 23 infografias oriundas de nove jornais digitais, Rodrigues conclui que a infografia interativa em base de dados promove uma "ruptura qualitativa com relação aos modelos estáticos de narrar o fato infograficamente" (2009, p. 106).

Na web, é possível encontrar iniciativas de experimentações com infografias interativas em base de dados. Um dos projetos frequentemente citados (RODRIGUES, 2009; BARBOSA, 2007) é o Many Eyes[28], da IBM. No site do projeto, em que qualquer usuário pode criar sua visualização ou explorar visualizações criadas por outros usuários, são apresentadas várias formas de visualização para uma mesma coleção de dados. Entre estas formas de visualização, que geralmente são interativas e dinâmicas, encontramos mapas, taxonomias (*word tree*), gráficos em barra (*bar chart*), gráficos em pizza (*pie chart*), gráficos em bolhas (*bubble chart*), diagramas em rede (*network diagram*) entre outros. Fernanda Viégas, cientista brasileira que faz parte do projeto, explica a importância da visualização para a compreensão humana:

> Basicamente, metade de nosso cérebro é um hardware para a visão. Visão é a maior largura de banda que nós temos, em termos de informação sensorial sobre o mundo exterior. Então a visualização significa aproveitar o fato de que nós somos tão programados para entender o mundo a nossa volta através do que nós enxergamos [29] (VIÉGAS, 2010, *online*).

Se a infografia digital em bases de dados apresentam uma maior complexidade técnica em relação aos modelos estáticos, mais complexos ainda podem ser os *software* aplicativos.

1.3.3 Aplicativos jornalísticos

Alberto Cairo, que já foi responsável pelas infografias do jornal El País e diretor de infografia e multimídia da Editora Globo, ao tratar sobre o impacto da interatividade na visualização de informações jornalísticas, acredita que a complexidade dos infográficos podem alcançar o aprimoramento técnico do *software* aplicativo:

> Adicionar interatividade, mesmo em pequenas quantidades, significa assumir um novo paradigma: **compreender os gráficos on-line como *ferramentas de software*, e não como apresentações estáticas; o leitor se transforma em *usuário* e a**

[28] Disponível em: <http://www-958.ibm.com/software/data/cognos/manyeyes/>. Acesso em: 13 nov 2011.
[29] *Basically, half our brain is hardwared for vision. Vision is the biggest bandwidth that we have, in terms of sensory information to the outside world. So visualization is taking advantage of the fact that we are so programmed to understand the world around us in terms of what we see.*

> **infografia, em *aplicativo***. Esta pequena mudança de esquema mental ajuda a entender melhor o caminho a seguir: em um mundo onde o software está ao mesmo tempo se tornando cada vez mais sofisticado e fácil de usar, as expectativas de qualidade e de capacidade de controle sobre os programas do leitor/usuário são incrementadas. Como jornalistas, devemos atender a estas exigências[30] (CAIRO, 2008, p. 4, grifo do autor).

O conceito de produto jornalístico como *software* pode ir além de uma interface mais complexa, mais interativa e que oferece maior controle. Atualmente, é possível encontrar iniciativas de organizações jornalísticas que chegam a oferecer API[31] de seus sistemas aos usuários. Um caso ilustrativo é o do jornal britânico The Guardian, que possui um site chamado *Open Plataform*[32] destinado a disponibilizar serviços que permitem aos usuários criarem aplicativos com os conteúdos jornalísticos armazenados nas bases de dados do jornal. Entre os serviços disponibilizados, se destacam: o *Content API*, que é um mecanismo que permite ao usuário selecionar e coletar conteúdos do jornal (aproximadamente um milhão de artigos desde 1999, além de imagens, vídeos e *tags*), e o *Data Store*, um diretório de coleções de dados já estruturados para serem utilizados por aplicativos, como, por exemplo, em formato de planilha. Além destes dois serviços, o site *Open Plataform* ainda apresenta uma galeria de aplicativos desenvolvidos por usuários da web que utilizaram os serviços do referido site.

O conceito apresentado por Cairo sobre a infografia interativa como *software* aplicativo é significativo porque nos leva à reflexão sobre uma possível tendência da produção de conteúdos jornalísticos em formato de *software*. Manovich (2008) defende que o *software* é o elemento que caracteriza a "sociedade da informação global", assim como a eletricidade e o motor a combustão tornaram possível a sociedade industrial. De acordo com o autor, os principais *players* que fazem a economia da sociedade da informação, tais como os "trabalhadores do conhecimento", os "analistas de símbolos" e as "indústrias criativas", só existem porque o *software* permite. Para Manovich, o *software* é o centro das atividades

---

[30] *Añadir interactividad, aun en cantidades pequeñas, implica asumir un nuevo paradigma: comprender los gráficos online como herramientas de software, y no como presentaciones estáticas; el lector se transforma en usuario y la infografía, en aplicación. Este pequeño cambio de esquema mental ayuda a entender mejor hacia dónde avanzar: en un mundo en el que el software se hace cada día más sofisticado y sencillo de usar al mismo tiempo, las expectativas de calidad y capacidad de control sobre los programas del lector/usuario se incrementan. Como periodistas, debemos satisfacer estas exigencias.*

[31] API é a sigla para "*Application Programming Interface*" (interface para programação de aplicação). É um recurso utilizado para que diferentes aplicativos ou serviços se comuniquem entre si. Através da API, os desenvolvedores podem manipular os dados dos respectivos serviços ou sites e, então, desenvolver *mashups* ou aplicativos específicos para o serviço/site em questão. Como exemplo, podemos citar os diversos *software* aplicativos independentes utilizados para a publicação de mensagens do site twitter.com: tais aplicativos só estão aptos a acessarem os dados do Twitter porque este disponibiliza uma API para os desenvolvedores.

[32] Disponível em: <http://www.guardian.co.uk/open-platform>. Acesso em: 26 nov. 2011.

globais nas áreas da economia, cultura, vida social e, cada vez mais, da política. Por isso, o autor utiliza o termo *cultural software*; cultural no sentido de que o *software* é usado "por milhares de milhões de pessoas e que ele carrega 'átomos' de cultura (mídia e informação, além das interações ao redor dessas mídias e informações)" [33] (MANOVICH, 2008, p. 3). Embora o autor cite, principalmente, os *software* aplicativos utilizados para produção de conteúdos, tais como Microsoft Word, Adobe Photoshop ou Adobe Flash, ele também considera o próprio conteúdo midiático um *software*, já que "as próprias interfaces das mídias - ícones, pastas, sons, animações e interações do usuário - são também *software* cultural, já que estas interfaces mediam as interações das pessoas com mídias e outras pessoas[34]" (MANOVICH, 2008, p. 13).

O crescimento exponencial na venda de aparelhos smartphones[35] nos últimos anos ajudou a proliferar os aplicativos para aparelhos móveis (conhecidos como *apps*); e entre eles, estão os aplicativos jornalísticos (WANGLON, 2010), geralmente utilizados para distribuir notícias de um jornal específico ou para agregar notícias de vários jornais. Além dos aplicativos nativos (que funcionam em um sistema operacional específico), a W3C defende que o futuro lançamento da HTML5 (atualização da atual versão da HTML) deverá atribuir às páginas da web algumas das características dos programas de computador (LAMMEL, 2010), tornando-as aplicativos compatíveis com diferentes sistemas e plataformas. Hoje, há organizações jornalísticas que já desenvolvem produtos com esta tecnologia, tais como a The Economist[36], a BBC[37] e a Folha de S. Paulo[38].

Essa aproximação do jornalismo ao conceito de *software* demonstra um movimento dos tradicionais documentos hipertextuais da web para produtos mais complexos, que utilizam dados de forma mais intensa. Percebemos que os produtos digitais jornalísticos se enveredam por caminhos que os tornam geradores e consumidores de dados, pois tanto os *software* aplicativos quanto os produtos da web, na concepção do *data journalism*, geram e

---

[33] […] *cultural in a sense that it is directly used by hundreds of millions of people and that it carries "atoms" of culture (media and information, as well as human interactions around these media and information)* […].

[34] Moreover, the media interfaces themselves – icons, folders, sounds, animations, and user interactions - are also cultural software, since these interface mediate people's interactions with media and other people.

[35] Somente no terceiro trimestre de 2011, houve crescimento de 42% na venda de *smartphones* no mundo (GARTNER, 2011).

[36] "The Economist explains its Electionism HTML5 app for iPad and Android". Disponível em: <http://www.guardian.co.uk/technology/appsblog/2012/jan/18/economist-electionism-html5-tablet-app>. Acesso em: 07 fev. 2012.

[37] "BBC switches to HTML5 for mobile News vídeo". Disponível em: <http://www.zdnet.co.uk/blogs/communication-breakdown-10000030/bbc-switches-to-html5-for-mobile-news-video-10025070/>. Acesso em: 07 fev. 2012.

[38] "Folha lança novo aplicativo para tablets e smartphones em HTML5". Disponível em: <http://www1.folha.uol.com.br/mercado/1022054-folha-lanca-novo-aplicativo-para-tablets-e-smartphones-em-html5.shtml>. Acesso em: 07 fev. 2012.

demandam uma alimentação constante de dados estruturados. Tal situação confirma o que já foi predito por Barbosa: que o cenário em que emerge uma quarta geração do jornalismo digital se caracteriza pela "consolidação das bases de dados como estruturantes da atividade jornalística e como agentes singulares no processo de convergência jornalística" (BARBOSA, 2008a, p. 9).

Ao mesmo tempo em que se consolida um ambiente tecnológico cada vez mais dependente das bases de dados, percebe-se que o principal sistema de armazenamento de dados da atualidade, a World Wide Web, ainda mantém na sua essência a mesma lógica de funcionamento idealizada em sua origem: a de um repositório de documentos hipertextuais. Embora diversas tecnologias tenham surgido no decorrer dos anos e expandido as funcionalidades da rede (tais como as linguagens de script PHP, ASP e JavaScritp, as folhas de estilo CSS, a plataforma multimídia Flash, a linguagem de marcação XML e as próprias bases de dados relacionais), a web ainda demonstra limitações técnicas quando a questão é a integração das diferentes BDs com os dados não estruturados e em formatos não padronizados, como é o caso dos documentos hipertextuais. Em outras palavras, surgem dúvidas sobre como tantos sites, *software* aplicativos e infográficos interativos podem aproveitar a imensa quantidade de dados e informações armazenadas na web ao longo de mais de 20 anos, pois grande parte destes conteúdos está "enclausurada" dentro de documentos ou de diferentes bases de dados que não se comunicam entre si. O modelo relacional de BD não foi projetado para resolver esta questão. Segundo Mike Loukides, vice-presidente de estratégias de conteúdo da O'Reilly Media[39]:

> A maioria das organizações que construíram plataformas de dados acha que é necessário ir além do modelo relacional de base de dados. Os tradicionais sistemas de bases de dados relacionais deixaram de ser efetivos nessa escala [de quantidade de dados]. Gerenciar *sharding*[40] e replicação de uma horda de servidores de bases de dados é difícil e lento [41] (LOUKIDES, 2011, edição para Kindle, *location* 185).

O problema da grande quantidade de dados na atualidade vai muito além da velocidade de processamento. Além deste problema de ordem quantitativa, que atinge a eficiência do sistema, há também problemas de ordem qualitativa, que atinge a eficácia: como

[39] A O'Reilly Media é uma empresa especializada em livros técnicos sobre programação e desenvolvimento web. Seu fundador, Tim O'Reilly, foi a responsável por cunhar o termo "Web 2.0".
[40] *Sharding* é uma técnica de separação de tabelas de bases de dados relacionais em partes menores, permitindo a replicação destas partes entre bases de dados diferentes.
[41] *Most of the organizations that have built data platform have found it necessary to go beyond the relational database model. Traditional relational database systems stop being effective at this scale. Managing sharding and replication across a horde of database servers is difficult and slow*.

se obter melhores resultados na busca de informações significativas e no inter-relacionamento destas mesmas informações em um ambiente como a atual web, saturada de dados em diferentes formatos e muitas vezes não estruturados?

Para que as máquinas tenham a capacidade de processar e combinar quantidades tão grandes de dados, publicadas diariamente de forma esparsa entre diferentes produtos digitais, uma série de autores, empresas e profissionais, encabeçados pelo cientista Tim Berners-Lee, afirma que tais máquinas deveriam portar a capacidade de compreender o significado destes dados, para que seja possível, então, a execução de operações automatizadas de identificação, associação e combinação de dados. Essa proposta de solução tecnológica é denominada "Web Semântica". Nela, busca-se substituir a lógica de publicação de documentos pela lógica de publicação de dados (BERNERS-LEE et al, 2001), em uma estrutura padronizada entre os *sites* da web, de maneira que todos possam, então, compartilhar estes mesmos dados (pois estão estruturados em um mesmo modelo padrão), o que permite a interoperabilidade entre os diferentes produtos digitais (W3C, 2001a). Segundo Berners-Lee, essa padronização torna a web uma única "base de dados gigante" (SIEGEL, 2009, p. 6; OLAVSRUD, 2003, *online*). Além da interoperabilidade, a Web Semântica oferece recursos para que as máquinas possam "compreender" o significado das informações publicadas, permitindo, assim, que elas realizem operações automatizadas no gerenciamento dos dados. Tal cenário pode vir a contribuir nas categorias do Jornalismo Digital em Base de Dados, e essa é a preocupação central do presente trabalho. No próximo capítulo, apresentamos mais detalhadamente o conceito da Web Semântica. Abordaremos a visão original de Tim Berners-Lee, o funcionamento das tecnologias que tornam a proposta viável e alguns exemplos reais de aplicação que ilustram alguns dos benefícios desta ideia.

2 WEB SEMÂNTICA

A Web Semântica (WS) é um conceito de uma rede digital de dados estruturados de tal forma que tanto humanos quanto máquinas tenham a capacidade de identificar o significado dos dados publicados, o que permitiria o desenvolvimento de aplicações mais inteligentes, capazes de realizar determinadas operações de forma automatizada.

Ainda que já se falasse em tecnologias semânticas para a web na metade da década de 1990[42], o marco que impulsionou os debates sobre esse conceito foi a publicação de um artigo, em 2001, de autoria de Berners-Lee, junto com os autores Hendler e Lassila, em que apresentavam a WS como um passo evolutivo da atual web. Desde então, a proposta encontra-se em desenvolvimento através dos esforços da W3C e de diferentes profissionais, estudiosos, empresas e entusiastas que trabalham, principalmente, com atividades relacionadas às áreas de ciência da computação e sistemas de informação.

Embora seja possível encontrar discursos que a tratem como uma nova web, ela não é uma rede separada da atual web, mas uma extensão dela (BERNERS-LEE et al, 2002); ou seja, a WS não apenas funciona de forma agregada à web atual, como necessita do seu aporte tecnológico. A partir dessa asserção, julgamos necessário retomar uma breve apresentação da tecnologia por trás da atual web, para que possamos, mais adiante, tecer comparativos e compreender a proposta (e o diferencial) da WS.

2.1 A web atual: uma rede de documentos

A internet é uma estrutura tecnológica que permite a transmissão de dados entre redes de computadores que utilizam o mesmo protocolo de comunicação. Ela não é a interface gráfica de apresentação de tais dados, pois eles podem ser recuperados e apresentados de diferentes maneiras pelos computadores. A formatação visual-gráfica destes dados fica a cargo de outras tecnologias que funcionam a partir da estrutura da internet. Entre vários sistemas já utilizados para a recuperação e apresentação de dados na internet (tais como o correio eletrônico, o FTP e o Gopher), destaca-se o mais popular: a World Wide Web (ou simplesmente web), que é "um sistema de armazenamento, recuperação e exibição de

[42] Em 1996, ao refletir sobre o future da web no artigo "The World Wide Web: Past, Present and Future", Tim Berners-Lee afirma que as máquinas poderiam participar de processos de análises automatizados, mas, para isso, os dados publicados na web precisariam ser apresentados também em formatos interpretáveis pelas máquinas e com semânticas definidas (BERNERS-LEE, 1996, *online*).

informações que combina recursos de texto, hipermídia, imagens e som" (AUDY, 2005, p. 186). Ela foi proposta pelo físico britânico Tim Berners-Lee, entre os anos de 1989 e 1991[43], como um projeto paralelo que o cientista desenvolvia enquanto trabalhava na Organização Europeia para a Pesquisa Nuclear (CERN). Antes da web, grande parte dos sistemas que funcionavam na internet apresentava uma interface complexa para o usuário comum, como, por exemplo, a interface com linhas de comando, que poderiam exigir conhecimentos de UNIX (LEÃO, 1999). A web passou a funcionar como uma interface gráfica para a internet, que possibilitou a criação, a publicação e a visualização de documentos digitais hipertextuais e multimídia. Nesses documentos, o jornalismo encontrou um novo espaço para a distribuição de sua produção jornalística tradicional e, mais tarde, um meio para o trabalho de apuração jornalística (MACHADO, 2002).

O sistema da web foi tecnicamente viável devido à união de três recursos básicos: o HTTP[44], a URI[45] e a HTML[46] (CECCONI, 2010). Entre as três tecnologias, a linguagem de marcação HTML é a que determina as possibilidades e as limitações na apresentação das informações, pois tem como função a montagem dos documentos digitais hipertextuais. A linguagem oferece diversos códigos (conhecidos como elementos, *tags* ou etiquetas) para a formatação dos documentos publicados na web (conhecidos como páginas), o que permite a criação de conteúdos ricos em recursos visuais e multimídia. Entre estes códigos, podemos citar alguns exemplos, como o elemento <b> (de *bold*, utilizado para aplicar o efeito de negrito a um texto), o elemento
 (de *line brake*, utilizado para inserir uma quebra de linha) ou o elemento <a> (de *anchor*, utilizado para a inserção de um link) (W3C, 1999). A função do *software* navegador é a de interpretar tais códigos e, a partir disso, gerar e disponibilizar uma página digital para o usuário final.

Em um caso de publicação de um artigo na web, por exemplo, seria possível criar um documento digital com o uso de códigos HTML, em que o título do artigo poderia ser destacado com o efeito negrito (através do elemento <b>) e os subsequentes parágrafos poderiam ser delimitados espacialmente no documento com o uso da quebra de linha (através

[43] A W3C disponibiliza uma página com um breve histórico da World Wide Web, em que lista acontecimentos importantes em ordem cronológica. O desenvolvimento inicial da WWW (das primeiras anotações até a sua publicação na internet) compreende um processo de várias etapas entre os anos de 1989 e 1991. Disponível em: <http://www.w3.org/History.html>. Acesso em: 5 jun 2011.

[44] HTTP (HyperText Transfer Protocol) é um protocolo de transferência de dados entre computadores; permite que as máquinas se comuniquem utilizando "a mesma língua".

[45] URI (Uniform Resource Identifier) é um esquema único de nomes para localização de recursos da rede, como os endereços de páginas que começam com o "www".

[46] HTML (HyperText Markup Language) é a linguagem de marcação utilizada para a montagem de páginas da web. É formada por códigos padronizados (*tags*) que executam comandos de formatação ao conteúdo, como negrito e itálico, e que permitirem a inserção de hiperlinks e metadados nas páginas.

do elemento
). Entretanto, sabe-se que o efeito negrito não é necessariamente sinônimo de título, pois outros elementos do texto podem receber o negrito (como as legendas das fotos ou nomes dos autores), assim como a quebra de linha não é sinônimo de parágrafo, pois outros elementos também podem ser delimitados pela quebra de linha (como as imagens e suas respectivas legendas). Em outras palavras, a maior parte dos elementos HTML[47] geralmente não traduz o significado dos elementos que fazem parte do texto. Citamos "a maioria dos elementos" porque existem alguns deles que permitem a associação de significados. Por exemplo, o elemento <h1> (de *header*) significa título, logo poderia ser utilizado para definir o título no nosso exemplo anterior, no lugar do elemento <b>. Mesmo que o resultado final não seja evidente para a leitura humana (de uma forma ou de outra, o título ficaria visualmente destacado no documento), a vantagem dessa prática é que, neste último caso, as máquinas também poderiam compreender que aquele elemento é um título e não apenas uma parte do documento destacado com efeito negrito. Entre as inúmeras utilidades dessa situação, podemos citar o caso dos sites de busca: se o usuário deseja encontrar páginas da web utilizando determinada palavra-chave, ele poderia escolher entre: a) encontrar resultados que considerassem todo o documento, ou b) encontrar resultados que considerassem a referida palavra-chave apenas nos títulos. Logo, é vantajoso associar aos dados significados que possam ser interpretados pelas máquinas.

A HTML tem poucos elementos que indicam o significado do conteúdo. Citamos o exemplo do elemento <h1> para título, mas não poderíamos citar exemplos para elementos que identifiquem legendas de fotos, resumos ou sobrenomes de autores, porque tais elementos não existem. Embora pareça simples solucionar esse impasse com a criação de novos elementos semânticos em futuras atualizações do HTML (como <legenda>, <resumo> ou <sobrenome>), este problema se torna ainda mais complexo se considerarmos que há incontáveis possibilidades de associações semânticas além da estruturação de um documento, como, por exemplo, o reconhecimento do tipo de entidades[48] tratadas no conteúdo do texto (se é uma pessoa, um animal, um lugar, um objeto, uma empresa etc), além das características desta entidade (caso seja uma pessoa, como ela é? Quem é ela? Caso seja um lugar, onde ele fica? Qual é a língua oficial? etc). Seria inviável criar um elemento HTML para cada um destes itens. Como não há uma forma da HTML associar um significado a cada elemento

[47] Consideramos aqui a HTML 4.01, versão mais atual do código até o presente momento (dez. de 2011). A W3C está em processo de desenvolvimento de uma atualização da linguagem (HTML5), que deverá trazer novos elementos semânticos (LAMMEL, 2010).

[48] Nos estudos sobre Web Semântica, é utilizado o termo em inglês *entity* para referenciar as unidades individuais que possuem propriedades e que podem ser relacionadas. Como exemplo de entidade, podemos citar pessoas, lugares e objetos.

presente em uma página, então a interpretação dos significados destes elementos fica a cargo do usuário final, que lê tais páginas e interpreta de acordo com sua capacidade intelectual e seu repertório cultural.

Se, por um lado, o ser humano tem a capacidade de distinguir o significado dos elementos presentes em um documento através da livre interpretação do texto publicado, por outro, falta esta faculdade às máquinas (BREITMAN, 2005). Em outras palavras: a web atual é uma rede de documentos, e documentos são feitos para serem lidos por humanos e não por máquinas (BERNERS-LEE et al, 2002). Tal situação resulta em certas limitações ao sistema da web (SILVA FILHO, 2004), principalmente os relacionados às operações automatizadas e à interoperabilidade em um sistema com bilhões[49] de documentos feitos para humanos. Para que os computadores tenham a capacidade de processar o significado de tanto conteúdo, seria mais apropriado termos uma rede de dados estruturados, ao invés de documentos.

2.2 Web Semântica: uma rede de dados

No início da década de 2000, a web estava em vertiginosa expansão e já fazia parte da rotina de muitas empresas, instituições e usuários particulares. Ainda assim, Tim Berners-Lee apresentou, em um artigo escrito em 2001 com os autores James Handler e Ora Lassila[50], uma proposta de mudança na forma de publicar as informações na rede. Para ele, a web foi originalmente concebida como uma rede de documentos digitais, mas documentos são feitos para serem lidos por humanos, não por máquinas, e isso gera algumas dificuldades no processamento automatizado de dados e na interoperabilidade dos mesmos. Os autores propuseram repensar a ideia de **rede de documentos** para o conceito de **rede de dados** (BERNERS-LEE et al, 2002). A diferença entre as duas concepções é que os documentos são escritos em linguagem natural para que sejam lidos por humanos; já os dados podem ser manipulados pelas máquinas (SHADBOLT et al, 2006). Dados podem ser categorizados, classificados, filtrados, enfim, manipulados automaticamente por computadores (BERNERS-LEE, 2009). Podem ser inter-relacionados de acordo com critérios lógicos, como tamanho, formato, quantidade, igualdade, semelhança ou diferença.

---

[49] O site http://www.worldwidewebsize.com apresenta estatísticas sobre a quantidade de páginas indexadas nos principais sites de buscas. Em dezembro de 2011, o Google listava aproximadamente 50 bilhões de páginas.

[50] O artigo "The Semantic Web – A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities" foi publicado em 2001, na revista American Scientific, por Tim Berners-Lee, James Handler e Ora Lassila. O artigo pode ser encontrado em: <http://www.med.nyu.edu/research/pdf/mainim01-1484312.pdf>. Acesso em: 11 jun 2011.

Ao caracterizar a atual web como "de documentos", como se estivéssemos no início da década de 1990, quando a World Wide Web era, de fato, uma rede de documentos estáticos, Tim Berners-Lee parece generalizar um sistema que, atualmente, já se encontra em um estágio muito mais complexo. Hoje, a maior parte da organização dos dados na web é baseada em sistemas de gerenciamento de bancos de dados em detrimento das páginas HTML estáticas (KASHYAP et al, 2008). No entanto, quando Berners-Lee descreve uma rede inteligente, em que máquinas têm a capacidade de identificar os significados dos dados, o conceito é mais rico do que uma rede de dados ordenados em BDs. Para o autor, mais do que um armazenamento ordenado, a Web Semântica é a proposta de um sistema em que os dados são publicados de uma forma padronizada entre os sites, possibilitando a interoperabilidade entre eles. Além disso, outra diferença da proposta de Web Semântica é que os dados publicados não são apenas formatados para a leitura humana: eles também são formatados para a interpretação por parte das máquinas, o que possibilitaria aos computadores inferir o significado das informações publicadas. Este sistema permitiria a execução de operações automatizadas na manipulação dos dados na interoperabilidade desses dados entre diferentes sistemas (SHADBOLT et al, 2006).

Então, embora tenhamos hoje uma rede de dados que funciona de forma concomitante à rede de documentos (ou seja: BDs com páginas da web), não temos uma padronização na forma como estes dados devem ser interpretados e compartilhados pelas máquinas, e aí está o diferencial na proposta da WS. Daí o termo **semântica**, que, no dicionário de língua portuguesa, é "o estudo do significado da palavra, que explica a origem e as variações da significação vocabular" (BUENO, 1996, p. 598); já para a linguística, é "o estudo *sistemático* do sentido nas línguas naturais" (PIETROFORTE e LOPES, 2003, p. 114, grifo do autor). Seguindo a linha dos termos linguísticos, a web atual poderia ser comparada à sintaxe, pois, nas inter-relações de dados, os computadores consideram mais as construções sintáticas das palavras do que os seus significados (como ocorre no gerenciamento dos bancos de dados ou nos motores de busca). Nas palavras de Breitman, a atual web pode ser denominada sintática porque "nela os computadores fazem apenas a apresentação da informação, porém o processo de interpretação fica a cabo dos seres humanos" (2005, p. 2); e a autora ainda se pergunta: "a questão é: por que os computadores não podem realizar esse trabalho para nós?" (idem, p. 2). Poderíamos responder a Breitman que os computadores não realizam este trabalho para nós porque eles não compreendem a língua natural dos humanos. O professor de computação Akerkar endossa esta resposta:

> [...] humanos conseguem fazer uso de sua intuição para obter sentido dos documentos e processá-los adequadamente, mas a ausência de informações processáveis por máquinas para descrever o conteúdo é um enorme obstáculo para se automatizar a gerência do conhecimento presente na web [51] (AKERKAR, 2009, p. 12).

Embora a WS busque associar significados aos conteúdos publicados, ela também precisa manter o atual sistema de documentos, a fim de continuar proporcionando suporte à leitura humana. Por isso, a Web Semântica busca uma forma de publicar conteúdos que sustentem dois requisitos ao mesmo tempo: um modo compreensível aos humanos e outro compreensível às máquinas (KASHYAP et al, 2008, p. 24). Logicamente, relativizamos o termo "compreensível às máquinas", pois essa proposta não busca alcançar uma capacidade cognitiva/racional aos computadores. Para Berners-Lee, "[...] na verdade, o computador não 'compreende' qualquer destas informações, mas agora ele pode manipular os termos muito mais eficientemente, de um modo que sejam úteis e significativas para o usuário humano" [52] (BERNERS-LEE et al, 2002, p. 27).

A manipulação automatizada de dados, em um processo que também leva em consideração seus significados, diversifica as potencialidades desse atual sistema informacional que é a web. Como exemplo para ilustração, Berners-Lee (et al, 2002) apresenta um caso hipotético: eu contrato uma empresa de envio de mensagens para enviar congratulações aos meus clientes em seus respectivos aniversários. Para isso, informo à empresa as datas dos aniversários e os endereços dos meus clientes, armazenados em uma tabela de minha base de dados. Poderia ocorrer de a respectiva empresa copiar a coluna de endereços da minha tabela para a coluna de endereços da tabela de sua base de dados, para que os seus mensageiros pudessem encontrar tais clientes. Porém, ocorre que o sistema daquela empresa utiliza em sua coluna o termo "Endereço" para identificar a rua onde residem os clientes; já a minha empresa utiliza o termo "Endereço" para identificar os locais de cobrança dos clientes, e não propriamente de suas residências. Como consequência desse engano, os mensageiros acabariam se direcionando para as caixas postais dos correios e congratulando carteiros, ao invés de se encontrarem com os clientes em suas residências. Devido à impossibilidade do sistema reconhecer automaticamente a diferença de significados entre as duas colunas denominadas "Endereço", seria necessário que ocorresse uma

---

[51] *Humans can make use their intuition to make sense of the documents and process them accordingly, but the absence of machines processable information to describe the content is a huge hindrance to automating the management of knowledge present in the Web.*

[52] *The computer doesn't truly "understand" any of this information, but it can now manipulate the terms much more effectively in ways that are useful and meaningful to the human user.*

intervenção humana na manipulação desses dados. Esse caso seria uma realidade para a atual web, pois os conteúdos apresentados na rede são formatados para serem compreendidos por humanos e não por máquinas. Afinal, as atuais bases de dados são criadas de forma arbitrária, pois cada desenvolvedor escolhe por conta própria os termos que ele considera mais apropriados para identificar as colunas da BD que funcione em seu *site*.

A situação descrita no exemplo dos mensageiros foi aplicada a um caso bastante específico que envolve uma empresa; contudo, a mesma ideia poderia ser aplicada a um usuário comum da web que utiliza diversos serviços da rede que manipulam seus dados pessoais. Por exemplo, com as tecnologias da Web Semântica em uma situação de funcionamento ideal, o serviço de agenda online de um determinado usuário poderia interagir com o serviço de compras de passagens online de uma forma tal que, no momento em que o usuário fizesse uma solicitação de compra de passagem, o sistema poderia automaticamente alertá-lo que a transação não deveria ocorrer, porque no período de viagem solicitado haveria algum compromisso previamente marcado em sua agenda pessoal. Ao mesmo tempo, o sistema poderia sugerir outras datas mais apropriadas, de acordo com as informações pessoais armazenadas na agenda.

No site da W3C há uma página especial para a Web Semântica que apresenta outro exemplo hipotético utilizado para auxiliar na compreensão deste conceito:

> A Web Semântica é uma rede de dados. Existem grandes quantidades de informações que todos nós utilizamos todos os dias, e que não fazem parte da web. Eu posso ver o extrato do meu banco na web, e também as minhas fotografias, e eu posso ver minhas anotações em um calendário. Mas eu poderia ver minhas fotos em um calendário para ver o que eu estava fazendo quando eu as fotografei? Eu poderia ver as linhas do meu extrato bancário em um calendário? Por que não? Porque nós não temos uma rede de dados. Porque os dados são controlados por aplicativos, e cada aplicativo mantém tais dados para si [53] (W3C, 2001b, *online*).

Para que os diferentes aplicativos e serviços da web possam integrar suas funcionalidades, é preciso que as máquinas possam reconhecer os significados e os tipos de relacionamentos dos dados disponibilizados, através do fornecimento de metadados.

---

[53] *The Semantic Web is a web of data. There is lots of data we all use every day, and it is not part of the web. I can see my bank statements on the web, and my photographs, and I can see my appointments in a calendar. But can I see my photos in a calendar to see what I was doing when I took them? Can I see bank statement lines in a calendar? Why not? Because we don't have a web of data. Because data is controlled by applications, and each application keeps it to itself.*

2.2.1 Metadados

Se a Web Semântica é uma forma de apresentar informações[54] compreensíveis tanto para humanos quanto para máquinas, então, além da apresentação tradicional de um conteúdo ao usuário, também é necessário fornecer, ao mesmo tempo, dados **extras** especificamente para as máquinas. Esses dados, que geralmente não são mostrados na página formatada em HTML, são denominados "metadados" e são um requisito para o funcionamento da Web Semântica. O termo metadados significa "dados sobre dados". Segundo Manovich, "metadado é o que permite aos computadores 'enxergarem' e recuperarem dados, movê-los de um lugar a outro, comprimi-los e expandi-los, conectar dados com outros dados, e assim por diante" [55] (2002, p. 1). Manovich, ao relatar como os metadados auxiliam na automação do processamento de vídeos digitais pelos computadores, afirma que tal automação requer um novo formato de mídia, que inclua metadados que descrevam a semântica dos dados. Assim, pode-se perceber que os metadados também são importantes para que as máquinas interpretem o significado de mídias diversas, inclusive aquelas que não podem ser estruturadas da mesma forma que um texto escrito, como é o caso do audiovisual.

Em outras palavras, enquanto a web tradicional é formada principalmente por informações compreensíveis aos humanos, a WS é formada por estas mesmas informações, porém associadas a metadados interpretáveis pelas máquinas que descrevem a elas o que está sendo mostrado ao usuário. A partir deste conceito, Kashyap et al (2008, p. 25) apresenta a seguinte fórmula para descrever o tipo de conteúdo que forma a Web Semântica: "*Semantic Web Content = Data + Metadata*".

Para a construção das páginas na atual web, a linguagem HTML oferece alguns elementos de metadados. Por exemplo, com o elemento <meta>, é possível identificar o autor de uma página, ou a descrição do documento ou ainda as palavras-chaves relacionadas ao conteúdo (W3C, 1999). Esses metadados auxiliam os motores de busca a identificarem o conteúdo das páginas, além de ajudarem os navegadores em determinadas funções. Entretanto, há poucos destes elementos HTML, por isso são metadados limitados (HEBELER

[54] Para o entendimento completo do conceito de metadados, é importante esclarecer a acepção dos termos dado e informação. Entendemos que "o dado consiste em um fato bruto (nome de um funcionário, número de matrícula de um aluno, código de um produto etc.) ou suas representações (imagens, sons, números, etc.) que podem ou não ser úteis ou pertinentes para um processo particular" (AUDY et al, 2005, p. 93). Já "informação é uma coleção de fatos organizados de forma a possuir um valor adicional aos fatos em si. Em outras palavras, são dados concatenados, que passaram por um processo de transformação, cuja forma e conteúdo são apropriados para um uso específico" (idem, 2005, p. 93).

[55] *Metadata is what allows computers to "see" and retrieve data, move it from place to place, compress it and expand it, connect data with other data, and so on.*

et al, 2009). Na WS, os metadados são mais complexos e diversos e não são associados apenas ao documento inteiro, mas, também, aos dados presentes neste documento. Mais do que indicar as partes de um texto (como título, autor etc), na WS os metadados podem identificar os significados das informações publicadas. Como exemplo hipotético, ao publicarmos a frase: "Tobby está com raiva" em uma página HTML, uma possível linguagem de marcação semântica poderia permitir uma descrição com a seguinte estrutura: "<cachorro>Tobby</cachorro> está com raiva". Desta maneira, ao procurarmos pelo termo "Tobby" em um site de busca, poderíamos, por exemplo, indicar à máquina que tal busca deve ser apenas sobre cachorros, evitando, assim, a apresentação de resultados irrelevantes, como pessoas com o apelido Tobby ou personagens de desenho animado com o mesmo nome, entre outros. O mesmo ocorreria para o termo "raiva": a palavra se refere ao sentimento ou à doença? Enquanto os humanos obtêm o significado para o termo a partir do contexto em que a frase se encontra, as máquinas obteriam o significado através do processamento dos metadados associados ao termo em questão.

2.2.2 Um modelo padronizado para os metadados: o padrão RDF

Seguindo o exemplo anterior, embora a estratégia de se utilizar um elemento HTML <cachorro> pareça uma proposta funcional para a incorporação de semântica ao conteúdo em HTML, essa não é uma solução viável, porque não existem elementos HTML para cada uma das possíveis propriedades que um termo pode ter. Na medida em que associamos mais características a Tobby, mais elementos HTML seriam necessários, como, por exemplo: <idade>, <raça>, <cor> etc. Porém, a HTML é uma linguagem de marcação que, por padrão, possui uma quantidade limitada de *tags*. Essa situação se tornou um problema para os desenvolvedores.

> A quantidade de usuários a descobrir e utilizar a Web tem crescido quase que exponencialmente desde início da década de 1990, quando ela começou a tornar-se popular. Paralelamente ao crescimento de navegadores, têm surgido novas aplicações, e isso demanda mais e mais recursos da linguagem HTML, que tem sido empregada, popularmente, para a editoração de páginas para a Web. Como resultado, as limitações da linguagem têm sido evidenciadas, causando frustração àqueles que elaboram documentos para Web e motivando a necessidade de extensões (SILVA FILHO, 2004, p. 50).

Uma solução encontrada pela W3C para resolver a limitação da HTML foi o desenvolvimento de uma nova linguagem de marcação semelhante, mas que permite ao

desenvolvedor criar suas próprias *tags*. Essa linguagem é a XML (eXtensible Markup Language), uma "linguagem de editoração que oferece um formato universal para a estruturação de documentos e dados na Web" (SILVA FILHO, 2004, p. 6). Ao contrário da HTML, a XML é utilizada apenas para se estruturar o conteúdo, por isso não possui recursos para alterar as características gráficas do mesmo. Logo, com a XML é possível, por exemplo, criar elementos personalizados para conteúdos (ex.: título, autor, subtítulo, parágrafo etc), mas não há a possibilidade de determinar o tipo e o tamanho da letra, a cor de fundo etc. Nem seria preciso, pois para isso existe outra tecnologia: o CSS[56]. Com a XML, é possível, por exemplo, criar elementos como <titulo>, <autor>, <legenda> e <foto>, aplicando-se desta forma valores semânticos à estrutura da página. Entretanto, ainda assim, a técnica de se criar uma etiqueta para cada possível propriedade seria insustentável, pois, ao se permitir que os desenvolvedores criem suas etiquetas arbitrariamente, não haveria uma padronização universal de propriedades e descrições. Sem padronização, alguns desenvolvedores criariam o elemento <cachorro>, outros criariam <cao>, ou <dog>, ou <canino> e assim por diante. Dessa forma, ao buscarmos pelo termo 'Tobby' em um serviço de busca, e ao delimitarmos que os resultados devem ser obrigatoriamente relacionados ao conceito de "cachorro", o sistema não saberia quais dos elementos HTML citados anteriormente deveriam ser considerados.

A proposta da Web Semântica oferece uma solução com uma lógica diferente para esse impasse. Ao invés de se criar um elemento HTML para cada propriedade na linguagem de marcação, foi proposto um modelo de dados padronizado que permite a associação dos dados presentes na página a coleções de propriedades externas à página da web (chamados "vocabulários"). A vantagem deste modelo é que tais vocabulários podem ser compartilhados na web para que outros sites também os utilizem como referência na associação de significados. Ainda seguindo o exemplo anterior, em vez de se criar o elemento <cachorro> (ou <cao>, ou <dog> etc.), bastaria associar o termo "Tobby" ao conceito de cachorro presente em um vocabulário compartilhado por todos os sites da WS.

Este modelo de dados funciona em uma lógica padronizada chamada "triplas" (*triples*), pois trabalha sempre com uma estrutura que relaciona três unidades: um *sujeito*, um *predicado* e um *objeto* (KASHYAP et al, 2008). Cada tripla forma uma frase com um sentido (*statement*), logo, para se associar significados a um termo, bastaria tomar tal termo como o sujeito da tripla e então construir uma ou mais frases que o descrevam (ver Figura 6).

[56] O *Cascading Style Sheets* (CSS) é uma linguagem simples, sugerida pelo W3C, utilizada para definir os elementos visuais de uma página da Web. Com o CSS, é possível, por exemplo, definir o tamanho de uma caixa, a sua posição na página e a sua cor de fundo e da borda.

Figura 6 – Estrutura da tripla

Como exemplo, vamos considerar um blog da internet que denominaremos aqui como "blog X". Para descrevermos qual é a autoria do blog, criaríamos a frase: **o blog X é escrito por João**. Nesse caso, o sujeito seria "blog X", o predicado seria "escrito por" e o objeto seria "João" (ver Figura 7).

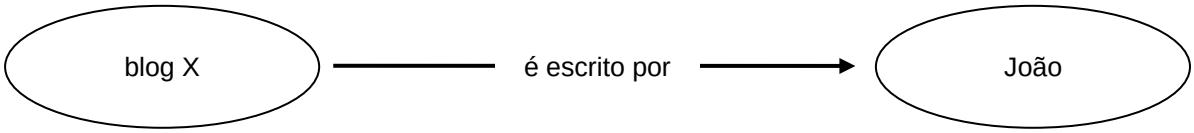


Figura 7 – Exemplo de tripla

Em outras palavras, a tripla cria relações entre entidades, como pessoas, lugares, instituições ou objetos, e o tipo de relação é definido pelo predicado, que também é conhecido pelos termos "verbo" (BERNERS-LEE et al, 2002) ou "propriedade" (AKERKAR, 2009). A proposta tecnológica para a Web Semântica seria bastante limitada caso permitisse que as associações semânticas fossem expressas apenas em triplas isoladas. Por isso, a lógica das triplas permite que elas sejam associadas entre si, formando, assim, redes de triplas, conhecidas como grafos (*graphs*). Na Figura 8, é possível observar um exemplo de grafo, formado pela associação de um sujeito a dois objetos, inter-relacionados através de predicados diferentes. Neste exemplo, o grafo indica que o **blog X** tem como autor o João e, também, indica que o site foi publicado no ano de 2011.

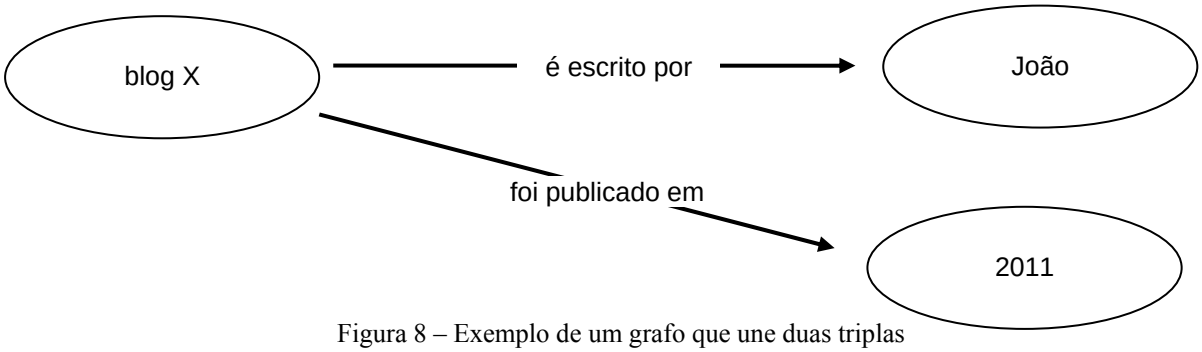


Figura 8 – Exemplo de um grafo que une duas triplas

Sob outro ponto de vista, o grafo da Figura 8 pode ser considerado a união das duas triplas citadas abaixo:

- blog X -> é escrito por -> João
- blog X -> foi publicado em -> 2011

Segundo Segaran et al (2009), diferentes grafos podem ser combinados ou separados, pois as triplas continuam mantendo os seus significados após a separação. Na Figura 9, os autores apresentam um exemplo de grafo com maior complexidade, em que é possível identificar as seguintes triplas:

- São Francisco -> tem o prefeito -> Gavin
- São Francisco -> tem população -> 774.000
- São Francisco -> está em -> Califórnia
- Califórnia -> está em -> Estados Unidos
- São Francisco -> é localizado na longitude -> -122.4183
- São Francisco -> é localizado na latitude -> 37.775

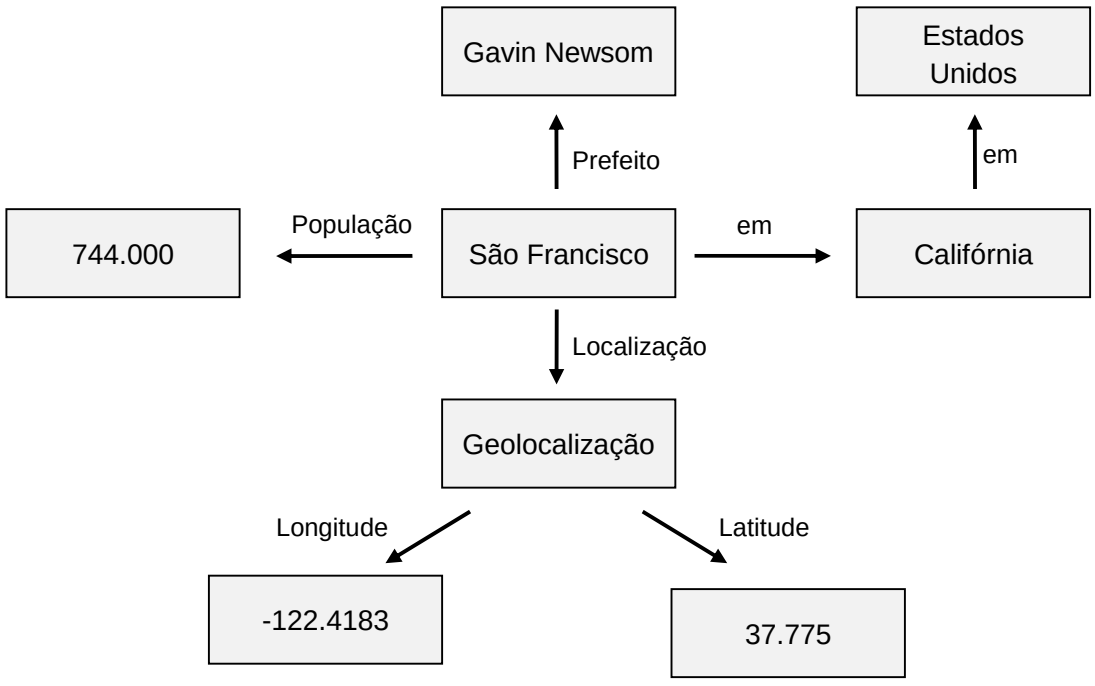


Figura 9 – Exemplo de grafo mais complexo. Adaptada de Segaran (et al, 2009, p. 30)

Para Tim Berners-Lee (2007), o inter-relacionamento dos grafos presentes na web formaria uma grande rede de grafos, o que seria o modelo ideal de organização de dados na

WS, assim como a web atual tem como modelo os documentos vinculados. Em uma comparação com as três grandes redes digitais desde a internet até a WS, Berners-Lee associa uma palavra para cada uma das redes: a internet é uma rede de computadores, a web é uma rede de documentos e a Web Semântica é uma rede de grafos. Seguindo essa linha, o autor desenvolve um jogo de siglas para facilitar a identificação de cada rede: a internet seria a "III", ou seja, *International Information Infrastructure*; a web manteria a sigla "WWW" (*World Wide Web*) e a Web Semântica poderia ser identificada como "GGG", sigla para *Giant Global Graph* (BERNERS-LEE, 2007).

Os agrupamentos de triplas formam repositórios de dados inter-relacionados. São como bases de dados, porém organizados em grafos, e não como tabelas do modelo relacional, amplamente utilizado na web atualmente. Esses repositórios de dados em triplas são denominados de *triple store* (SHADBOLT et al, 2006). Assim como ocorre nas bases de dados relacionais, é possível realizar buscas dentro dos *triple stores*, porém, é permitido utilizar comandos de busca mais complexos que aqueles normalmente utilizados nos atuais motores de busca da web. Tomamos como exemplo o grafo da Figura 9: um serviço ideal de busca semântica possibilitaria a execução da seguinte pesquisa: "Quantas pessoas vivem na cidade em que Gavin Newsom governa?", e o resultado seria gerado de forma automatizada. Em um grafo mais complexo sobre a cidade de São Francisco, as perguntas poderiam ser ainda mais variadas: "Qual a temperatura em São Francisco hoje?", "Qual foi a temperatura média em São Francisco em 1970?", "Quem foi o prefeito de São Francisco em 1970?", e assim por diante. Todavia, tais pesquisas ainda poderiam gerar resultados ambíguos, afinal, há a probabilidade de existir mais do que uma cidade denominada São Francisco no mundo. Da mesma forma, repetem-se nomes de pessoas, de empresas ou de lugares. Para que a Web Semântica funcione como um sistema com capacidade de gerar inferências, é necessário que exista um modo de identificar sem ambiguidades as entidades presentes nas triplas.

Para materializar a proposta de modelo de dados em triplas sem ambiguidades, a W3C desenvolveu a especificação RDF (*Resource Description Framework*), que é uma linguagem para representar informações sobre recursos na World Wide Web (W3C, 2004a). Os "recursos na web" são quaisquer elementos passíveis de descrição. Para descrevê-los, o RDF emprega o (já mencionado) modelo das triplas: o recurso a ser descrito ocupa o lugar do sujeito, o predicado é uma propriedade do recurso, e o objeto é o valor atribuído ao predicado. Porém, para que não ocorra ambiguidade, os dados inter-relacionados na tripla do RDF são devidamente identificados com o uso de um identificador único, denominado URI (*Uniform Resource Identifier*), que nada mais é do que um endereço único que aponta para determinado

recurso. A web atual já utiliza a URI em seu funcionamento básico, pois, para se acessar uma página da web, é necessário inserir no navegador um endereço único, denominado URL (*Uniform Resource Locator*), que é um tipo de URI (BERNERS-LEE et al, 2002). Em outras palavras, a URI é um índice que pode ser representado de diversas formas (tais como palavras, códigos ou números), e uma dessas formas é a URL, que é um endereço único para um recurso da web (que geralmente inicia com a combinação "http://www.").

Tomemos novamente como exemplo a tripla que satisfaz a frase **o blog X é escrito por João**. O sujeito da tripla (blog X) é o recurso a ser descrito e pode ser identificado pelo endereço (URL) do próprio blog, afinal, uma URL é um URI, logo, é um identificador único. Da mesma forma, o objeto da tripla (João) poderia ser simplesmente um valor escrito (no caso, a palavra "João") ou poderia ser um outro recurso disponível na web e identificado com uma URI que apontasse para esse recurso. Desta maneira, o RDF permite criar relações entre recursos da web (ver Figura 10).

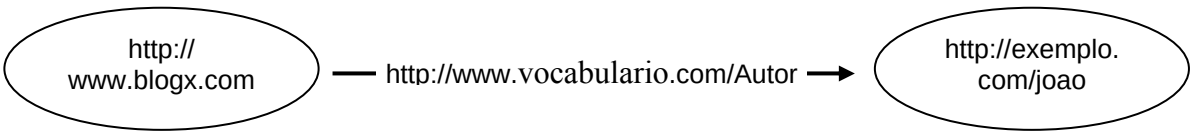


Figura 10 – Exemplo de tripla com sujeito, predicado e objeto identificados através do uso de URI

Se o blog X é identificado pelo URI do próprio blog, e se João é identificado por uma página que o representa na web, como identificar o predicado "é escrito por"? Como encontrar um recurso na web que representaria uma propriedade? Podemos citar exemplos de possíveis propriedades, como "nome", "localização", "ano de surgimento", entre muitos outros. As propriedades que definem as relações entre sujeito e objeto devem ser padronizadas e compartilhadas na web a fim de se manter um ambiente propício para o intercâmbio de dados e de seus significados. Assim, ao se definir um número limitado de possíveis propriedades para a tripla, haveria a certeza de que diferentes sites e aplicativos estariam a utilizar as mesmas lógicas no relacionamento entre sujeitos e objetos. Essa padronização de propriedades ocorre com a publicação das mesmas em vocabulários disponíveis na web e compartilhados entre os sites.

Já que esses vocabulários são publicamente disponíveis na web, então uma tripla pode indicar a sua propriedade através de uma URI que aponte para um vocabulário que possua tal propriedade. Por exemplo, se desejamos utilizar a propriedade "autor" e se essa propriedade

está presente no vocabulário disponível no endereço fictício “http://www.vocabulario.com”, então a tripla poderia utilizar a URL “http://www.vocabulario.com/Autor” (ver Figura 10).

O vocabulário fictício apresentado no exemplo da Figura 10 poderia fornecer outros tipos de propriedades, tais como “http://www.vocabulario.com/Endereco” ou “http://www.vocabulario.com/Data-publicacao”. Um exemplo real de vocabulário disponível atualmente na web é o Dublin Core[57], um projeto que publica na web uma lista de 15 categorias aplicáveis na organização de publicações, tais como título, autor, assunto, descrição, editora, data, formato, língua, entre outros. O projeto não foi originalmente desenvolvido para a Web Semântica. Ele surgiu ainda em 1995 na forma de uma lista de metadados aplicáveis na catalogação de recursos editoriais, como livros em bibliotecas. Entretanto, o projeto se transformou em uma lista padronizada de metadados aplicáveis em diferentes tipos de projetos, entre eles a própria WS. Então, se um site decide utilizar o padrão Dublin Core como vocabulário de propriedades, o exemplo da Figura 10 utilizaria como predicado o seguinte endereço: “http://purl.org/dc/elements/1.1/author” (endereço real do projeto Dublin Core para a propriedade “autor”, em setembro de 2011).

Nas bases de dados tradicionais em modelo relacional, é possível consultar os dados armazenados com o uso da linguagem SQL (*Structured Query Language*), utilizada para a execução de determinadas operações de consulta e escrita de dados, tais como seleção (SELECT), adição (INSERT), exclusão (DELETE) e alteração (UPDATE). Se as triplas em RDF formam bases de dados em estrutura de grafo, então deve haver uma forma de consultar estes dados. E, de fato, é possível realizar buscas dentro dos *triple stores* em RDF com o uso da linguagem SPARQL (*Simple Protocol and RDF Query Language*), que oferece uma gama de possíveis operações nos grafos, como a seleção, o filtro e a comparação de dados, entre outras (SEGARAN et al, 2009). Entretanto, ao contrário do SQL, o SPARQL apenas oferece opções de consulta aos dados (leitura), enquanto o SQL também permite opções que modificam os dados nas bases de dados relacionais (escrita). Embora pareça uma limitação, tal situação pode ser vantajosa para a Web Semântica, pois permite aos sites que disponibilizem publicamente na web suas bases de dados em grafos e os abram para consulta realizada por terceiros, sem o receio de que um agente externo modifique os dados ali armazenados (SEGARAN et al, 2009). Desta forma, os grafos da Web Semântica têm o potencial de formar uma grande base de dados em comum, pois o SPARQL permite a seleção e comparação de dados armazenados em grafos diferentes.

[57] Disponível em: <http://dublincore.org/>. Acesso em: 17 set 2011.

Um exemplo de associação de triplas em RDF é o site Data.gov, lançado pelo governo dos EUA. No site, são publicadas grandes coleções de dados sobre diversas áreas da administração pública daquele país, tais como saúde, educação e gastos militares. Grande parte destes dados foi convertida para um formato compatível com o padrão RDF e, por isso, há dados estruturados na lógica das triplas (sujeito, predicado e objeto). Segundo o próprio site, a soma dos arquivos em RDF já contava 6,4 bilhões de triplas disponíveis para download em setembro de 2011[58]. As coleções de dados em RDF podem ser acessadas por qualquer usuário ou site da web (ver Figura 11).

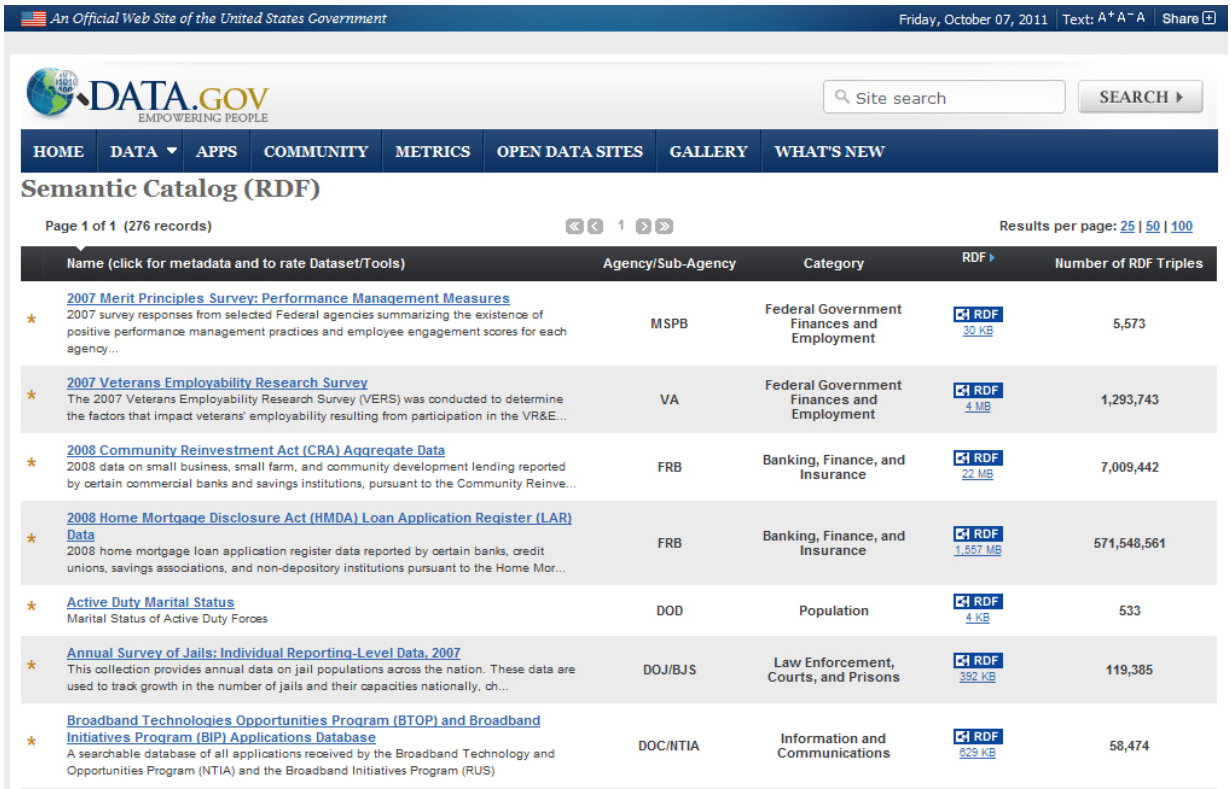


Figura 11 – Lista de coleções de dados em RDF disponíveis para download no site Data.gov[59]

As coleções de RDF deste site não são apresentadas em uma formatação "amigável" para a leitura do usuário comum da web, pois estão disponíveis em grandes blocos de código, não formatados, ainda "crus", que misturam conteúdo com marcações XML[60]

[58] Informação disponível em: <http://www.data.gov/semantic>. Acesso em: 18 set 2011.
[59] Disponível em: <http://www.data.gov/semantic/data/alpha>. Acesso em: 07 out 2011.
[60] O RDF está representado na linguagem XML, porque o RDF não é uma linguagem com sintaxe própria, ele é um modelo de dados, que pode ser representado em diferentes formatos. Essa representação do modelo em um formato escrito é chamada de "serialização". Entre vários tipos de serializações, podemos citar: a N-Triples, a N3, a RDF/XML (apresentada na Figura 12) e a RDFa, utilizada dentro do código HTML (SEGARAN et al, 2009; W3C, 2004a). Logo, é possível escrever dados no modelo RDF com a linguagem XML.

(como é possível observar na Figura 12). São dados disponíveis para serem lapidados por outros aplicativos.

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#" xmlns:rdfs=
"http://www.w3.org/2000/01/rdf-schema#" xmlns:socrata="http://www.socrata.com/rdf/terms#"
xmlns:dcterm="http://purl.org/dc/terms/" xmlns:geo="http://www.w3.org/2003/01/geo/wgs84_pos#"
xmlns:skos="http://www.w3.org/2004/02/skos/core#" xmlns:dcat="http://www.w3.org/ns/dcat#"
xmlns:foaf="http://xmlns.com/foaf/0.1/" xmlns:dsbase="http://data.medicare.gov/views/" xmlns:ds=
"http://data.medicare.gov/views/_4cpu-68he/">
<dsbase:_4cpu-68he rdf:about="http://data.medicare.gov/views/_4cpu-68he/rows/1">
    <socrata:rowID>1</socrata:rowID>
    <rdfs:member rdf:resource="http://data.medicare.gov/views/_4cpu-68he"/>
    <ds:provider_number>015010</ds:provider_number>
    <ds:nursing_home_name>COOSA VALLEY NURSING FACILITY</ds:nursing_home_name>
    <ds:street>315 WEST HICKORY STREET</ds:street>
    <ds:city>SYLACAUGA</ds:city>
    <ds:state>AL</ds:state>
    <ds:zip_code>35150</ds:zip_code>
    <foaf:phone rdf:resource="tel:2562495604"/>
    <ds:enforcement_type>CMP </ds:enforcement_type>
    <ds:civil_money_penalty>12250</ds:civil_money_penalty>
    <ds:survey_date>20090901</ds:survey_date></dsbase:_4cpu-68he>
<dsbase:_4cpu-68he rdf:about="http://data.medicare.gov/views/_4cpu-68he/rows/2">
    <socrata:rowID>2</socrata:rowID>
    <rdfs:member rdf:resource="http://data.medicare.gov/views/_4cpu-68he"/>
    <ds:provider_number>015015</ds:provider_number>
    <ds:nursing_home_name>PLANTATION MANOR NURSING HOME</ds:nursing_home_name>
    <ds:street>6450 OLD TUSCALOOSA HIGHWAY   P O BOX 97</ds:street>
    <ds:city>MC CALLA</ds:city>
    <ds:state>AL</ds:state>
    <ds:zip_code>35111</ds:zip_code>
    <foaf:phone rdf:resource="tel:2054776161"/>
    <ds:enforcement_type>CMP </ds:enforcement_type>
```

Figura 12 – Visualização parcial de uma das coleções de dados em RDF/XML disponíveis para download no site Data.gov

O próprio site Data.gov incentiva programadores a desenvolverem aplicativos que utilizem as coleções de dados estruturados em triplas e os convertam em informações adaptadas para a leitura dos usuários na web. O site disponibiliza uma página em que são apresentados aplicativos (sites da web) desenvolvidos por terceiros e que se utilizam de tais dados para oferecer informações específicas. Muitos desses aplicativos são *mashups*[61] que misturam os dados do Data.gov com outras fontes de dados disponíveis na web, ou que os aplicam em mapas interativos, como o Google Maps.

Em 2009, o Data.gov lançou uma competição para incentivar o desenvolvimento de aplicativos que utilizassem o conteúdo do site[62] de maneira criativa, e um dos três vencedores

[61] Na computação, *mashups* são aplicativos que combinam dados ou serviços oriundos de diversas fontes para criar um novo produto ou serviço.
[62] Competição "Apps for America 2". Disponível em: <http://sunlightfoundation.com/blog/2009/09/10/apps-for-america-2-winners/>. Acesso em: 18 set 2011.

foi o site *This We Know*[63], que utiliza os dados estruturados do Data.gov para apresentar estatísticas sobre diferentes áreas da administração pública dos EUA, tudo de forma automatizada. Na página inicial do site (ver Figura 13), são apresentados alguns rankings de cidades norte-americanas em relação a determinados temas, como, por exemplo, a lista das cinco cidades com maior quantidade de toxinas no meio ambiente ou as cinco cidades com menores índices de desemprego. Na Figura 13, é mostrada a página inicial do site, em que destacamos com um círculo a lista das cinco cidades com maior incidência de câncer (e que indica Los Angeles como a cidade que apresenta o maior índice).

Figura 13 – Página inicial do site This We Know, em que são apresentadas listas com *rankings* entre cidades norte-americanas[64]

[63] Disponível em: <http://thisweknow.org/>. Acesso em: 13 dez. 2011.
[64] Disponível em: <http://thisweknow.org/>. Acesso em: 13 dez. 2011.

Além dos *rankings* de cidades, organizados por assuntos ou temas, o site também constrói e apresenta, de forma automatizada, páginas com dados estruturados para cada uma das cidades, mostrando estatísticas de interesse público. Na Figura 14, é possível observar o resultado para a busca na cidade de Bridgeport (estado de Connecticut). Nestes resultados, são apresentados dados como a quantidade de fábricas ("Há 15 fábricas", na 1ª linha, marcada com a letra A), de crimes violentos ("1603 crimes violentos ocorreram ou 11,6 por pessoa", na 4ª linha, marcada com a letra B), de desempregados ("36369 pessoas desempregadas, enquanto 443028 possuem empregos", na 6ª linha, marcada com a letra C), entre outras informações de interesse público:

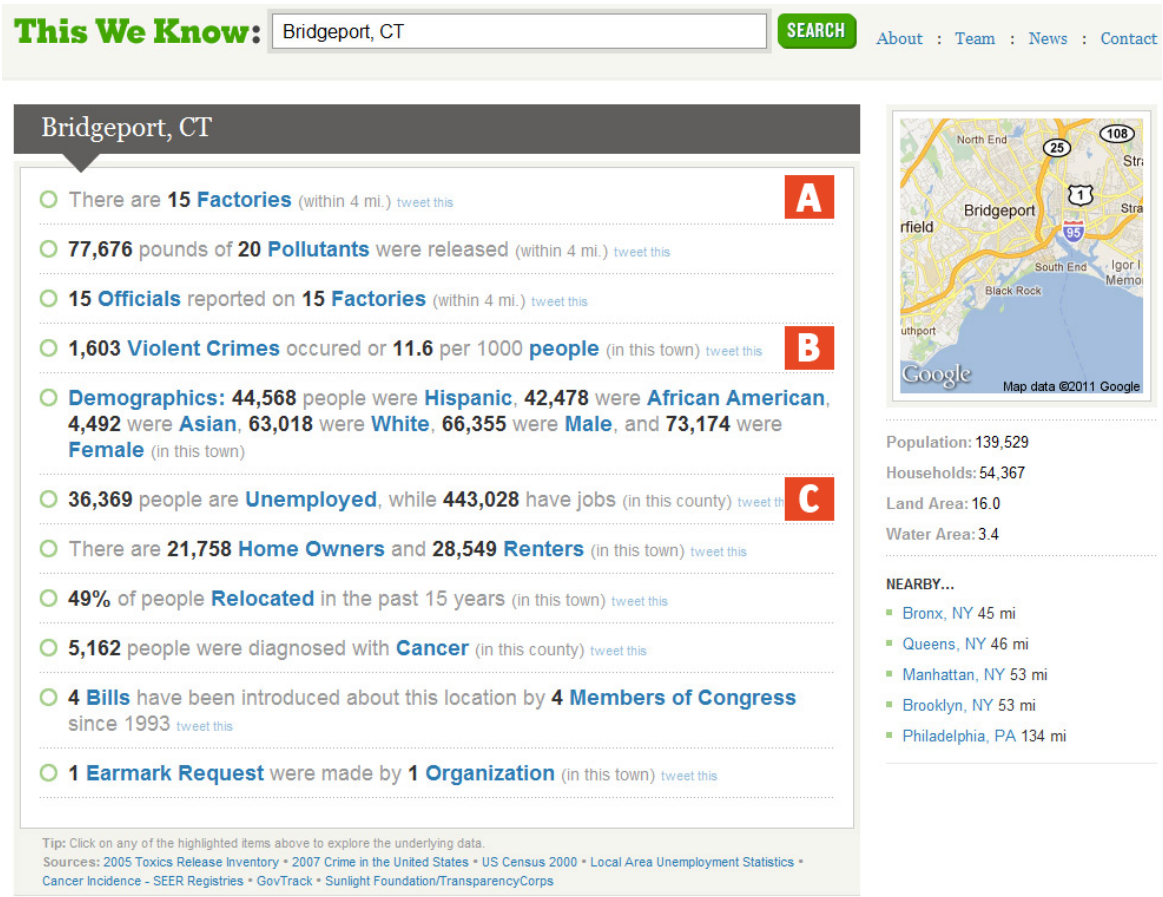


Figura 14 – Página do site This We Know, que apresenta números sobre uma cidade dos EUA, como o número de fábricas (A), de crimes violentos (B) e de empregados x desempregados (C)[65]

Os bancos de dados relacionais, largamente utilizados pelos atuais sites dinâmicos, também oferecem a funcionalidade de armazenamento e cruzamento de dados. Porém, os desenvolvedores do site *This We Know* justificam o uso das tecnologias padronizadas da Web Semântica neste projeto:

[65] Disponível em: <http://thisweknow.org/>. Acesso em: 13 dez. 2011.

> Uma vantagem em armazenar as informações do data.gov usando RDF é que a base de dados e os aplicativos podem prontamente se expandir na medida em que novas fontes de dados são adicionadas ao catálogo, sem requerer nova digitação de código ou revisões do código existente. Em uma base de dados relacional, as conexões entre as informações teriam de ser feitas com antecedência, revisões seriam necessárias assim que novas bases de dados fossem carregadas, e o modelo final de dados se tornaria extremamente largo e pesado se milhares de bases de dados tivessem que ser modelados como uma única base de dados [66] (THIS WE KNOW, *online*).

Apresentamos, até aqui, alguns conceitos-chaves para o entendimento do que é a proposta da WS e das condições que a tornam viável: a necessidade dos metadados para as máquinas, o modelo de metadados em triplas (RDF) e a identificação de recursos com o uso de identificadores únicos (URI). Há projetos da Web Semântica que utilizam basicamente estas tecnologias, e que já apresentam resultados ricos, como no caso do site *This We Know*. Entretanto, além destes conceitos, o ideal de Web Semântica proposto por Tim Berners-Lee ainda propõe um recurso mais complexo que, além de relacionar dados a significados, permite às máquinas identificarem regras de relacionamento entre esses dados publicados na web. A identificação dos tipos de relacionamentos permite às máquinas realizarem inferências sobre tais dados. Na concepção de Berners-Lee et al (2002), as regras de relacionamento entre entidades devem ser formalizadas através de um recurso denominado ontologia.

2.2.3 Ontologias

Antes da ideia de WS, já era possível realizar o inter-relacionamento de dados a metadados através de outras tecnologias, como a dos bancos de dados relacionais. A proposta da Web Semântica apresenta um recurso ainda mais complexo utilizado para explicitar os relacionamentos desses dados a determinados significados, o que possibilita, teoricamente, que esse sistema gere inferências sobre determinadas situações. Ao considerarmos a frase exemplo "todos os humanos são mamíferos", conclui-se que para seguir a lógica da Web Semântica, é necessário informar à máquina o significado do termo **humanos** e do termo **mamíferos**. Embora os metadados possam indicar à máquina o significado dos dois termos, como seria possível indicar os tipos de relacionamentos possíveis entre eles? Se todos os humanos são mamíferos, então seria correto fazer a relação inversa e afirmar que todos os mamíferos são humanos?

---

[66] *An advantage of storing the data.gov information using RDF is that the database and applications can readily expand as new data sources are added to the catalog, without requiring new coding or revisions to existing coding. In a relational database, the connections between information will need to be made in advance, revisions will be necessary as new databases are loaded, and the data model will become extremely large and unwieldy if thousands of databases were to be modeled in a single database.*

Para os humanos, a identificação de significados e de suas inter-relações parte das suas experiências com a realidade; ou seja, a partir do conhecimento adquirido. Pode-se citar como exemplo o fato de um ser humano compreender que na relação pai-filho o pai sempre será o mais velho, pois, segundo o seu conhecimento adquirido, na relação entre pais e filhos, a regra será sempre que o primeiro é o mais velho. Então, o sistema da Web Semântica também deveria ter como base para essas inferências algum tipo de relação com a realidade e com o conhecimento. E, na concepção da WS proposta por Tim Berners-Lee, esse processo de fato ocorre através da associação das informações a vocabulários padronizados e compartilhados na web. Esses vocabulários são arquivos, interpretáveis pelas máquinas, que descrevem os termos empregados em um domínio específico do conhecimento. Quando os vocabulários compartilhados apresentam regras formais de relacionamentos entre tais termos (através de classes, subclasses, funções etc), são chamados de "ontologias". Por exemplo, uma ontologia especificamente para o domínio farmacêutico poderia descrever formalmente as regras de relacionamento que existem nas interações entre os fármacos e suas substâncias ativas. Por essa razão, as ontologias podem ser consideradas representações abstratas do conhecimento, geralmente desenvolvidas para determinados domínios do conhecimento humano.

O termo "ontologia" vem da filosofia grega e, segundo Berners-Lee et al, significa a "teoria sobre a natureza da existência, sobre 'que tipos de coisas' existem" (2002, p. 27). Ainda segundo o autor, os pesquisadores da inteligência artificial e da web adaptaram o termo da filosofia e o tomaram como um jargão para fazerem referência ao "documento ou arquivo que define formalmente as relações entre os termos. O tipo de ontologia mais representativo para a web possui uma taxonomia e uma coleção de regras de inferência" [67] (2002, p. 27). As taxonomias definem as classes dos objetos e as relações hierárquicas entre essas classes, para, assim, permitirem a geração de inferências lógicas e consistentes.

Ontologias não surgiram com a Web Semântica. Antes do surgimento da web, elas já eram estudadas e aplicadas na área de inteligência artificial. Ainda em 1992, Thomas Gruber, pesquisador da área, apresentou um conceito de ontologia bastante citado por autores que pesquisam a Web Semântica (KASHYAP et al, 2008; AKERKAR, 2009; KASHYAP et al, 2008, BREITMAN, 2005): para ele, trata-se de "uma especificação explícita de uma conceituação" (GRUBER, 1993, p. 2). Para facilitar o entendimento do conceito, é necessário compreender o que é "conceituação", que para Gruber é "uma visão abstrata e simplificada do

[67] *Artificial-intelligence and Web researchers have co-opted the term for their own jargon, and for them an ontology is a document or file that formally defines the relations among terms. The most typical kind of ontology for the Web has a taxonomy and a set of inference rules.*

mundo que queremos representar por alguma razão" (1993, p. 2). Essa "visão do mundo" é formada por objetos, conceitos e outras entidades que presumidamente existem em alguma área de interesse, além dos relacionamentos que existem entre eles. Essa coleção de objetos e de seus relacionamentos é formalizada em um vocabulário utilizado para representar o conhecimento humano; porém, especificamente na área em que esses objetos fazem parte. Como ilustração, podemos citar uma ontologia para uma sala de aula: ela especifica os principais tipos de elementos existentes no domínio (alunos, professores, carteiras, cadeiras etc), especifica as propriedades desses elementos a partir de classe e subclasses (como em uma taxonomia que classifica os seres vivos) e determina as regras de relacionamento entre esses elementos, permitindo a geração de inferências (ex.: uma sala pode conter alunos, mas não ocorre o inverso). Para Gruber (1993), é muito dispendioso de se construir, testar e manter os sistemas e serviços baseados em conhecimento humano, e é por isso que as ontologias são necessárias, porque são representações complexas que, depois de produzidas, podem ser compartilhadas e reutilizadas pelos *software* aplicativos e sistemas inteligentes.

Entende-se, então, que as ontologias funcionam como vocabulários precisos, que expressam regras formais de relacionamentos para inferências (SEGARAN et al, 2009), que podem ser utilizados para diversas aplicações em que há vantagens em associar entidades a significados. Nas palavras de Berners-Lee:

> Ontologias podem enriquecer o funcionamento da web de várias maneiras. Elas podem ser utilizadas como uma simples forma de aprimoramento na precisão de ferramentas de busca da web – o programa de busca pode procurar somente por páginas que se referem precisamente a um conceito específico ao invés de todas as outras páginas que estejam utilizando palavras-chaves ambíguas. Aplicativos mais avançados irão utilizar ontologias para relacionar a informação de uma página às estruturas de conhecimentos associadas e às regras de inferência [68] (2002, pg. 28).

Nas triplas em RDF, fazemos referências a termos (como "São Paulo" e "Brasil") e às relações entre tais termos (como "faz parte de", "pertence a", "é autor de", "é igual a"). Já a ontologia faz uma classificação desses termos e de seus relacionamentos como se fossem regras para um processo de inferência. A Web Semântica não propõe uma ontologia única e geral para todo o sistema, mas diferentes ontologias para diferentes domínios, e os termos descritos pela ontologia devem ser de comum aceitação dentro da comunidade que faz parte do domínio (AKERKAR, 2009).

---

[68] *Ontologies can enhance the functioning of the Web in many ways. They can be used in a simple fashion to improve the accuracy of Web searches – the search program can look for only those pages that refer to a precise concept instead of all the ones using ambiguous keywords. More advanced applications will use ontologies to relate the information on a page to the associated knowledge structures and inference rules.*

A ideia de uma lista de itens categorizados com suas definições pode trazer a imediata lembrança de uma taxonomia[69] ou de um tesauro[70]; porém, embora semelhantes, as ontologias são propostas mais completas, pois definem regras complexas de relacionamento entre os itens categorizados, tais como ambiguidades, semelhanças etc. Ainda assim, embora as taxonomias e os tesauros não apareçam nas listas de "principais tecnologias da Web Semântica", eles ainda pertencem ao cenário da WS (AKERKAR, 2009). Tanto que, para Berners-Lee (et al), "o tipo de ontologia mais representativo para a web possui uma taxonomia e uma coleção de regras de inferência" [71] (2002, p. 27).

2.2.4 Uma linguagem para construção de ontologias: o padrão OWL

De acordo com Breitman (2005), existem diferentes linguagens que possibilitam o desenvolvimento de ontologias aplicadas à Web Semântica, tais como a *Ontology Inference Layer* (OIL), desenvolvida por um consórcio da Comunidade Europeia; a *DARPA Agent Markup Language* (DAML), desenvolvida pela agência norte-americana DARPA (*Defense Advanced Research Projects Agency*); ou ainda o próprio RDF[72], pois como ele é um modelo de dados (modelo em triplas) que pode ser utilizado para modelar regras, então pode relacionar termos a predicados e conceitos (ex.: X / pode fazer parte de / Y). Em certo momento, os desenvolvedores europeus da OIL e norte-americanos da DAML uniram esforços para formular uma linguagem em comum para ontologias (DAML+OIL), integrando nesta mesma linguagem as funcionalidades de cada uma, tais como elementos de classe, expressão de classes e propriedades (BREITMAN, 2005). Em busca de uma linguagem para ontologias aplicadas à web, a W3C realizou uma revisão da linguagem DAML+OIL e desenvolveu a OWL (*Web Ontology Language*), uma linguagem de marcação semântica utilizada para recursos da web que possui classes, subclasses, propriedades, subpropriedades e restrições de propriedades (Akerkar, 2009).

---

[69] "Taxonomia é um vocabulário controlado hierarquicamente organizado. O mundo tem muitas taxonomias, porque o ser humano naturalmente classifica as coisas. Taxonomias são semanticamente fracas e são normalmente usadas quando se navega sem se preocupar em se ter uma precisão na pesquisa" (AKERKAR, 2009, p. 76, tradução nossa).

[70] "Tesauro é um vocabulário controlado e arranjado em uma ordem e uma estrutura já conhecidas, que as equivalências e as relações homográficas, hierárquicas e associativas entre os termos são apresentadas claramente e identificadas por indicadores de relacionamento padronizados" (AKERKAR, 2009, p. 76, tradução nossa).

[71] *The most typical kind of ontology for the Web has a taxonomy and a set of inference rules*.

[72] A W3C desenvolveu um modelo de dados que facilita a descrição de vocabulários com o RDF. A esse modelo, é denominado RDF Schema. É uma extensão ao RDF, pois além do modelo sujeito – propriedade – objeto, também inclui a funcionalidade de descrição mais detalhada sobre a propriedade (W3C, 2004b), recurso esse necessário para se criar vocabulários mais complexos.

Nas definições da W3C (2004c), a OWL é um modelo baseado em RDF e RDFS, e possui quatro elementos básicos: classes, propriedades, instâncias de classes e relacionamentos. Abaixo, apresentamos uma breve descrição para cada elemento:

- **Classes**: são grupos que abrigam unidades individuais que compartilham das mesmas características.
- **Instâncias de classes**: são as unidades individuais que fazem parte das classes.
- **Propriedades**: são atributos aplicados a toda a classe ou apenas às instâncias de classes.
- **Relacionamentos**: são as regras formais que se aplicam no relacionamento entre as instâncias.

Para ilustração, podemos citar o seguinte exemplo: se definirmos que "mamíferos" é uma classe, podemos considerar que "leão" é uma instância desta classe. Podemos considerar que a classe "mamíferos" tem como propriedade em comum a presença da mama. Logo, todas as instâncias que pertencem à classe devem herdar tal propriedade. Poderíamos, também, definir como propriedade "tem juba", porém aplicaríamos apenas à instância "leão", pois nem sempre os animais mamíferos têm esta característica.

Segundo Kashyap et al., as ontologias em OWL conseguem representar restrições e axiomas e, a partir deles, as máquinas teriam a capacidade de inferir "relacionamentos equivalentes entre dois conceitos além de mutuais contradições entre conceitos, se eles existirem" (2009, p. 32). Como exemplo de regras semânticas da OWL, citamos as seguintes lógicas:

> Filiação a uma classe. Se x é uma instância da classe C, e C é uma subclasse de D, então nós podemos inferir que x é uma instância de D.
> Equivalência de classes. Se a classe A é equivalente à classe B, e a classe B é equivalente à classe C, então A é equivalente a C, também.
> Consistência. Suponha que nós declaremos x como uma instância da classe A e que A é a subclasse de B ∩ C, A é uma subclasse de D, e B e D são disjuntos. Então nós temos uma inconsistência porque A deveria ser vazio, mas tem a mesma instância de X. Essa é uma indicação de um erro na ontologia.
> Classificação. Se nós temos declarado que certos pares de propriedade-valor são uma condição suficiente para a filiação em uma classe A, então se um indivíduo x satisfaz tal condição, nós podemos concluir que aquele x precisa ser uma instância de A (ANTONIOU et al., 2004, p. 110, tradução nossa).

O desenvolvimento de ontologias pode ser, muitas vezes, um trabalho árduo, pois além de exigir do desenvolvedor o conhecimento técnico da linguagem OWL, ainda há a tarefa de modelar a representação de uma área do conhecimento, que geralmente é formada por uma grande variedade de objetos e de seus relacionamentos. Para facilitar o trabalho na

modelagem de ontologias, um grupo de pesquisadores da Universidade de Stanford disponibiliza na web um editor gratuito de ontologias chamado *Protégé*[73], que permite a criação e também a visualização da ontologia em diferentes representações visuais, como listas ou mapas mentais. Na Figura 15, é mostrada uma tela do software *Protégé* com uma ontologia em OWL que pesquisadores da área da saúde desenvolveram para auxiliar médicos e seus pacientes portadores de diabetes a controlarem o consumo de alimentos, baseados em uma dieta apropriada para diabéticos. Para isso, foram descritos diversos alimentos e suas propriedades, como tipos e quantidades de nutrientes. A ontologia determinou certas regras de relacionamento entre estes nutrientes e as recomendações médicas (CANTAIS et al, 2005).

Na Figura 15, o software *Protégé* apresenta parte da ontologia: na caixa demarcada com a letra A, são listadas as classes, subclasses e suas relações (neste caso, alguns tipos de alimentos, tais como "frutas", "carne" e "vegetais"); na caixa demarcada com a letra B, são listadas algumas possíveis propriedades da classe, como "tem álcool" (*hasAlcohol*) e "tem gordura animal" (*hasAnimalFat*); e na caixa C, as regras para a classe (tais como restrições, condições etc), que a máquina utiliza para executar inferências.

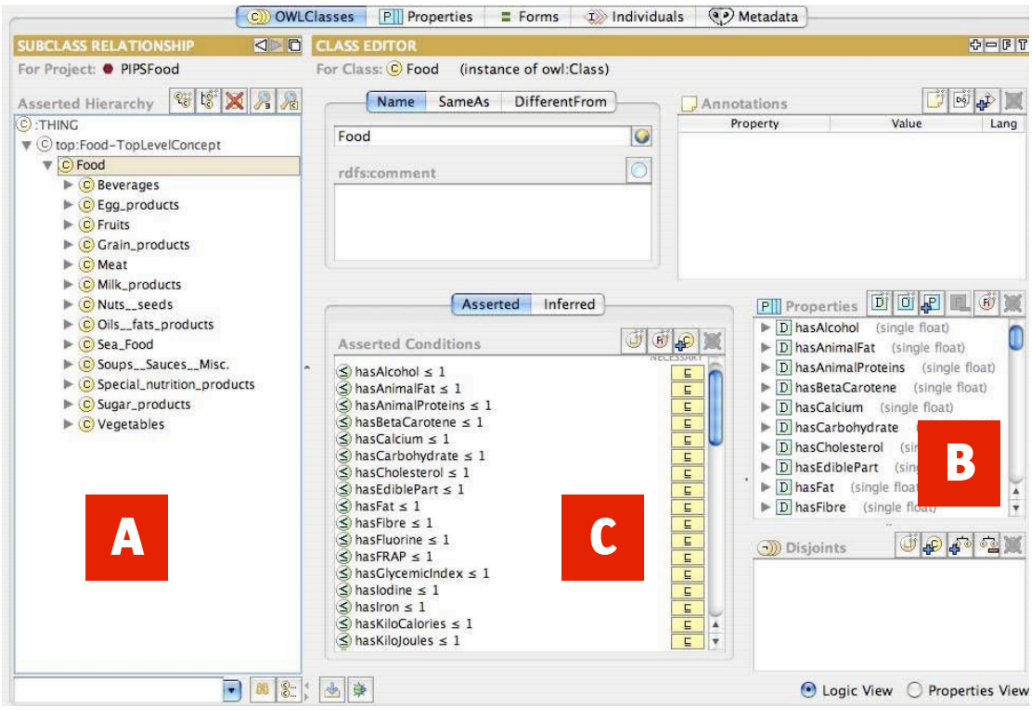


Figura 15 – Tela do software Protégé que mostra parte de uma ontologia em OWL (CANTAIS et al., 2005)

Atualmente, os projetos inseridos no âmbito da Web Semântica que utilizam recursos como triplas em RDF não necessariamente utilizam as ontologias devido à complexidade do

[73] http://protege.stanford.edu/

desenvolvimento deste recurso. Ainda assim é possível realizar inferências em um grafo sem o uso de ontologias, pois as regras de inferências podem estar implícitas nos comandos de pesquisa (*queries*) realizadas na recuperação dos dados do grafo (SEGARAN, 2009). Por exemplo, em um grafo que relaciona filmes e atores através de triplas, seria possível realizar a seguinte pesquisa: "listar os filmes em que o ator Jack Nicholson atuou no período entre 1980 e 1990". O resultado será uma resposta lógica, devido às regras impostas no comando de pesquisa. Entretanto, ao se compartilhar os dados deste grafo com outros aplicativos, tais aplicativos não teriam como identificar as regras de relacionamento entre as entidades presentes no grafo. Por isso, as ontologias são fundamentais para a visão de uma Web Semântica plena, visto que não haveria como diferentes aplicativos processarem inferências com uma mesma lógica se não houvesse um vocabulário de termos e de regras em comum entre eles. Por essa razão que as ontologias devem ser formalizadas, explícitas e compartilhadas, pois dessa maneira poderão ser utilizadas por diferentes máquinas (sites, serviços, agentes, entre outros) de modo que seja minimizada a ocorrência de ambiguidades. Para Akerkar (2009, p. 74), "metadados e ontologias são complementares e constituem os blocos de construção da Web Semântica. Eles evitam ambiguidades nos significados e proveem respostas mais precisas" [74]. No entanto, como o desenvolvimento de uma ontologia é um trabalho complexo, grande parte dos produtos experimentais da Web Semântica ainda não utiliza o recurso para seu funcionamento e, como consequência, oferece às máquinas menor capacidade de geração de inferências.

2.2.5 As máquinas tomam a iniciativa: os agentes inteligentes

Por fim, para que a Web Semântica seja possível na visão de Berners-Lee, além da estruturação de dados, das ontologias e dos metadados, outro recurso importante para esse sistema é o conceito de agentes. Berners-Lee (et al, 2002) afirma que a Web Semântica só será possível quando as "pessoas" (desenvolvedores da web) criarem programas que, de forma autônoma, coletem conteúdos de diversas fontes da web, processem tais informações e então troquem os resultados com outros programas (ou seja, outro agentes).

Os agentes não seriam exatamente essas pessoas, mas esses programas criados por elas, automatizados e autônomos, como no caso já citado neste trabalho de uma agenda online

[74] *Metadata and ontologies are complementary and constitute the Semantic Web's building blocks. They avoid meaning ambiguities and provide more precise answers.*

que se comunica com um site de compras de passagens aéreas. Para Berners-Lee, os agentes seriam os impulsionadores da Web Semântica. Eles também seriam responsáveis por averiguar a confiabilidade da fonte dos conteúdos, pois informações erradas trariam prejuízos às associações de dados.

2.2.6 Extração de conceitos em conteúdos não estruturados

Com a combinação das tecnologias semânticas até aqui apresentadas, como o RDF e as ontologias, é possível desenvolver aplicações que trabalham de modo automatizado com os dados publicados na web. Porém, para que isso seja possível, é necessário indicar às máquinas quais são os significados destes dados, através de metadados que os descrevam.

Em um conteúdo estruturado, como em um texto fragmentado e ordenado em uma planilha, a associação de metadados às partes do texto é facilitada. Por exemplo: nessa situação, é possível indicar à máquina que uma determinada coluna da planilha deve ser associada a alguns metadados específicos, que, por sua vez, podem indicar às máquinas alguns significados para os conteúdos que fazem parte daquela coluna.

No jornalismo, a situação é bastante diferente. Geralmente, a produção jornalística resulta em narrativas não estruturadas, escritas exclusivamente em linguagem natural, ou seja, não preparadas para a compreensão por parte das máquinas. Sem essa capacidade de interpretação, os computadores não têm como identificar os conceitos presentes nas narrativas. De nada valeria um sistema semântico capaz de gerenciar automaticamente a organização de conteúdos a partir de seus conceitos se não há uma maneira de identificar quais conceitos estão presentes no conteúdo em questão. Por isso, é preciso associar a estas narrativas os metadados que descrevam os conceitos ali presentes, pois são com estes metadados que as máquinas identificam significados. Atualmente, existem técnicas para a extração dos conceitos presentes em conteúdos não estruturados. A seguir, apresentamos duas delas: a técnica de *tagging* e a de extração automática de conceitos via *software*.

2.2.6.1 Técnica de *tagging*

Segundo Bertocchi (2009), uma maneira de se atribuir metadados a uma produção jornalística é pela técnica de *tagging*, que, do inglês, podemos traduzir como “etiquetagem”. A técnica nada mais é do que a associação de palavras-chaves (*tags*) a um conteúdo, para sugerir significados ou conceitos relacionados. Estas palavras-chaves podem ser atribuídas

pelo próprio autor da informação (como o jornalista) ou pelos usuários do site em que tal informação esteja publicada. A autora apresenta três tipos de "*tagging*":

1) ***folksonomia***, quando os usuários podem criar *tags* livremente, sem a necessidade de uma regra rígida, oferecendo maior liberdade, porém com o risco de se gerar indefinições linguísticas, como polissemias, diversidade de sinonímias e homonímias etc;
2) **taxonomia**, quando a classificação é realizada com *tags* hierárquicas e já previamente existentes em um vocabulário definido por uma equipe; e
3) ***folksonomia* controlada**, uma forma híbrida entre as duas anteriores, pois funciona a partir de uma taxonomia pré-definida, mas também permite contribuições de usuários.

Segundo Bertocchi, este terceiro modo de *tagging* seria o mais indicado para o jornalismo digital, já que possibilita a utilização e a integração de um repertório definido por um corpo editorial (jornalistas) e, também, de um repertório sugerido pelos leitores, e assim:

> [...] as livres associações de termos criadas pelos usuários refletirão a linguagem comum da coletividade ao mesmo tempo em que o vocabulário controlado da redação jornalística evidenciará suas predileções editoriais, conforme estratégia comunicativa previamente identificada (BERTOCCHI, 2009, p. 17).

Por ser um processo manual, a técnica de *tagging* exige certa dedicação de tempo do jornalista no processo de anotação do conteúdo. A seguir, descrevemos uma técnica automatizada de extração de conceitos.

2.2.6.2 *Software* de análise automática

Existem *software* aplicativos especializados em analisar conteúdos não estruturados para extrair conceitos do mesmo de forma automatizada. Para ilustrar esta técnica, podemos citar o site Calais[75], serviço lançado pela Thomson Reuters (mesma organização detentora da agência de notícias Reuters) que oferece ao usuário o serviço de análise e extração automática de significados presentes em textos escritos. Ao se submeter um texto para o site, ele realiza uma leitura automática e então o serviço identifica determinadas palavras-chaves incluídas no conteúdo e as compara com uma ontologia, conseguindo, assim, retornar ao usuário a

[75] Disponível em: <http://www.opencalais.com/>. Acesso em: 29 set. 2011.

identificação de diversas entidades presentes no site, como pessoas, lugares, organizações, eventos, livros etc., além de links para locais do ciberespaço que contenham descrições sobre tais entidades. O serviço não apenas identifica termos e conceitos, como também retorna metadados para cada entidade identificada, que podem ser utilizados na associação com outros dados da web. Segundo descrição do próprio site,

> O metadado oferece a você a possibilidade de construir mapas (ou gráficos ou redes) conectando documentos a pessoas a companhias a lugares a produtos a eventos a geografias a... qualquer coisa. Você pode usar estes mapas para aprimorar a navegação do seu site, prover distribuições contextualizadas, etiquetar e organizar seu conteúdo, criar folksonomias estruturadas, filtrar e reduplicar feeds de notícias, ou analisar um conteúdo para observar se ele contém o que você procura [76] (OPEN CALAIS, *online*)[77].

Ainda na descrição do produto, o site apresenta um gráfico que simplifica como o processo de extração ocorre. Apresentamos a imagem na Figura 16, traduzida por nós.

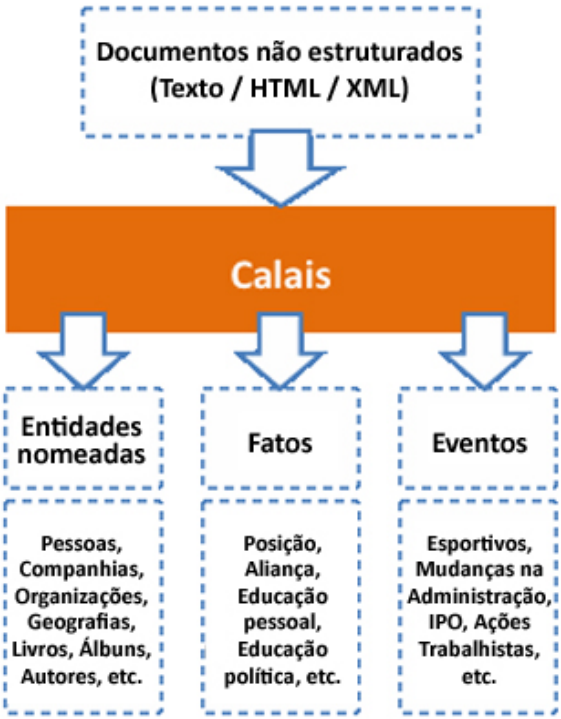


Figura 16 – Processo de extração de conceitos no serviço Calais[78]

[76] *The metadata gives you the ability to build maps (or graphs or networks) linking documents to people to companies to places to products to events to geographies to… whatever. You can use those maps to improve site navigation, provide contextual syndication, tag and organize your content, create structured folksonomies, filter and de-duplicate news feeds, or analyze content to see if it contains what you care about.*
[77] Disponível em: <http://www.opencalais.com/about>. Acesso em: 29 set. 2011. Tradução nossa.
[78] Disponível em: <http://www.opencalais.com/about>. Acesso em: 29 set. 2011. Tradução nossa.

Para ilustração, citamos um caso hipotético: uma ferramenta semântica que utiliza o serviço do site Calais poderia, por exemplo, realizar uma análise automática de uma notícia e apresentar no resultado da análise um resumo sobre as principais informações do lide, como a) “quem está envolvido no fato”, b) “onde ocorreu o fato”, c) “quando ocorreu o fato” etc; e ainda relacionar tais resultados com outras informações presentes na web, como a) outras notícias envolvendo os atores deste fato, b) informações extras sobre o local onde ocorreu o fato, c) lista de notícias que ocorreram no mesmo período deste fato etc.

Até o momento, apresentamos os principais conceitos sobre Web Semântica, de acordo com a visão de Berners-Lee et al (2002): metadados, triplas no modelo RDF, definição de conceitos e relações com ontologias, agentes inteligentes que trocam dados entre si. A seguir, apresentamos um movimento, também liderado por Tim Berners-Lee, que tem como objetivo criar uma rede de sites e serviços na web que utilizam de forma padronizada as tecnologias semânticas recomendadas pela W3C, e que, mais do que isso, têm como mote a prática da abertura e compartilhamento de seus dados.

2.3 Linked Data

Para que a WS cresça e se consolide, é necessário que surjam na web repositórios de grafos interligados, pois assim se cria um ambiente propício para a interoperabilidade de dados e de seus significados. Ou seja: sem dados estruturados de forma padronizada, não há uma rede semântica de dados. Porém, uma barreira para esse crescimento são os repositórios não padronizados e os repositórios fechados, que não permitem o acesso de sites e serviços externos a seus dados. Além de existirem maneiras diferentes de se publicar dados estruturados, também ocorrem práticas não recomendadas (ou mal executadas) na construção destes repositórios, que podem prejudicar a manutenção dos padrões.

Preocupado com a sustentabilidade do projeto da WS, Berners-Lee (2006) propôs uma série de processos padrões na publicação de dados estruturados em triplas. A essa prática, ele denominou Linked Data. Segundo Bizer et al. (2009), essas práticas padronizadas se referem basicamente a dados que: 1) sejam publicados na web de tal forma que possam ser lidos pelas máquinas, 2) seus significados sejam explicitamente definidos, 3) sejam lincados a outros repositórios externos de dados, e 4) permitam aos repositórios externos que se conectem a eles. Para que isso seja possível, os sites devem seguir quatro princípios básicos:

1. Use URIs como nome para as coisas.
2. Use HTTP URIs e então as pessoas poderão procurar por aqueles nomes.

> 3. Quando alguém procurar por uma URI, ofereça informações úteis, utilizando os padrões (RDF, SPARQL).
> 4. Inclua links para outras URIs, para que então os usuários possam descobrir mais coisas [79] (BERNERS-LEE, 2006, *online*).

Em outras palavras, o Linked Data é uma recomendação de boas práticas, em que os projetos envolvidos publicam seus dados dentro dos padrões da W3C e buscam vincular seus dados a repositórios externos que também seguem estas mesas práticas padronizadas. Assim, cria-se uma grande rede de grafos interligados, em que qualquer um dos projetos envolvidos pode utilizar livremente[80] os dados dos outros repositórios, formando uma espécie de banco de dados mantido por diversas fontes (BIZER et al., 2009). Em uma visão otimista de crescimento do Linked Data, esse sistema tende a se tornar o já citado *Giant Global Graph* (GGG), a "versão semântica" da rede *World Wide Web* (WWW).

No decorrer dos anos, surgiram diversos projetos com a preocupação de publicar seus dados e metadados em conformidade com esses padrões e, ainda, visando à abertura destes dados para outros sites. Por isso, o termo também é conhecido como *Linked Open Data*, ou seja, dados abertos e lincados. Desta maneira, a web se auto-organiza para o desenvolvimento de um ambiente propício ao compartilhamento (e reuso) de dados.

Atualmente, é possível encontrar na web diferentes projetos em desenvolvimento que buscam estruturar grandes quantidades de dados já existentes na rede para a lógica da Linked Data. Alguns projetos focam seus repositórios para determinados domínios (ex.: apenas para a saúde ou para conteúdos relacionados à música); porém, dois grandes projetos se destacam por terem já estruturadas grandes quantidades de dados de múltiplos domínios: o Freebase[81] e o DBpedia[82]. Ambos fazem uma reestruturação dos dados publicados na Wikipédia e os publicam em formatos compatíveis com o RDF. Embora semelhantes, os dois projetos apresentam algumas diferenças[83]: enquanto o DBpedia tem como única fonte de dados a Wikipédia, o Freebase também toma como fonte de dados outros sites da web. Outra diferença é que cada projeto utiliza o seu próprio "*schema*", ou seja, cada um possui uma estrutura própria de propriedades (predicados). Na Figura 17, é possível observar parte dos

---

[79] *1. Use URIs as names for things. 2. Use HTTP URIs so that people can look up those names. 3. When someone looks up a URI, provide useful information, using the standards (RDF*, SPARQL). 4. Include links to other URIs. so that they can discover more things.*

[80] Uma questão fundamental para o funcionamento desta grande base de dados é que, segundo Segaran et al. (2009), o Linked Data não fornece mecanismos para que sites externos insiram dados nos grafos que fazem parte desta rede de dados, fornece apenas a função de recuperação (*query* via SPARQL).

[81] http://www.freebase.com/

[82] http://www.dbpedia.org/

[83] Essas diferenças foram publicadas pelo site do Freebase. Disponível em: <http://wiki.freebase.com/wiki/DBPedia>. Acesso em: 17 set 2011.

dados estruturados relativos ao termo “São Paulo” no site do projeto DBpedia, apresentados em formato de tabela. Logo, se tomarmos São Paulo como sujeito, teremos, na coluna à esquerda, uma lista de propriedades (predicados) e, à direita, a lista dos valores correspondentes (objetos). Ainda na Figura 17, destacamos duas linhas da tabela: a oitava linha (ver 1ª flecha vermelha) indica que a propriedade *name* (nome) tem como objeto “São Paulo”, já a nona linha (ver 2ª flecha vermelha) tem como propriedade *nickname* (apelido) os objetos “Terra da Garoa” e “Sampa”.

Os dados estruturados disponíveis nos projetos Freebase e DBpedia podem ser utilizados como metadados por qualquer site. Logo, um site que aplica a lógica das triplas pode utilizar tais metadados no lugar do sujeito ou do predicado (BIZER et al., 2009). Isto pode ser vantajoso, pois se diferentes sites da web utilizam um mesmo endereço na referência a um sujeito ou a um objeto, então eles acabam por se referir ao mesmo significado para tal sujeito ou tal objeto.

Figura 17 – Tela que mostra parte dos dados estruturados relativos ao termo “São Paulo” no site do projeto DBpedia[84]

[84] Disponível em: <http://dbpedia.org/page/São_Paulo>. Acesso em: 17 set 2011.

A rede de iniciativas em conformidade com o Linked Data cresce a cada ano. Geralmente, essas iniciativas se conectam umas às outras para que os dados publicados em um domínio sejam aproveitados por outro domínio. Por exemplo: um repositório sobre músicas pode reaproveitar os dados de um repositório sobre eventos musicais, e vice-versa. Assim, com o crescimento do número de projetos e do número de relacionamentos, é criada uma rede semântica conhecida como *Linked Data Cloud* (nuvem de dados lincados), ou simplesmente *Cloud of Data* (SEGARAN et al., 2009). Na Figura 18, é possível perceber como era esta nuvem em maio de 2007 em um diagrama publicado pelo site do projeto[85]. Na Figura 19, está o mesmo diagrama, porém atualizado em 19 de setembro de 2011, ou seja, após quatro anos de crescimento. Entre os nós do diagrama de 2011, é possível encontrar sites como o DBpedia e o Freebase.

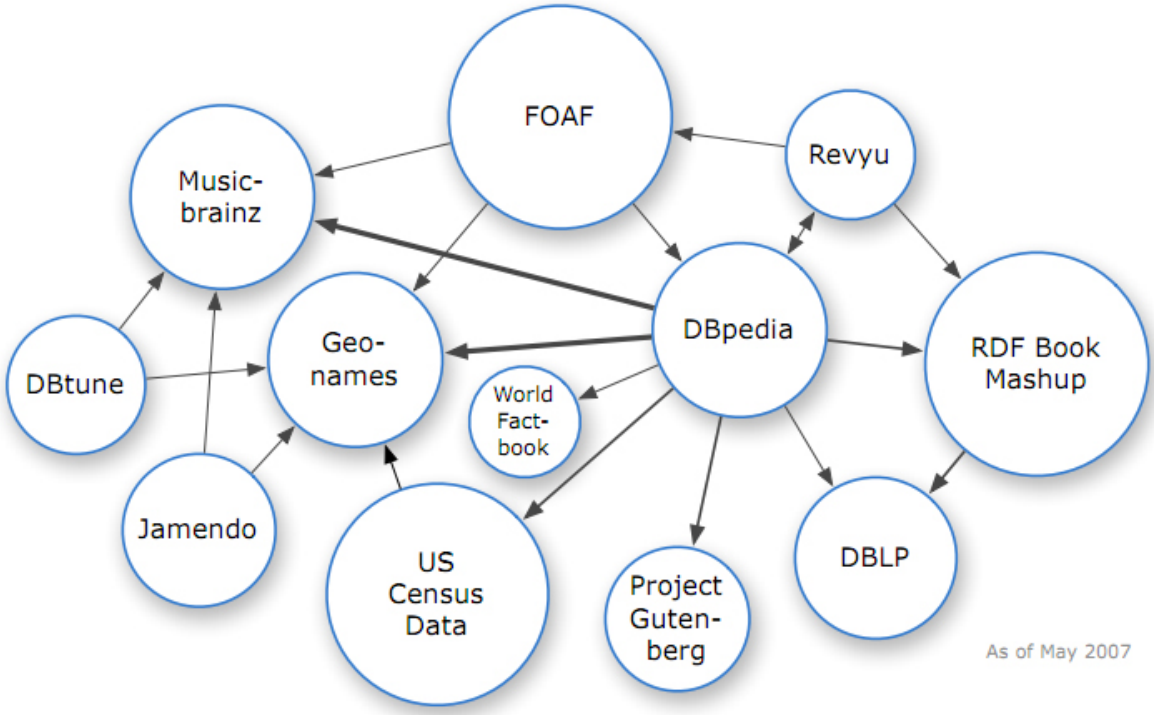


Figura 18 – Diagrama do Linked Data, atualizado em maio de 2007[86]

[85] Linked Data. Disponível em: <http://linkeddata.org/>. Acesso em: 12 jan. 2012.
[86] Disponível em: <http://richard.cyganiak.de/2007/10/lod/>. Acesso em: 12 jan. 2012.

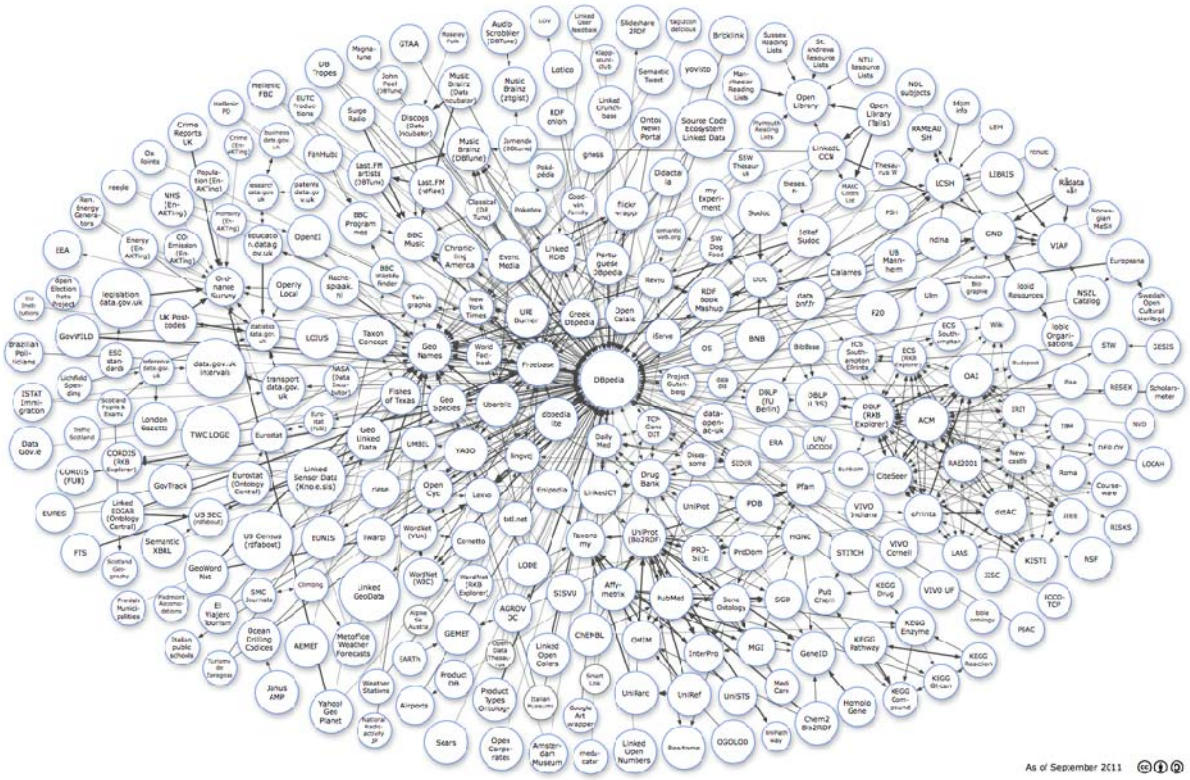


Figura 19 – Diagrama do Linked Data, atualizado em 19 de setembro de 2011[87]

A nuvem de dados do Linked Data está amarrada pelas triplas RDF que utilizam como identificadores as URIs dos projetos envolvidos, e é por isso que essas triplas funcionam como links RDF (BIZER et al., 2009). Tal situação permite que os serviços que acessam ao menos um dos projetos envolvidos possam seguir as trilhas destes links RDF e assim encontrar os dados dos outros projetos vinculados. Segaran et al. (2009) apresenta um exemplo hipotético que envolve quatro iniciativas que fazem parte do Linked Data: o DBpedia, o Freebase, o MusicBrainz (uma enciclopédia sobre música, que armazena dados e metadados sobre entidades do mundo da música e os identifica com um identificador padrão) e o site de *reviews* sobre música da BBC, que identifica os artistas, as bandas e as músicas com identificadores (IDs) do MusicBrainz. O autor apresenta um aplicativo fictício que responde a questões relacionadas à música. Então, o autor formula a seguinte pesquisa: buscar *reviews* de álbuns musicais produzidos por artistas de um determinado país que não seja os EUA. Tais *reviews* estão armazenados no site da BBC; contudo este site não armazena o local de origem das bandas, então é necessário buscar esta informação em outro local da nuvem de dados. Logo, para buscar a resposta, o aplicativo realiza pesquisas em diversas coleções de

[87] Disponível em: <http://richard.cyganiak.de/2007/10/lod/>. Acesso em: 12 jan. 2012.

metadados do Linked Data, detalhadas na Figura 20. Para um melhor entendimento, inserimos marcações na figura e apresentamos a explicação do processo em um passo-a-passo:

1) O processo começa com a entrada do nome do país no aplicativo.
2) É realizada uma pesquisa em SPARQL no repositório da DBpedia por bandas localizadas no referido país. A pesquisa encontra resultados, porém o DBpedia não utiliza IDs do MusicBrainz, necessários para que possamos encontrar os *reviews* na BBC, já que os *reviews* são indexados com IDs do MusicBrainz.
3) Como o Freebase é compatível tanto com o MusicBrainz quanto com o DBpedia, então o aplicativo fictício recebe os resultados da pesquisa no DBpedia em formato de IDs do Freebase.
4) Os IDs do Freebase são enviados como uma nova pesquisa ao respectivo repositório.
5) São recebidos novos resultados, porém no formato de ID do MusicBrainz.
6) Por fim, com as identificações das bandas selecionadas no formato de ID do MusicBrainz, basta enviar estes IDs como nova pesquisa contra o repositório da BBC.
7) O aplicativo recebe finalmente os *reviews* solicitados.

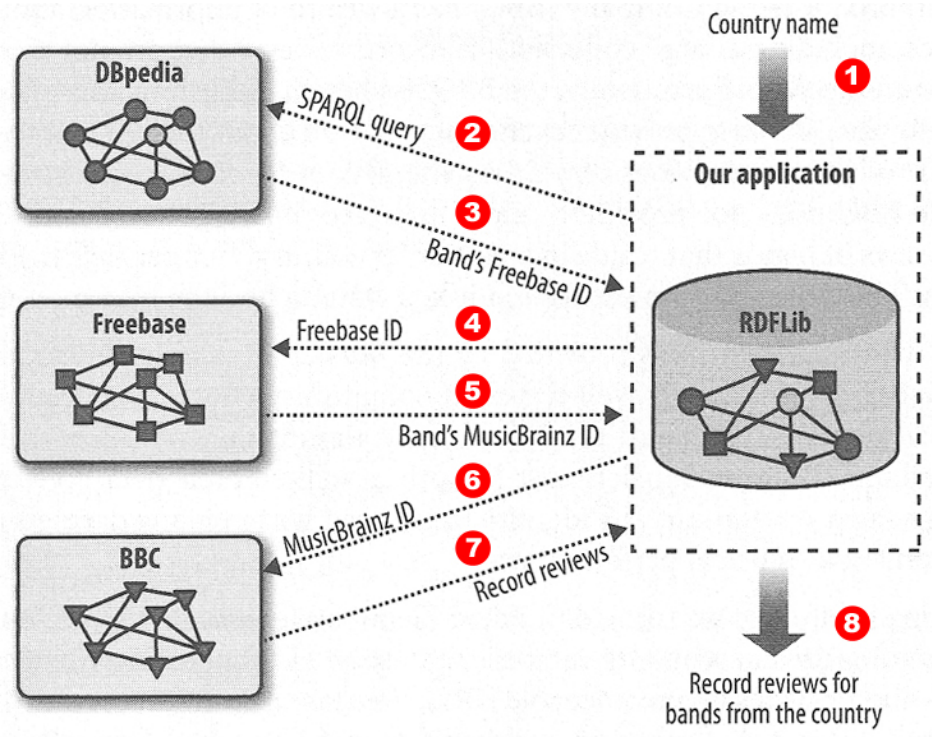


Figura 20 – Diagrama com fluxo de pesquisas na nuvem de dados para aplicativo fictício (SEGARAN et al., 2009, p. 112) com marcações que indicam a ordem das pesquisas (marcação nossa)

Embora pareça um processo complicado, nem sempre é necessário realizar um percurso burocrático como este, pois o exemplo foi apresentado pelos autores como exercício de compreensão sobre o funcionamento do Linked Data.

Neste capítulo, apresentamos uma explanação detalhada sobre alguns dos principais conceitos que constituem a Web Semântica: os metadados para as máquinas, as triplas em RDF e as ontologias em OWL. Buscamos tratar sobre os termos que surgem na análise dos casos aqui estudados e apresentados no próximo capítulo. Existem outras concepções e propostas tecnológicas para a Web Semântica além da proposta da W3C (AKERKAR, 2009), entretanto, mantivemos o foco nas tecnologias empregadas pelos produtos da BBC, escolhidos como casos para estudo desta dissertação. A seguir, partimos para a apresentação dos casos estudados e para a análise dos dados.

3 A WEB SEMÂNTICA NO JORNALISMO DIGITAL

As vantagens de um sistema semântico global alcançariam diversas áreas que trabalham informação. Souza e Alvarenga (2004) citam algumas dos benefícios esperados para a atividade dos profissionais da ciência da informação, tais como:

- projetos de novos e melhorados motores de busca,
- construção de interfaces com o usuário para sistemas de informação,
- construção automática de tesauros e vocabulários controlados,
- indexação automática de documentos,
- gestão do conhecimento organizacional,
- gestão da Informação Estratégica e da Inteligência Competitiva.

Especificamente no jornalismo digital, podemos especular diversos destes benefícios, devido à natureza informativa da área. Bertocchi (2010) cita pelo menos cinco formas como as tecnologias semânticas podem colaborar na produção e distribuição das narrativas jornalísticas em ambiente digital:

> [...] na entrega informativa (como os dados chegam aos usuários, em quais dispositivos e com qual aparência); na pesquisa de dados (como as notícias são recuperadas pelos usuários); na exploração e visualização dos dados (como as informações são visualmente apresentadas aos usuários) e, ainda, na forma de percepção do texto (como as notícias são compreendidas pelos usuários) (BERTOCCHI, 2010, p. 8).

É possível perceber que grande das vantagens esperadas de uma rede semântica se refere à organização e ao gerenciamento das informações. Com esse pressuposto, partimos para o seguinte problema de pesquisa: quais seriam as potencialidades que a Web Semântica ofereceria para a organização e o gerenciamento dos conteúdos jornalísticos? Para isso, propusemos analisar dois casos que já tenham utilizado as tecnologias semânticas para esse gerenciamento. A seguir, descrevemos como foi o processo de seleção dos dois casos: o BBC World Cup 2010 e o BBC Wildlife.

3.1 Seleção do *corpus* da pesquisa

Para a seleção do *corpus* a ser analisado, partimos das indicações encontradas em pesquisas bibliográfica e documental, como em artigos, documentos, entrevistas, debates e

apresentações disponibilizados na web por autores, especialistas, desenvolvedores, jornalistas e entusiastas que trabalham com o tema da Web Semântica. Para delimitarmos o universo da análise, determinamos que os casos a serem selecionados devessem ser produtos desenvolvidos por iniciativas oriundas do *mainstream* jornalístico, ou seja, de organizações consolidadas que possuam capacidade de investimento em pesquisa e tecnologia e que já apresentam uma grande audiência. Dessa forma, nos certificamos de que os produtos fazem parte de um projeto editorial de jornal e que tenham passado pelo crivo de uma base consistente de usuários. Após listarmos casos de referência populares em citações encontradas nas pesquisas bibliográfica e documental, foi realizada uma observação livre nos produtos pré-selecionados. Esta primeira etapa do processo de seleção do *corpus* de pesquisa resultou em produtos experimentais desenvolvidas por duas organizações europeias (BBC e The Guardian) e uma norte-americana (The New York Times).

A análise de produtos da Web Semântica exige mais do que a observação direta dos mesmos a partir de suas interfaces. É necessária, também, uma investigação sobre o funcionamento interno destes produtos. Tal situação ocorre porque nem sempre é possível observar as tecnologias semânticas em funcionamento a partir do produto final, pois geralmente tais tecnologias executam operações no servidor e esse, por sua vez, envia ao *software* navegador apenas o resultado final das operações semânticas[88]. Este mesmo problema ocorre no estudo aprofundado de produtos jornalísticos com arquitetura da informação baseada em bases de dados: em tais produtos, a análise exige o conhecimento tanto da estrutura interna (*back-end*) quanto da interface externa e pública do produto (*front-end*) (PALACIOS e NOCI, 2009). Ao considerarmos estas restrições, concluímos que os casos analisados deveriam suprir ao menos uma das duas condições seguintes:

- ser acessível ao pesquisador o suficiente para possibilitar a coleta de dados primários junto aos funcionários da empresa, a fim de se compreender o funcionamento das tecnologias semânticas empregadas no produto;
- apresentar dados secundários consistentes e diversificados, tais como depoimentos, bibliografias, documentos, debates, apresentações e outros textos que abordem a funcionalidade das tecnologias semânticas empregadas no produto.

[88] Em uma página especial da W3C que disponibiliza perguntas e respostas sobre a Web Semântica, há a confirmação desta situação: na pergunta "eu vou 'enxergar' a Web Semântica no meu navegador do dia a dia?", a resposta apresentada pela organização foi: "não necessariamente, ao menos não diretamente. As tecnologias da Web Semântica podem agir por baixo dos panos, resultando em uma melhor experiência do usuário, ao invés de influenciar diretamente no 'visual' do navegador" (tradução nossa). Disponível em: <http://www.w3.org/2001/sw/SW-FAQ#swonbrowser>. Acesso em: 27 nov 2011.

Ao considerarmos tais condições, concluímos que a primeira condição não seria viável, devido às diferenças geográficas e culturais e ao curto período da pesquisa, o que, ao combinarmos tais entraves, previmos que impossibilitariam a realização de uma eventual série de entrevistas com diversos funcionários da organização. Tomamos, então, como requisito para a seleção do *corpus* a segunda condição, ou seja, a existência de grande quantidade de dados secundários que abordem o funcionamento do produto. Por esta razão, entre as organizações pré-selecionadas, decidimos pela BBC, por apresentar não apenas maior quantidade de dados secundários, mas também por ter demonstrado o uso de tecnologias semânticas em mais de um produto digital. Outra justificativa pela sua escolha é pelo fato de que os dados secundários foram produzidos diretamente pelos funcionários envolvidos no desenvolvimento dos produtos, em relatos dispersos na web, caracterizando tais dados como verdadeiros depoimentos, o que nos aproxima da qualidade dos dados primários.

A BBC é a maior emissora de rádio e televisão do Reino Unido (de acordo com a própria BBC, é a maior do mundo[89]). A organização tem tradição na implantação de tecnologias digitais em seus produtos, como câmeras de alta definição para documentários e canais de televisão interativos. Para realizarmos a nossa investigação, selecionamos dois produtos digitais da BBC, cada um deles como um caso a ser estudado: o site **BBC World Cup 2010**[90] e o site **BBC Wildlife**[91]. A seguir, passamos para a identificação e descrição de cada um dos casos. Após esta descrição, apresentamos uma análise sobre como as tecnologias semânticas identificadas nos casos estudados atuam nas categorias do Jornalismo Digital em Base de Dados e como contribuem para a organização e o gerenciamento do conteúdo jornalístico.

3.2 Caso BBC World Cup 2010

O BBC World Cup 2010 é um site jornalístico especial da BBC para a Copa do Mundo de 2010. Funciona como um portal para abrigar todo o conteúdo jornalístico da BBC relacionado ao evento (notícias, blogs, perfis, imagens, vídeos e estatísticas). Embora a Copa tenha sido finalizada há aproximadamente um ano e meio desde a produção desta pesquisa, o site continua *online*, com todas as suas funcionalidades. A publicação e a organização do

[89] "*The BBC is the largest broadcasting organisation in the world. Its mission is to enrich people's lives with programmes that inform, educate and entertain*". Disponível em: <http://www.bbc.co.uk/aboutthebbc/purpose/what.shtml>. Acesso em: 4 dez 2011.
[90] http://news.bbc.co.uk/sport2/hi/football/world_cup_2010/
[91] http://www.bbc.co.uk/nature/wildlife

conteúdo são realizadas de forma automatizada, graças às tecnologias semânticas. Embora publique conteúdos unicamente da editoria de esportes, foi um produto jornalístico que se aproximou do modelo de publicação de *hard news*[92], devido à alta frequência de visitação de usuários e à intensa produção de conteúdos no período de cobertura do evento esportivo.

No APÊNDICE B, apresentamos os profissionais que serviram como fontes de dados secundários para a identificação e descrição das tecnologias semânticas. Além das produções dos profissionais, foram consultados documentos disponibilizados pela própria BBC, como a página em que é descrita a ontologia do BBC World Cup 2010.

3.2.1 Descrição do produto

O site possui dois tipos de conteúdos: o jornalístico (informativo e opinativo), que na época da Copa do Mundo 2010 era constantemente atualizado pelos jornalistas, e um conteúdo permanente de referência utilizado para descrever três grupos-chave de assuntos relativo à Copa do Mundo, que são constantemente citados nas narrativas jornalísticas: **times, jogadores e grupos**. Este conteúdo permanente serve como uma base de conhecimento para a construção dinâmica das diversas páginas que fazem o site. Para cada unidade individual que faz parte dos elementos citados (ou seja, para cada time, cada jogador e cada grupo da Copa), existe uma página única que reúne, de forma automatizada, diversos tipos de conteúdos relacionados ao assunto da página. No decorrer da competição, foram criadas, também, páginas únicas para cada partida realizada. As páginas dos grupos, dos times, dos jogadores e das partidas somam ao todo 832 unidades[93].

As páginas dos conteúdos de referência (grupos, times e jogadores) apresentam interface semelhante umas com as outras: são três colunas, sendo que a primeira é igual para todas (links para as últimas partidas do evento), já as outras duas colunas reúnem, de forma automatizada, dados atualizados sobre o elemento em questão (são as colunas que nos interessam, pois é o local de publicação dinâmica do conteúdo contextualizado). O que

[92] Para Tuchman (1978), podemos identificar tipos de conteúdos jornalísticos. A autora destaca dois tipos principais: os *hard news*, que são notícias "importantes para os seres humanos" (TUCHMAN, 1978, p. 48, tradução nossa), ou seja, "informações que as pessoas deveriam ter para se tornarem cidadãs informadas" (idem); e as *soft news*, que são notícias "interessantes porque lidam com a vida dos seres humanos" (idem), ou, em outras palavras, "diz respeito às fraquezas humanas e à textura da nossa vida humana'" (idem). Entendemos neste trabalho as *hard news* como notícias factuais e de interesse público, e as *soft news* como notícias de interesses de públicos específicos, relacionados à vida privada ou a questões de interesse humano, e que não se encaixam em editorias de grande relevância no exercício da cidadania, tais como política, economia e geral.

[93] Cálculo baseado nos seguintes números: 32 times, 23 jogadores por time, 8 grupos da Copa, 6 partidas por grupo, 8 partidas da 2ª fase, 4 partidas das quartas-de-final, 2 partidas das semifinais, 1 partida da final, 1 partida do 3º colocado. Então: 32 times + 736 jogadores + 64 partidas = 832 páginas.

diferencia as páginas dos times das páginas dos jogadores ou de grupos é a inclusão de dados específicos para cada tipo de entidade, como estatísticas apropriadas para cada elemento.

A **página dos times** (ver indicações na Figura 21) apresenta, na coluna central: A) as últimas partidas da seleção em questão com os respectivos resultados, B) as últimas notícias, C) as últimas mídias, D) os últimos artigos de opinião, E) uma galeria de fotos, F) algumas estatísticas sobre a eficiência do time na competição, G) um perfil do time (com brasão oficial e links para perfil estendido e estatísticas estendidas), H) uma tabela com a lista de jogadores com informações básicas sobre os mesmos (cada nome de jogador é um link para a página do respectivo), I) uma lista maior das últimas notícias sobre o time e, por fim, na parte final da coluna, J) uma lista que mostram links para conteúdos relacionados ao time em questão. Na coluna da direita, a página apresenta: K) a tabela do grupo em que o time faz parte, L) uma lista com reportagens especiais, M) uma lista de links para conteúdos relacionados que estejam fora do site da BBC, e, por fim, N) a lista das cinco matérias mais lidas.

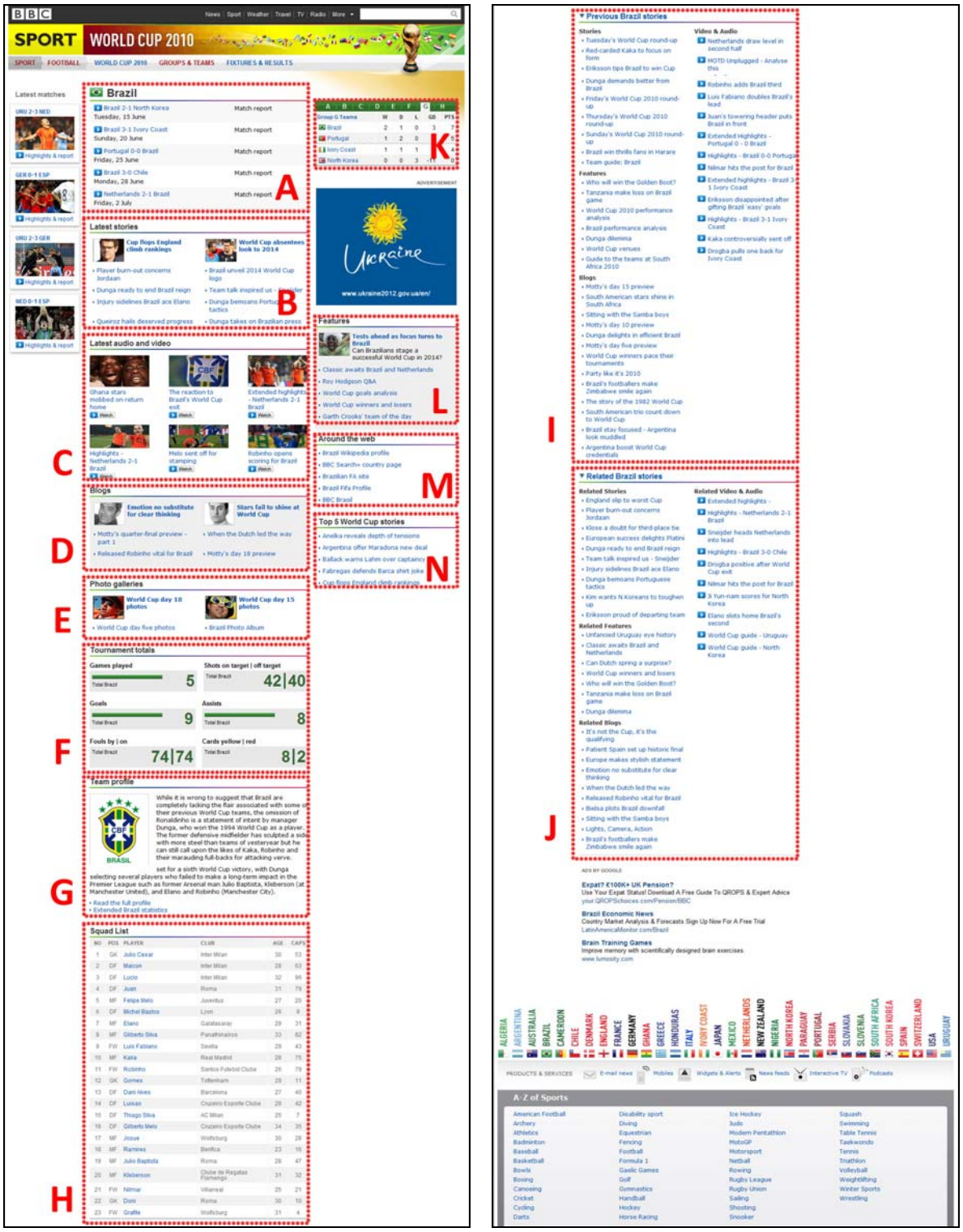


Figura 21 – Página dos times (Seleção brasileira), dividida em duas partes[94]

A **página dos jogadores** (ver Figura 22) apresenta, na coluna central: A) a identificação do jogador (nome, nacionalidade, posição, número da camisa, data de

[94] Disponível em: <http://news.bbc.co.uk/sport2/hi/football/world_cup_2010/groups_and_teams/team/brazil>. Acesso em: 11 dez. 2011.

nascimento e altura), B) as estatísticas do desempenho na competição, C) a lista de partidas (com resultados) em que jogou junto a sua seleção, D) as duas últimas notícias em que consta seu nome, E) últimas mídias em que é mencionado, F) posts opinativos de blogs em que é mencionado, G) um perfil biográfico do jogador e, por fim, H) uma lista maior de últimas entradas (notícias, mídias, posts etc) em que o jogador é mencionado. Na coluna à direita, I) há apenas a lista das cinco matérias mais relevantes sobre seu time (Top 5).

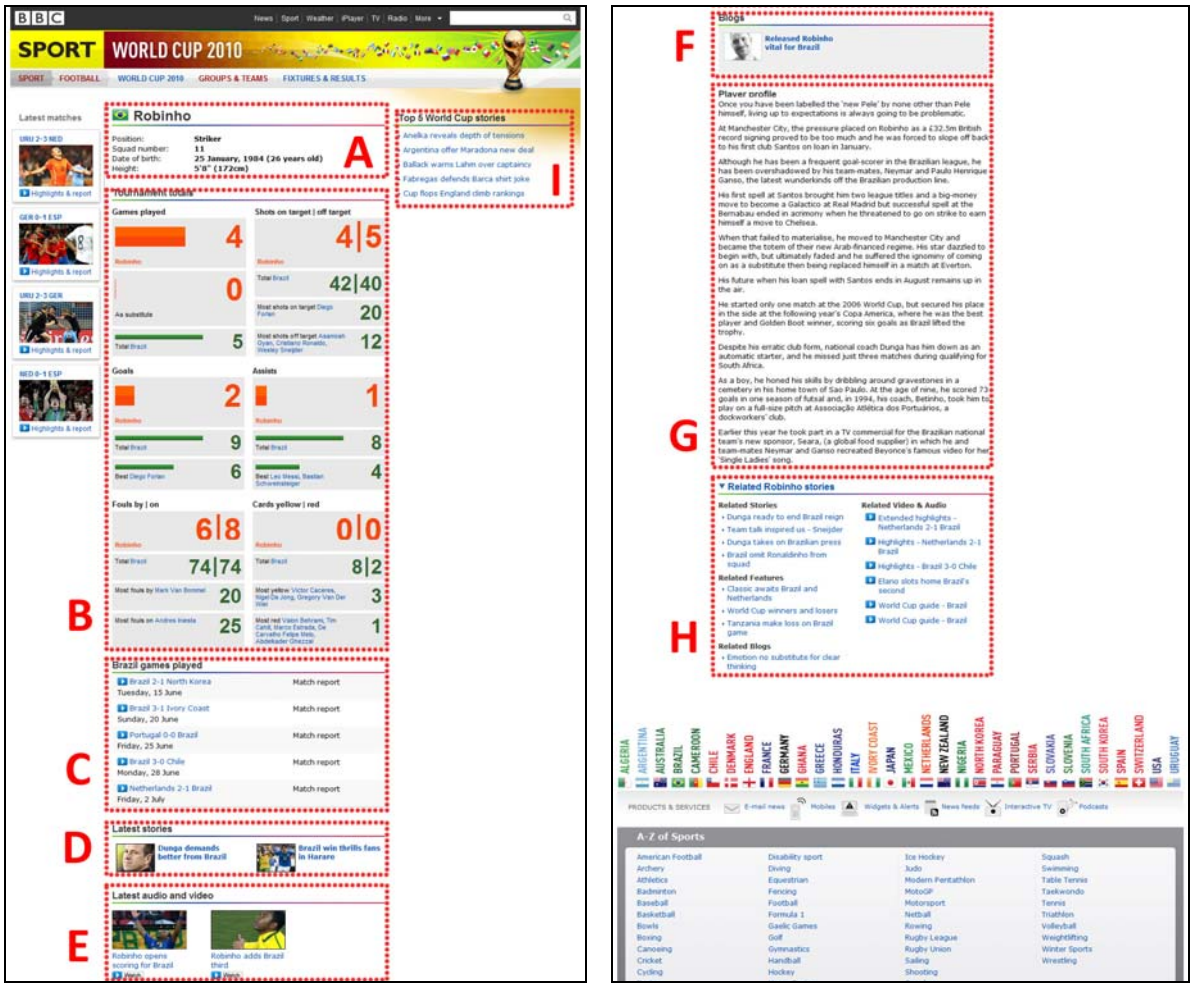


Figura 22 – Página dos jogadores (jogador Robinho), dividida em duas partes[95]

A **página dos grupos** (ver Figura 23) apresenta, na coluna central, os seguintes espaços: A) a tabela de times com estatísticas para cada seleção (jogos, vitórias, derrotas, pontos etc), B) a lista das últimas notícias, C) as últimas mídias (áudio e vídeo), D) os últimos artigos de opinião (posts de blogs), E) uma galeria de fotos, F) uma lista maior das últimas notícias sobre o grupo e, por fim, na parte final da coluna, G) há uma lista que mostram links

[95] Disponível em: <http://www.bbc.co.uk/sport/0/football/world_cup_2010/groups_and_teams/team/brazil/robinho/>. Acesso em: 11 dez. 2011.

para conteúdos relacionados ao grupo em questão. Na coluna da direita, a página apresenta: H) a lista de partidas do grupo, I) algumas reportagens especiais e J) a lista das 5 notícias mais relevantes (Top 5).

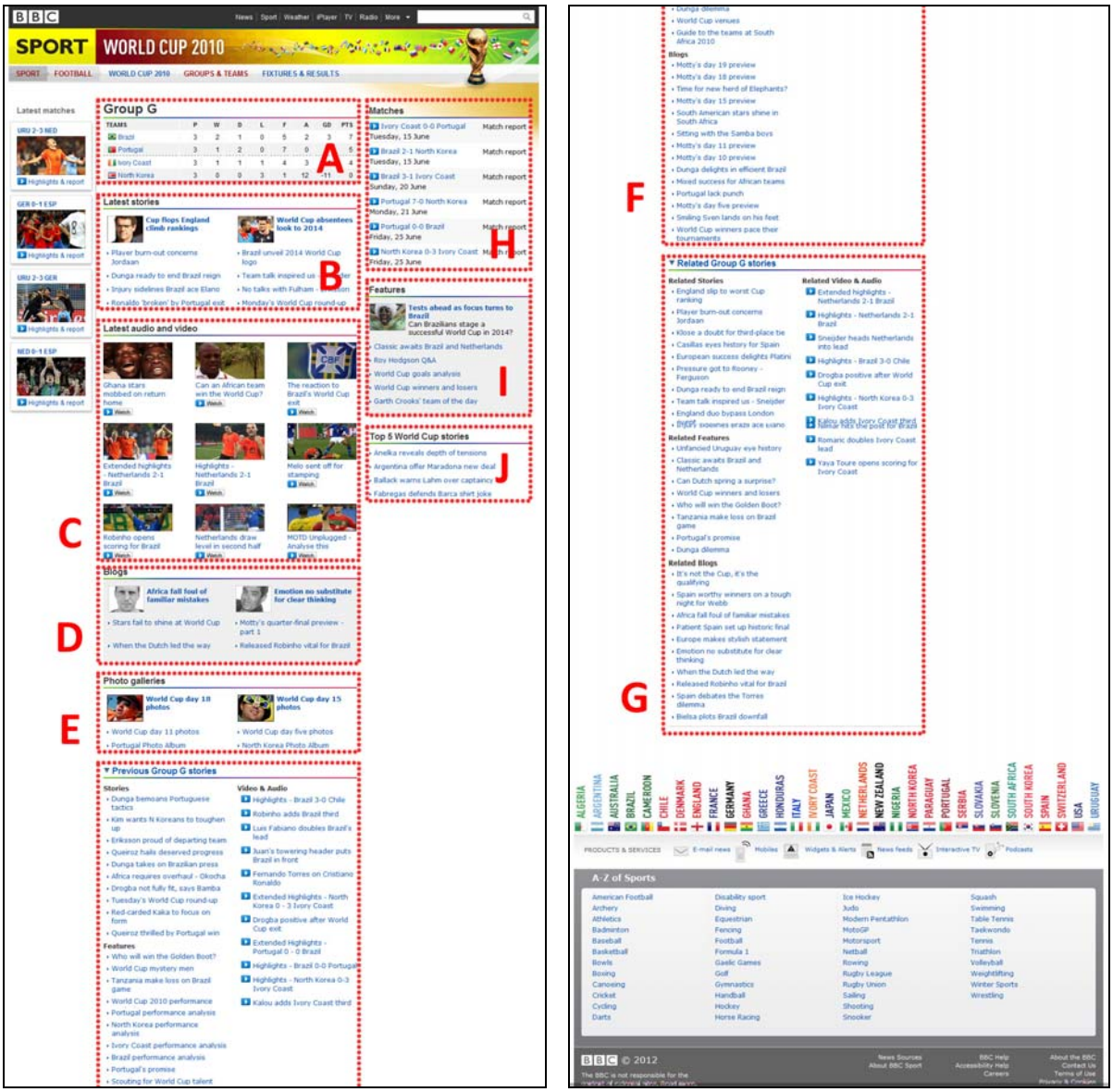


Figura 23 – Página dos grupos (grupo G), dividida em duas partes[96]

A **página das partidas** (ver Figura 24) segue um layout diferenciado em relação às outras. Além do resultado, ela apresenta duas grandes áreas: na área A, é mostrado o relato da partida, feita por um jornalista. Nesta mesma área, há uma aba intitulada "Comentário", que mostra um fluxo cronológico de mensagens publicadas no período do jogo, misturando relatos do narrador e comentários de jornalistas e usuários do site, originados do próprio site, de

[96] Disponível em: <http://www.bbc.co.uk/sport/0/football/world_cup_2010/groups_and_teams/group_g/>. Acesso em: 11 dez. 2011.

outros sites da BBC, da rede Twitter e do serviço de mensagens SMS. Na área B, há diversas estatísticas distribuídas em três abas: na 1ª, a tabela de informações sobre a partida (resultado final, jogadores que marcaram gol, escalação, cartões, jogadores substituídos, estádio, juiz e público total), na 2ª aba, são apresentadas estatísticas técnicas do jogo (tempo de possessão da bola, quantidade de escanteios e faltas etc), e na 3ª aba, a tabela do grupo projetada logo após o resultado da partida. Na coluna da direita (área C), há apenas informações não relacionadas à partida (anúncios, links para sites sobre a Copa e informações sobre como enviar mensagem para a aba "Comentário"). Na Figura 24, é mostrada a página de uma partida, com o relato do jogo (A) e as informações relacionadas à partida (B).

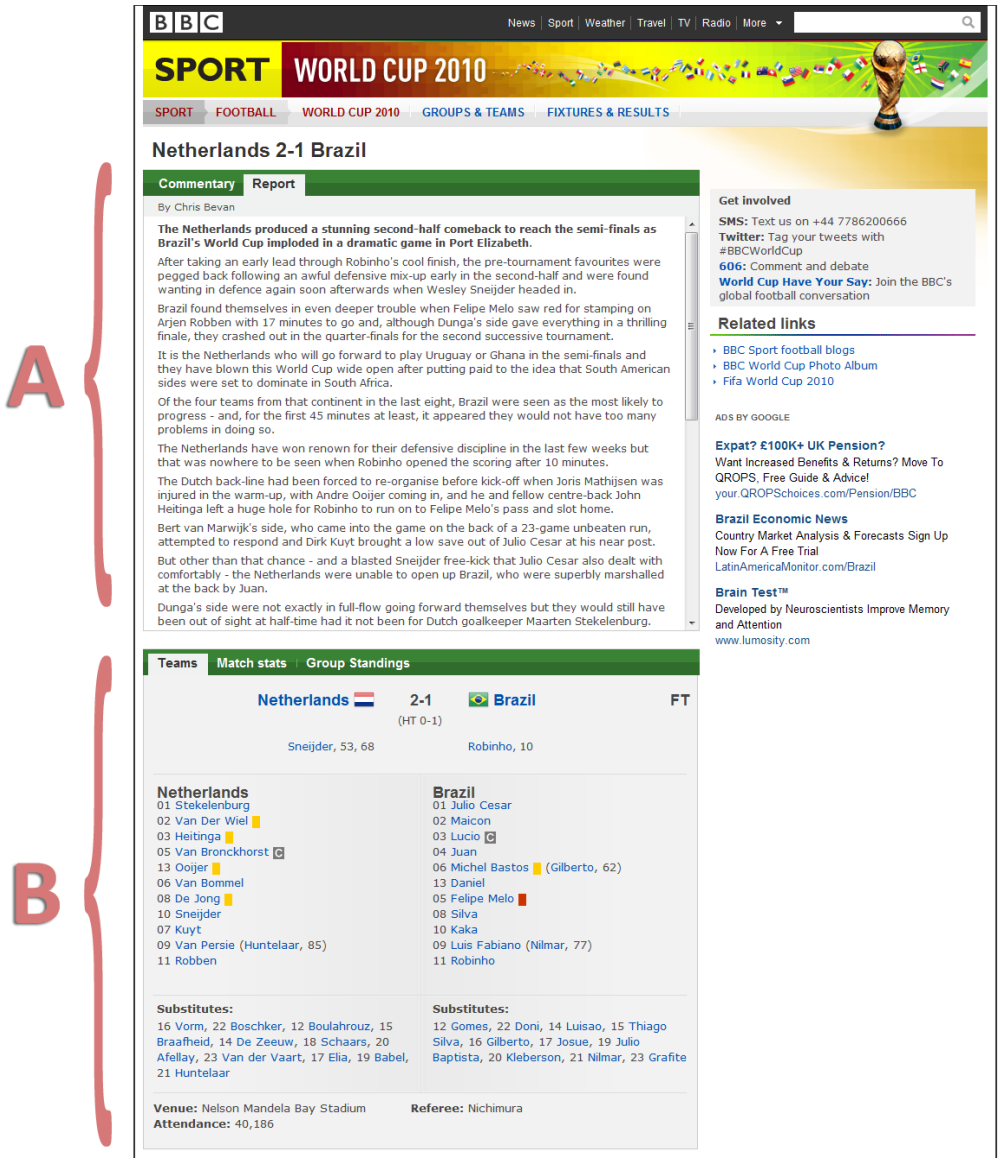


Figura 24 – Página das partidas, com o relato (A) e as informações (B) sobre o jogo[97]

[97] Disponível em: <http://news.bbc.co.uk/sport2/hi/football/world_cup_2010/matches/match_57/default.stm>. Acesso em: 12 dez. 2011.

A Figura 25 mostra a página da mesma partida, porém com outras abas selecionadas (comentários na indicação A e estatísticas na indicação B).

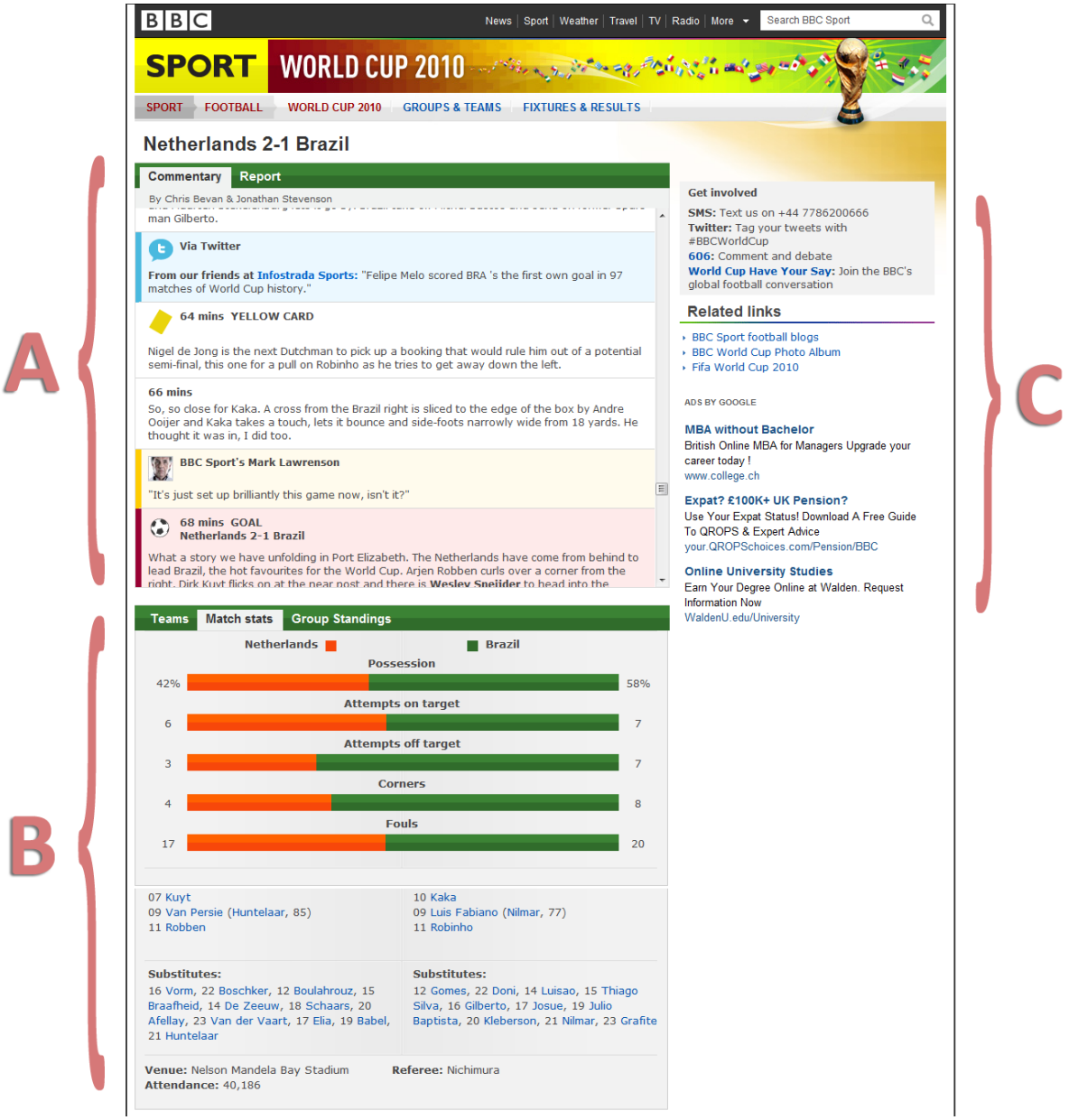


Figura 25 – Página da partida, com comentários (A) e estatísticas (B) sobre o jogo[98]

Nas **páginas das matérias** (Figura 26), em que são publicadas notícias e reportagens analíticas, a coluna central é ocupada apenas pela narrativa jornalística da matéria. Na coluna da direita, há três listas de links relacionados ao texto: A) para matérias do site World Cup

[98] Disponível em: <http://news.bbc.co.uk/sport2/hi/football/world_cup_2010/matches/match_57/default.stm>. Acesso em: 12 dez. 2011.

2010, B) para matérias de sites da BBC e C) para sites externos. A narrativa das matérias apresenta frequentemente elementos diferentes do textual, como imagens, vídeos e caixas (*box*). Embora exista um grande potencial para a lincagem do texto com as páginas dos times e dos jogadores, muitos textos não aproveitam este recurso, e são publicados sem link algum.

Figura 26 – Visão parcial da página de notícia, com marcações em três listas de *links*[99]

[99] Disponível em: <http://news.bbc.co.uk/sport2/hi/football/world_cup_2010/8823478.stm>. Acesso em: 11 dez. 2011.

Não existe uma página especial para os artigos de opinião, pois os links sempre remetem o usuário ao blog do respectivo colunista/articulador.

A **página inicial** (ver Figura 27) é um *hub* que reúne links para os conteúdos jornalísticos. É dividida em 3 colunas: na da esquerda, A) há uma lista dos artigos de opinião (blogs); na central, são mostradas B) as chamadas para as últimas matérias (notícias, reportagens), C) as chamadas para matérias aprofundadas, D) as últimas notícias sobre as seleções finalistas, E) notícias sobre a Copa em outros sites da web, e F) links para sites relacionados à Copa, como o site da FIFA. Na terceira coluna, são apresentadas: G) uma lista automática dos maiores goleadores da competição, H) uma lista das últimas mídias produzidas pela BBC (vídeos e áudios) e I) mídias e notícias oriundas de sites internacionais da BBC sobre o evento.

Figura 27 – À esquerda, uma visão parcial da página inicial do site World Cup 2010. À direita, a mesma página, porém completa e com marcações que indicam as áreas relatadas[100]

[100] Disponível em: <http://news.bbc.co.uk/sport1/hi/football/world_cup_2010/>. Acesso em: 11 dez. 2011.

Além da página inicial, há duas outras páginas que funcionam como *hubs*, porém para as páginas dos times, dos jogadores, dos grupos e das partidas. A página *Groups and Teams* (Figura 28) mostra todos os oito grupos da Copa, cada um em uma tabela, além do mapa dos confrontos realizados após a fase dos grupos A página *Fixtures and results* (Figura 29) apresenta um calendário com todas as partidas da Copa, em que os resultados são mostrados para os jogos já realizados.

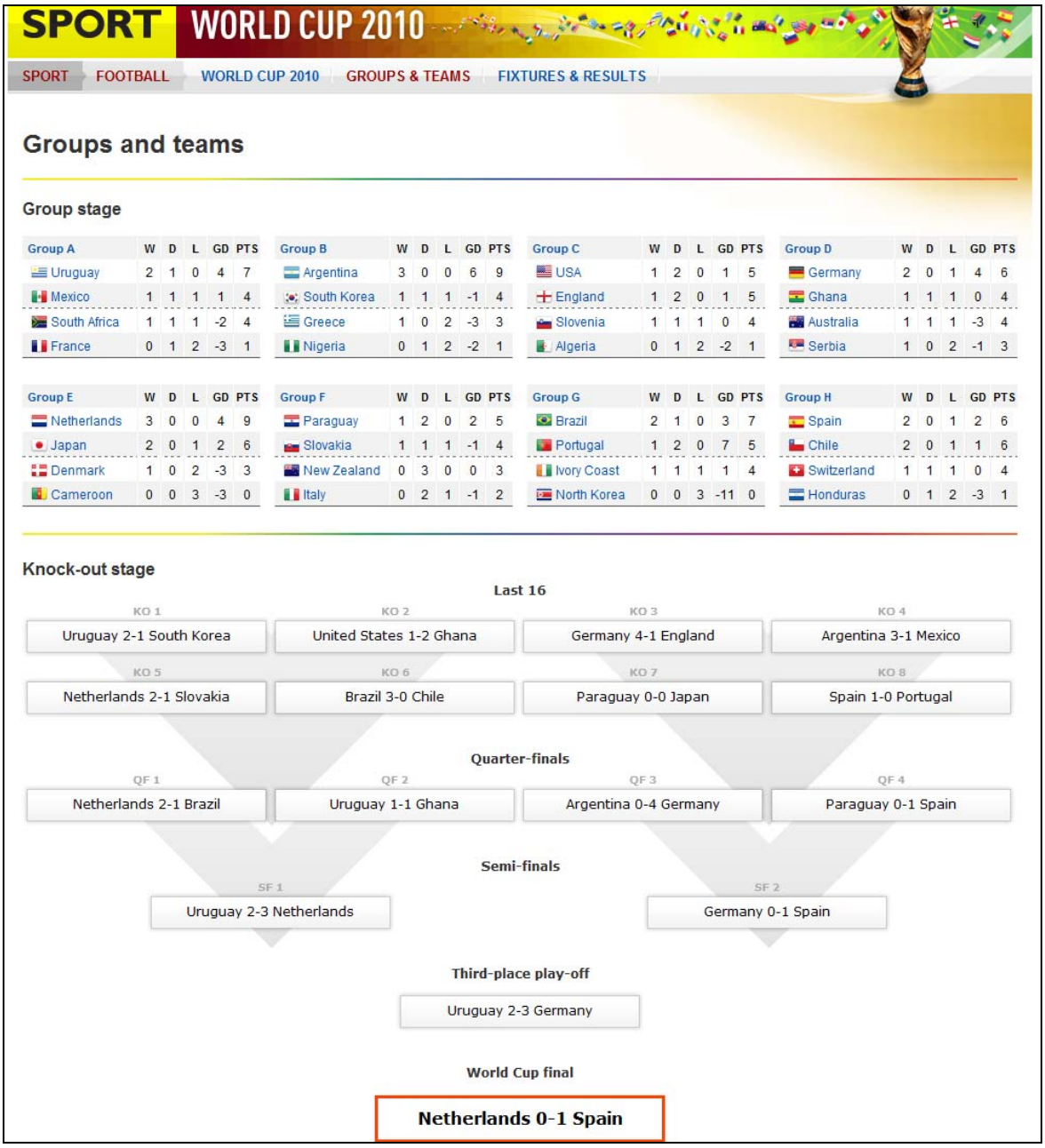


Figura 28 – Página *Groups and teams.* Na parte superior: os oito grupos da Copa. Na parte inferior: o mapa de confrontos pós-fase de grupos[101]

[101] Disponível em: <http://news.bbc.co.uk/sport1/hi/football/world_cup_2010/groups_and_teams>. Acesso em: 17 dez. 2011.

Figura 29 – Página *Fixtures and results*[102]

A **navegação do site** é realizada através de dois menus principais: um superior e outro na base da página. No menu superior (Figura 30) as três opções oferecidas direcionam o usuário às páginas *hubs*: o link **World Cup 2010** (página inicial), o link **Groups & Teams** (página que mostra os oito grupos e os confrontos) e **Fixtures & Results** (página que mostra as partidas em um calendário, com os devidos resultados). O menu inferior (Figura 31) lista as 32 seleções participantes do evento esportivo, em que cada seleção é um link para a página do respectivo time.

[102] Disponível em: <http://news.bbc.co.uk/sport1/hi/football/world_cup_2010/fixtures_and_results>. Acesso em: 17 dez. 2011.

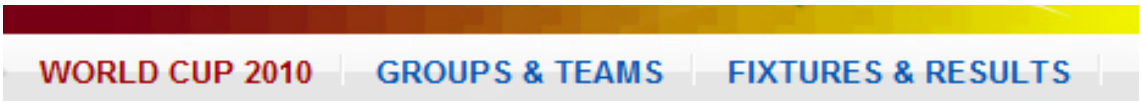


Figura 30 – Menu superior do site World Cup 2010

Figura 31 – Menu inferior do site World Cup 2010

Enquanto a opção World Cup 2010 do menu superior é a principal porta de entrada para os conteúdos jornalísticos (através das chamadas para as matérias), as outras opções dão acesso às páginas dos grupos, times, jogadores ou partidas. O abundante cruzamento de links nas páginas faz com que as próprias interfaces dos conteúdos se transformem em dispositivos de navegação. Por exemplo: a partir da página do time, é possível clicar em links para jogadores, partidas, notícias ou opiniões. Ou seja: todas as páginas do site seguem a estética base de dados, pois a estrutura visual é determinada por elementos formados a partir de pesquisas em BDs (*queries*), como listas de links, infográficos dinâmicos (não interativos) e caixas com dimensões delimitadas pelos dados dinâmicos.

Até aqui, descrevemos a interface e os funcionamentos do site BBC World Cup 2010. A seguir, passamos para a justificativa sobre a necessidade de se utilizar uma solução semântica na organização e gerenciamento do site.

3.2.2 Contexto e justificativa para uso das tecnologias semânticas

Os desenvolvedores da BBC encontraram o desafio de organizar e gerenciar um site com uma grande quantidade de conteúdos oriundos de diferentes setores da emissora, como *posts* de blogs, textos da redação da BBC News, textos do site BBC Sports e mídias de fotojornalistas e cinegrafistas. O desafio se tornou ainda maior ao considerarem que o evento envolve 32 seleções em oito grupos que somam 736 atletas e, para cada um, havia a necessidade tanto de informações permanentes (como os perfis biográficos) quanto de informações atualizadas frequentemente no período do evento (como as estatísticas e os

resultados). Segundo O'Donovan (2010), cada uma dessas páginas é uma *aggregation page* (que o autor denomina de *index page*, ou seja, uma página principal que agrega dados de um pequeno universo específico). As mais de 800 *index pages* do World Cup 2010 são em maior quantidade do que todos os *index pages* do site BBC Sports[103]. O'Donovan afirma que, normalmente, a administração de tantas *index pages* não seria possível, já que para cada uma delas seria necessário um editor com função de curador das informações publicadas, para configurar as regras de automação ou atualizar as *index pages* com as últimas matérias e estatísticas. Para ele, é clara a necessidade da automação, porém as tecnologias de busca e métodos empregados até o momento não tinham se mostrado precisos, logo seria um risco empregá-las em um sistema com tantas páginas. Como exemplo, ele afirma que não gostaria de ver informações misturadas entre páginas de jogadores com o mesmo sobrenome.

Os conteúdos produzidos pelos jornalistas já eram armazenados em bases de dados relacionais e continuaram sendo armazenadas desta maneira. O desafio não era o armazenamento, mas uma maneira de agregar estes conteúdos e construir as páginas de forma automatizada, ou seja, de publicar os conteúdos jornalísticos em determinadas páginas com o mínimo de intervenção humana.

Segundo Rayfield (2010), a escolha pelo sistema semântico na publicação de metadados, em detrimento das tradicionais bases de dados relacionais, se dá pela necessidade de interpretação dos metadados de acordo com um modelo de ontologia de um domínio, pois a ontologia permite um mapeamento inteligente dos conteúdos jornalísticos em relação a determinados significados. Rayfield exemplifica com a seguinte situação: se um jornalista associa o conceito do jogador inglês "Frank Lampard" a sua matéria, o sistema automaticamente cria inferências (através de triplas) e aplica a essa matéria conceitos como "Seleção da Inglaterra", "Grupo C" e "FIFA World Cup 2010".

Dimitrov (2010) cita a ferramenta como uma "plataforma de publicação dinâmica e semântica" (*dynamic semantic publishing platform*). Rayfield explica que o sistema semântico não seria tanto um espaço de publicação direta de conteúdos, como ocorre nos tradicionais sistemas de gerenciamento de conteúdo, mas seria mais um sistema de publicação de metadados, que permitiriam um relacionamento rico entre os conteúdos e, assim, uma navegação semântica. "Através de *queries* nesses metadados publicados, conseguimos criar dinamicamente páginas agregadas para times, grupos e jogadores" (Rayfield, 2010, *online*).

[103] http://news.bbc.co.uk/sport

3.2.3 Identificação de recursos e tecnologias semânticas utilizadas

Segundo os dados coletados a partir dos depoimentos dos desenvolvedores da BBC e de outros documentos, as principais tecnologias semânticas utilizadas no site foram as seguintes:

- Triplas em RDF, para relacionar recursos a objetos.
- Repositório semântico *triple store*[104] para gerenciamento de metadados em RDF. Foi utilizado um sistema privado, produzido pela empresa Ontotex, chamado BigOWLIM.
- Ontologia própria, de domínio (sobre a Copa do Mundo), em OWL.
- Sistema manual de etiquetagem de conteúdos (*tagging*), com auxílio de um software que já apresenta um vocabulário pré-definido (Graffiti).
- Sistema de extração automática de conceitos de conteúdos em linguagem natural (software IBM LanguageWare).
- SPARQL, para as pesquisas *query* no *triple store*.
- Dados e metadados disponíveis por terceiros na nuvem da Linked Data.

3.2.4 Descrição do funcionamento das tecnologias semânticas

O site World Cup 2010 da BBC reúne conteúdos de diversas fontes. Tais conteúdos são originalmente armazenados em bases de dados relacionais, pois são publicados via sistemas publicadores de conteúdo (CMS). O sistema semântico do site é responsável por recuperar tais conteúdos, associá-los a determinados conceitos (através de inferências automatizadas) e, a partir dessas associações, publicá-los nas páginas corretas, dentro de um universo de mais de 800 páginas. Além dos textos jornalísticos e das mídias, o sistema também é alimentado por informações estruturadas e constantemente atualizadas (*feeds*) oriundas de outros sites, como estatísticas produzidas pelo site de esportes da BBC.

Rayfield (2010) explica que o sistema de publicação dinâmico e semântico da BBC possui uma ontologia própria para o domínio do futebol, que define certos conceitos (e seus relacionamentos), tais como: jogador, time e grupo. Assim, segundo o exemplo apresentado pelo desenvolvedor, a ontologia pode inferir que "Frank Lampad" é parte do time "Seleção da Inglaterra", e que "Seleção da Inglaterra" compete no "Grupo C" da competição "FIFA World

[104] *Triple store* é a denominação dada aos repositórios de triplas em RDF. Eles são bancos de dados que, ao invés do modelo relacional (em tabelas), utilizam o modelo em *graph*. Os *triple stores* são utilizados para armazenar as ontologias e os metadados em tripla (RDF) utilizados pelo site em questão. É dentro dos *triple stores* que ocorrem as *queries* em SPARQL e as inferências nas relações entre triplas e ontologias.

Cup 2010". A ontologia também define os tipos de conteúdos que os jornalistas publicam (matérias, blogs, perfis, imagens, vídeos e estatísticas) e os relacionam com os conceitos sobre a Copa do Mundo. A BBC costuma disponibilizar na web suas ontologias, porém, atualmente, a ontologia desenvolvida para a Copa do Mundo de 2010 está mesclada a uma ontologia[105] mais geral sobre esportes, utilizada pela emissora para qualquer evento esportivo. Até o presente momento (2011), a ontologia de esportes contava com 21 classes[106] e 31 propriedades[107]. Cada entidade (também chamado *individual*) pode fazer parte de certas classes e possuir determinadas propriedades.

Para que o conteúdo jornalístico (matérias, mídias e *feeds*) possa ser associado às definições da ontologia, é necessário identificar a presença de determinados termos conceitos dentro do referido conteúdo, senão, do contrário, uma determinada matéria sobre "Seleção da Inglaterra" não poderia ser associada às páginas dos seus jogadores, do seu grupo e de suas partidas. De acordo com Rayfield (2010), para se **extrair conceitos dos conteúdos**, há dois processos complementares: um manual e outro automático. O processo manual é o de *tagging*, ou seja, o jornalista autor é responsável por associar palavras-chaves a sua matéria. Essa associação não é arbitrária: há o auxílio de uma ferramenta denominada Graffiti, utilizada para associações seletivas de determinados conceitos. Já no processo automático, um software analisa os textos e os compara aos conceitos da ontologia da Copa do Mundo. Esta análise é realizada por uma ferramenta desenvolvida pela IBM, o LanguageWare[108], um processador de linguagem natural responsável por extrair conceitos de conteúdos não estruturados (textos sequenciais, como documentos, relatórios, e-mails etc). Após essa associação automática, as *tags* são revisadas por um editor jornalista para que se mantenha a precisão e a qualidade dos metadados.

Após esta extração de conceitos, os metadados são passados para o modelo em tripla (RDF) e armazenados em um repositório *triple store*. Entre várias possíveis opções de sistemas para repositório de triplas RDF, a BBC optou por escolher uma solução comercial: o

---

[105] Ontologia desenvolvida por Jem Rayfield, Paul Wilton e Silver Oliver. Disponível em: <http://www.bbc.co.uk/ontologies/sport>. Acesso em: 7 fev. 2012.

[106] São elas: *Competition, CompetitionType, CompetitiveSportingGroup, CompetitiveSportingOrganisation, DivisionalCompetition, EventGender, FootballManagerRole, FootballPlayerRole, GroupCompetition, KnockoutCompetition, LeagueCompetition, Match, MultiRoundCompetition, MultiStageCompetition, RecurringCompetition, Round, Session, SportGoverningBody, SportingOrganisation, SportsDiscipline, UnitCompetition.* Disponível em: <http://www.bbc.co.uk/ontologies/sport>. Acesso em: 7 fev. 2012.

[107] São elas: *awayCompetitor, competesIn, competitionType, discipline, eventGender, firstRound, firstSession, firstUnitCompetition, hasRound, hasCompetitor, hasGroup, hasMatch, hasSession, hasStage, hasUnitCompetition, homeCompetitor, isCompetitiveSportingOrganisationOf, isGroupOf, isMatchOf, isRoundOf, isSessionOf, isStageOf, lastRound, lastSession, lastUnitCompetition, nextSession, nextUnitCompetition, prevSession, prevUnitCompetition, roundNumber, subDiscipline*. Disponível em: <http://www.bbc.co.uk/ontologies/sport>. Acesso em: 7 fev. 2012.

[108] Disponível em: <http://www-01.ibm.com/software/globalization/topics/languageware/index.html>. Acesso em: 27 jan. 2012.

triple store BigOWLIM[109], um sistema que, além de armazenar quantidades massivas de triplas, também tem a capacidade de gerar inferências (KIRYAKOV et al, 2010). Segundo Dimitrov (2010), o *triple store* BigOWLIM armazena ontologias, informações factuais sobre as entidades da Copa (jogadores, times, grupos, jogos etc) e os metadados associados aos conteúdos. Estes dados eram atualizados constantemente.

Para a publicação dinâmica e semântica dos conteúdos nas páginas, são realizadas *queries* (em SPARQL) no repositório *triple store* para gerar as inferências e obter os significados que determinam como as páginas deverão ser montadas. Segundo Dimitrov (2010), no período da Copa, eram realizados entre 1 e 2 milhões de *queries* por dia. Além da ontologia própria sobre a Copa do Mundo de 2010, o *triple store* também leva em consideração outras ontologias ou vocabulários externos, oriundas do Linked Data, como, por exemplo, na comparação entre o conceito de "uma seleção nacional" com os dados da DBpedia sobre a referida seleção. Então, em outras palavras, o *triple store* BigOWLIM armazena os metadados das matérias em triplas RDF e a ontologia sobre a Copa do Mundo em OWL, e no processo de inferência, integra os dados externos da Linked Data.

Oliver (2010a, 2010b) utiliza um gráfico (ver Figura 32) para demonstrar de forma simplificada como ocorre o processo de publicação semântica do site.

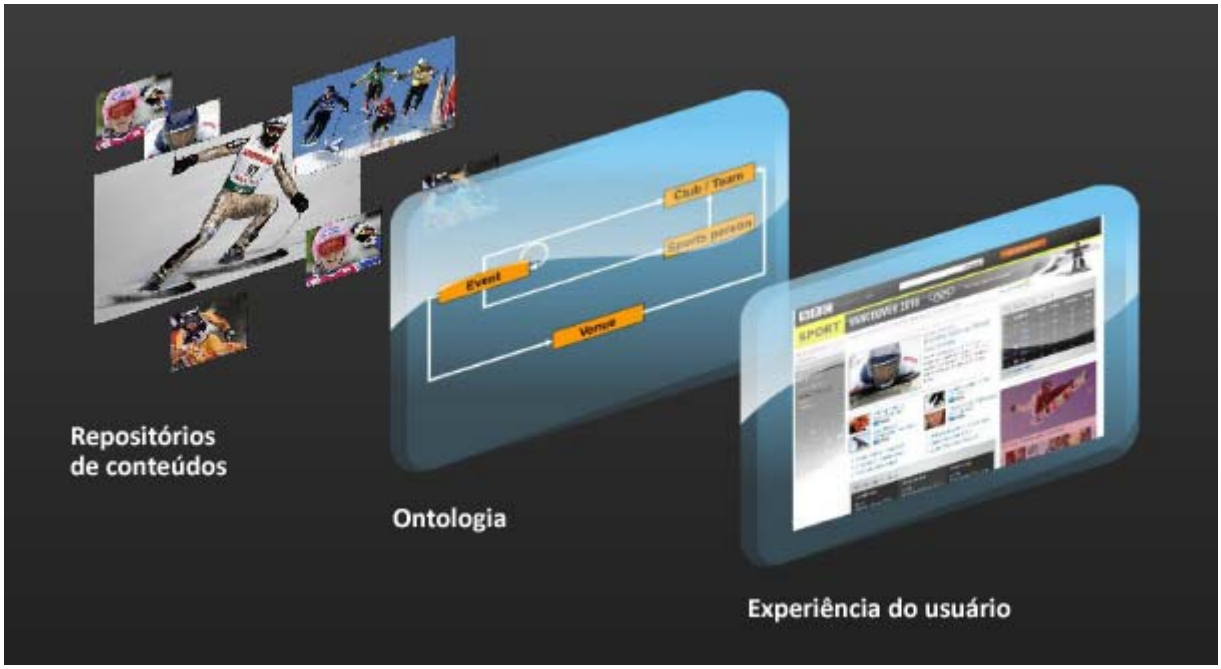


Figura 32 – Visão simplificada do processo de publicação semântica da BBC (OLIVER, 2010b, tradução nossa)

[109] *Triple store* desenvolvido pela empresa Ontotext. Segundo a empresa, a atual versão do BigOWLIN (denominado OWLIM-SE) é um repositório semântico com capacidade de carregar dezenas de bilhões de triplas. Disponível em: <http://www.ontotext.com/owlim>. Acesso em: 27 jan. 2012.

Na Figura 32, da esquerda para a direita: a primeira camada representa os repositórios de conteúdos, em formatos diversificados e oriundos de fontes internas e externas. Na camada intermediária, a ontologia do domínio "esportes", desenvolvida pela equipe da BBC, que serve como modelo para determinar os relacionamentos entre os conteúdos e, assim, definir a organização da publicação. Por último, a camada "Experiência do usuário", que nada mais é do que os documentos hipertextuais criados dinamicamente de forma automatizada. Segundo Oliver, para que a ontologia consiga determinar os relacionamentos, é necessário que os jornalistas associem *tags* consistentes aos conteúdos, que traduzam os conceitos dos mesmos.

O'Donovan resume o processo em uma frase: para ele, o ponto-chave é que "nós estamos usando alguns métodos avançados para analisar conteúdos e decidindo como rotular esse conteúdo com metadados precisos e lincados a conceitos únicos (um conceito é geralmente uma pessoa, um lugar ou uma coisa)" (O'DONOVAN, 2010, *online*). O autor também apresenta um gráfico que explica o processo de publicação dinâmica e semântica do site (Figura 33), porém de forma mais detalhada e complexa do que o gráfico da Figura 32.

Para fins de estudo, traduzimos as legendas presentes no gráfico. O processo mostrado na Figura 33 uma ordem de baixo para cima. O fluxo é formado por cinco caixas empilhadas, que representam as fases do processo. Cada uma das cinco fases está indicada com um número à direita (marcação nossa).

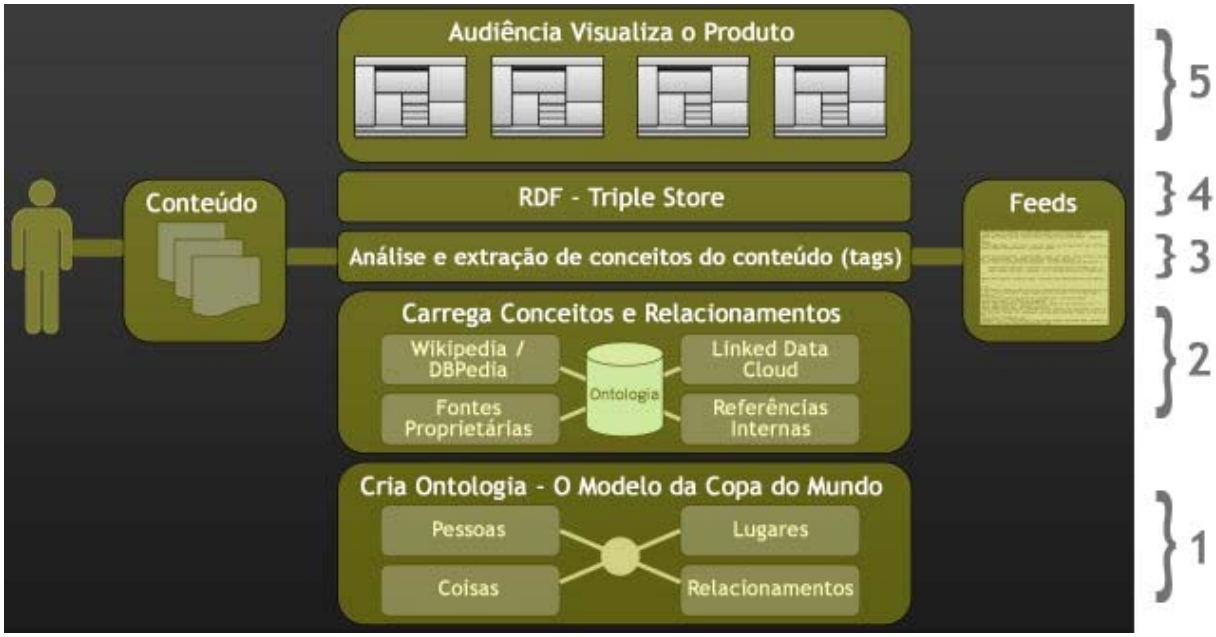


Figura 33 – Processo de publicação dinâmica e semântica da BBC (O'DONAVAN, 2010, tradução nossa, marcação nossa)

Na base da Figura 33 (indicada com o número 1), é representada a construção da ontologia, como um modelo para o domínio da Copa do Mundo. Nesta fase, são definidos os conceitos sobre as entidades que fazem parte deste domínio e como elas se relacionam entre si. Por exemplo: um jogador é uma pessoa; uma seleção nacional é um time; um jogador faz parte de um (e apenas um) time; e assim por diante.

Após a modelagem da ontologia, é necessário passar dados reais para este modelo. Na caixa acima (n.º 2), são carregados para a ontologia tais dados (conceitos e seus relacionamentos) oriundos de fontes internas e externas, tais como os dos *datasets* que fazem parte da Linked Data.

Em seguida (n.º 3), os conteúdos produzidos por jornalistas e as informações estruturadas de fontes externas (ex.: *feeds* de estatísticas sobre jogadores e seleções) são analisados e etiquetados (*tags*), a fim de se extrair conceitos destes conteúdos. Como já citamos, esta etiquetagem é realizada através de dois processos: o *tagging* manual (com auxílio de uma taxonomia pré-definida por um *software*) e a análise automática do conteúdo via *software* de reconhecimento de conceitos de textos em linguagem natural.

Na penúltima caixa (n.º 4), é representado o repositório semântico (*triple store*), que armazena os metadados gerados nas fases anteriores (inclusive a ontologia), organizados em grafos, que são utilizados na criação de inferências, necessárias para a publicação automaticamente dos conteúdos jornalísticos nas devidas páginas geradas de forma automatizadas e apresentadas aos usuários (fase n.º 5).

Finalizamos aqui a descrição do primeiro caso estudado. No próximo tópico, passamos para a análise das contribuições das tecnologias identificadas no caso em questão, baseando-nos nas categorias do JDBD propostas por Barbosa (2007, 2008a).

3.2.5 Contribuições das tecnologias semânticas ao atual paradigma do JDBD

Ao associarmos os depoimentos dos desenvolvedores da BBC com as categorias de análise elencadas por Barbosa (2007, 2008a), o sistema de publicação semântico apresentou possibilidades de potencialização em grande parte das categorias, principalmente na de automatização. A seguir, realizamos uma análise em cada categoria do JDBD baseados nos dois casos estudados.

3.2.5.1 Dinamicidade

O caso apresentou alto nível de dinamicidade, pois não há trabalho manual na manutenção de interfaces ou na inserção de conteúdos em códigos HTML. Não encontramos indícios de rupturas em relação aos atuais produtos jornalísticos em base de dados, já que, atualmente, grande parte dos sites que utilizam BDs já consolidou a lógica da separação entre conteúdo e apresentação, o que os torna altamente dinâmicos, pois a separação entre conteúdo e apresentação exige um sistema dinâmico de publicação. Entretanto, a autonomia das máquinas na associação de entidades (jogadores, times etc.) potencializa o caráter dinâmico do sistema semântico, pois a dinamicidade deixa de ser um atributo apenas das operações de publicação, e passa a ser um atributo das operações de decisão (ex.: a qual time um jogador deve ser associado?). Tal potencialização está diretamente relacionada à categoria de automatização.

3.2.5.2 Automatização

Praticamente, todo o site é organizado de forma automatizada. Segundo Barbosa (2007), existem três tipos de automatização: a parcial (parte do processo é automática), a procedimental (várias etapas do processo são automáticas) e a total (todo o processo é automático). No caso do site World Cup 2010, a técnica de *tagging* nos conteúdos é manual, mas poderia ser excluída do processo, pois ainda haveria as *tags* resultantes da extração automática de conceitos via *software*; porém, a permanência de um sistema manual de moderação de *tags* foi uma escolha da equipe, para que se mantenha a qualidade elevada do conteúdo (Rayfield, 2010).

D'Onovan afirma que a plataforma de publicação dinâmica e semântica desenvolvida pela BBC modificou o fluxo editorial ("*workflow*") da criação de conteúdos e gerenciamento do site: passa do modelo tradicional de "publicar matérias e páginas *index*" para o fluxo de "publicar conteúdos e checar se as sugestões de *tags* estão corretas", pois as publicações de matérias e páginas *index* são automatizadas, e foi esse novo fluxo editorial que permitiu a viabilidade de um projeto com mais de 800 páginas *index*.

3.2.5.3 Flexibilidade

A Web Semântica contribui na categoria da flexibilidade por diversos motivos. Um deles é a possibilidade de diferentes equipes produzirem conteúdos especializados em distintos locais de produção e, ainda assim, terem suas produções reunidas de forma automatizada em um produto, devido à associação dos metadados destes produtos ao modelo de conceitos da ontologia. Foi o que ocorreu no caso do World Cup 2010, que reuniu, de forma automática, conteúdos e *feeds* gerados por sites diferentes, tanto internos quanto externos à BBC.

Nas palavras de O'Donovan, o uso de RDF e Linked Data torna o sistema "incrivelmente flexível". Para Rayfield, "o modelo de tripla RDF também facilita a modelagem ágil, enquanto que a modelagem do esquema relacional tradicional é menos flexível e também incrementa a complexidade da *query*" (2010, *online*). Rayfield ainda afirma que a capacidade de gerar inferências torna as *queries* e o processo de *tagging* mais rápidos e simples que o modelo em SQL tradicional, além de aumentar a qualidade e a abrangência dos conteúdos no site. Para ele, além de ser mais flexível do que o tradicional SQL, o *triple store* empregado ainda permite futuras expansões na abrangência de dados relacionados, pois aceita a inclusão de novas ontologias e *datasets* da Linked Data. Ou seja: o modelo de organização dos dados não fica preso à rigidez de uma estrutura de BDs em tabelas.

3.2.5.4 Inter-relacionamento/Hiperlinkagem

A categoria de inter-relacionamento/hiperlinkagem, que é a "capacidade de identificar padrões combinatórios e inter-relacionamentos diversos entre as informações" (BARBOSA, 2007, p. 238), é reforçada pela capacidade do sistema semântico de identificar as entidades com o mínimo de ambiguidade através do uso das URI como identificadores únicos para todos os sites envolvidos no Linked Data (e para toda a web). Dessa maneira, uma página que cita o nome de um jogador, de uma seleção, de um grupo ou de uma partida poderá buscar dados sobre estes assuntos nos sites do Linked Data com menor chance de um erro de identidade no inter-relacionamento.

3.2.5.5 Densidade informativa

Na categoria da densidade informativa, a grande vantagem do sistema semântico é a convergência de conteúdos diversificados oriundos de sites externos. No caso World Cup 2010, o Linked Data contribuiu bastante para a maximização da densidade de informações. Segundo Barbosa (2007), um produto jornalístico que obtém dados de diversas fontes terá uma densidade informativa maior. No caso do site da BBC, as páginas são alimentadas com conteúdos jornalísticos, *posts* de blogs e *feeds* de diversas fontes internas da emissora, além da integração de dados e metadados oriundos de outros *datasets* disponíveis na internet e que respeitam as condições do Linked Data. Outra vantagem da Linked Data é a possibilidade de liberação dos repositórios da BBC para *queries* realizadas por sites externos, já que o SPARQL realiza apenas operações de recuperação de dados, ou seja, não realiza ações de inclusão, exclusão e *update* dos dados, como ocorre nas bases de dados relacionais com o uso de linguagens tradicionais de SQL (SEGARAN et al., 2009).

3.2.5.6 Diversidade temática

A categoria diversidade temática foi uma das que menos demonstraram vantagens, devido à natureza do site: todo ele é sobre esportes, mais especificamente sobre um evento único. Entretanto, embora Barbosa conceitue a categoria como diversidade de tematizações e ilustre essa diversidade temática com a listagem de editorias diferentes (como política, economia, cultura etc), poderíamos considerar que há uma diversidade de formatos e gêneros jornalísticos, como notícias, reportagens, artigos de opinião em blogs, mídias e estatísticas.

Em relação aos quesitos técnicos, consideramos que a diversificação de formatos é um desafio mais complexo do que a diversificação de temas. Por isso, o sistema semântico poderia manejar facilmente a integração de diversos temas em um mesmo produto jornalístico digital. Na Web Semântica, as entidades individuais são identificadas com URIs únicas, então, independentemente do tema tratado nos conteúdos, se a entidade estiver presente neles, tais conteúdos poderão ser recuperados e reunidos em uma mesma interface.

3.2.5.7 Visualização

Na categoria visualização, que para Barbosa são as diferentes maneiras de se representar na tela as informações jornalísticas armazenadas nas BDs, o sistema semântico do

site World Cup 2010 não demonstrou benefícios vantajosos em relação aos sistemas tradicionais, pois não houve o aproveitamento efetivo de recursos mais elaborados de visualização, como os infográficos interativos. Entretanto, ainda assim, pudemos observar páginas que apresentavam dados estruturados em formatos diferenciados, tais como as tabelas de resultados (Figura 28), calendários de jogos (Figura 29), e gráficos em barras com estatísticas (Figura 34) presente tanto na página dos jogadores quanto na página das seleções. Acreditamos que a vantagem apresentada no site para a categoria de visualização ocorre em uma etapa anterior à construção dos gráficos: ocorre na busca de dados em fontes internas e externas que ocorre graças às inferências, ou seja, na associação automática dos dados ao gráfico a partir de significados gerados pela máquina.

Figura 34 – Dados sobre jogador convertidos para o formato de gráficos em barra

3.2.5.8 Convergência

A categoria da convergência se beneficia pelo fato das mídias (áudios e vídeos) serem etiquetadas (*tagged*) com metadados ricos em semântica, o que possibilita maiores chances de reaproveitamento das mídias em diversas matérias jornalísticas (e outros espaços do site). Outra vantagem para esta categoria é o fato das páginas permitirem o reaproveitamento automático de informações presentes em outros sites da web, como no caso dos *feeds*. Se considerarmos que a convergência é mais do que a união de mídias em um mesmo espaço; que é, também, a ideia de convergir conteúdos de origens diversas em um mesmo local

agregador, então acreditamos que a categoria da convergência é uma das mais beneficiadas pelas contribuições da Web Semântica, pois está diretamente relacionada à ideia de interoperabilidade entre sites e serviços diferentes. A característica da interoperabilidade possibilita o compartilhamento de conteúdos, que por sua vez pavimenta o caminho necessário para o reaproveitamento de produções informacionais, midiáticas e intelectuais.

Finalizamos aqui a análise do caso BBC World Cup 2010. No próximo tópico, começamos o estudo do segundo caso selecionado para a pesquisa. No final deste capítulo, realizamos uma análise geral sobre as contribuições da Web Semântica identificadas em ambos os casos.

3.3 Caso BBC Wildlife

O BBC Wildlife é um portal que reúne uma grande produção de conteúdos sobre o mundo natural, mais especificamente biológico, como animais selvagens, plantas, fungos e, inclusive, seres pré-históricos. O site armazena e organiza seu conteúdo (textos, imagens, áudios e vídeos) sobre a natureza como se fosse uma enciclopédia multimídia, e utiliza esta base de conhecimento em matérias jornalísticas sobre o tema. Ao contrário do caso anterior, neste caso as tecnologias semânticas foram aplicadas para conteúdos mais "leves", conhecidos como *soft news* ou *feature*, e se aproximam de produtos como reportagens de revista, documentários e produtos informativos para educação e entretenimento.

No APÊNDICE C, apresentamos os profissionais que serviram como fontes de dados secundários para a identificação e descrição das tecnologias semânticas. Além dos profissionais, também encontramos informações importantes na página da ontologia desenvolvida para o site, na página de FAQ (questões frequentemente questionadas) e na página *Feeds and Data*, em que há indicações sobre algumas das tecnologias semânticas empregadas.

3.3.1 Descrição do produto

O BBC Wildlife reúne uma grande quantidade de conteúdos multimídia produzidos continuamente por diferentes programas da BBC (TV, rádio e digital) que contenham como temática a vida natural, mais especificamente os seres biológicos, como animais, plantas, e inclusive dinossauros. Só na parte de vídeo, são mais de 3000 clipes (de curta duração), oriundos de dezenas de programas em mais de 30 anos de produção televisiva da BBC. Na Figura 35, é mostrada a página inicial do BBC Wildlife.

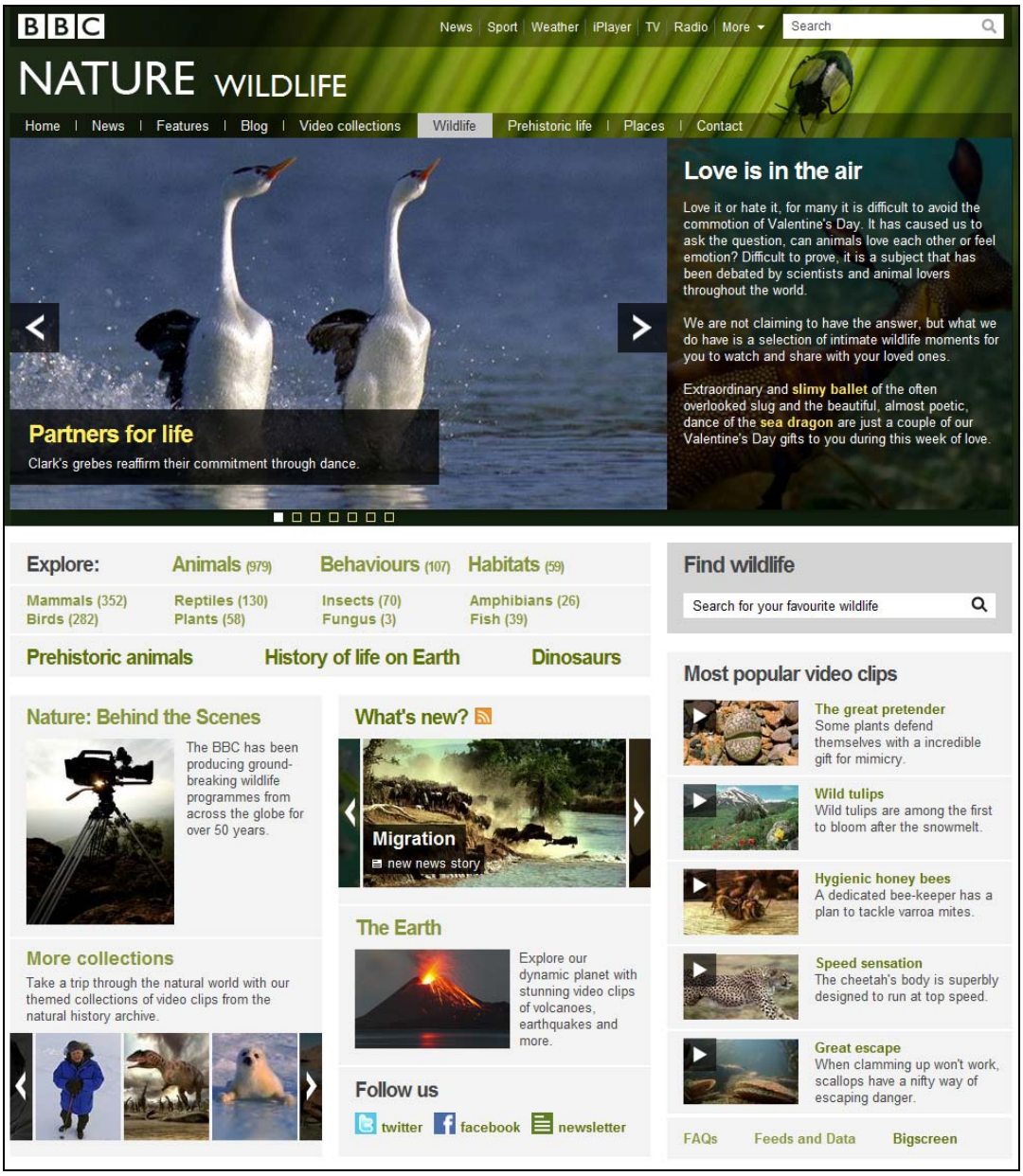


Figura 35 – Página inicial do BBC Wildlife[110]

Além do rico repositório de mídias, outro grande destaque do site é a organização das páginas. Para cada uma das mais de mil espécies, há uma página única gerada e atualizada de forma dinâmica, que agrega informações, áudios e vídeos sobre a espécie em questão. Além das páginas para as espécies, o site gera outras centenas de páginas para reunir conteúdos sobre animais que compartilham das mesmas características. Essas características são: habitat (ex.: floresta, deserto, marinho, urbano etc), comportamento/adaptação (ex.: se é carnívoro, se voa, se é noturno etc) e nível da classificação biológica (domínio, reino, filo, superclasse, classe, superordem, ordem, subordem, superfamília, família, gênero e espécie). Como

[110] Disponível em: <http://www.bbc.co.uk/nature/wildlife/>. Acesso em: 21 dez. 2011.

exemplo, podemos citar uma página que agrega vídeos e informações apenas sobre animais da classe dos insetos, outra apenas sobre animais e plantas com comportamento carnívoro, outra só com animais voadores, ou ainda uma página somente sobre animais e plantas que vivem no habitat urbano. As páginas apresentam diversos links umas para as outras, de forma dinâmica, de acordo com os tipos de relacionamentos entre os conceitos. Por exemplo: a página de uma determinada espécie mostra uma lista de características que esta espécie apresenta (ex.: é voador, é carnívoro etc.), e cada característica listada é um link que abre uma página sobre animais que também apresentam tal característica. Isso também ocorre com outras possíveis relações, como links de habitats na página das espécie, links de filos na página dos reinos etc.

A **navegação** do site não funciona com um menu central permanente, como ocorre em sites tradicionais. Há somente um menu na página inicial (Figura 36), que não é mostrado nas outras páginas, pois serve apenas como ponto de partida para a navegação pelos links das páginas internas.

Figura 36 – Menu na página inicial do site Wildlife. Marcações nossas

O menu aparece somente na página inicial porque é apenas uma porta de entrada para uma série de páginas internas que, por sua vez, apresentam diferentes listas de links, que servem como menus contextualizados para o conteúdo que mostram. Ainda na Figura 36, é possível observar um número entre parênteses próximo a cada uma das opções do menu, que representam a quantidade de links encontrados na página interna a que o referido link remete

(ex.: ao clicar em *Animals*, é aberta uma página que lista 979 espécies de animais, e cada item listado é um link para a página da respectiva espécie).

Além destas três categorias de navegação oferecidas no menu (por espécies, por comportamentos/adaptações e por habitats), as páginas internas do site ainda oferecem duas outras categorizações que possibilitam outra forma de navegação: por biorregião (oito grandes regiões do mundo, ou seja, regiões mais generalizadas do que os 59 habitats); ou pelos outros níveis da classificação dos seres vivos além das espécies (domínios, reinos, filos, superclasses, classes, superordens, ordens, subordens, superfamílias, famílias e gêneros), que geram centenas de páginas agregadoras de conteúdo.

O site Wildlife constrói uma **página dinâmica** para cada habitat, cada comportamento/adaptação, cada biorregião e cada nível da classificação biológica (espécies, domínios, filos etc) (ver Figura 37). As páginas são construídas a partir de um *template* padrão que divide o espaço da tela em duas grandes áreas: na parte superior (parte mais escura da Figura 37), são disponibilizados os conteúdos relacionados ao assunto em questão (título, texto descritivo, links e mídias) oriundos de diferentes fontes internas e externas à BBC. Já na parte inferior (parte mais clara da Figura 37), são apresentadas diversas listas de links, como se fossem novos menus para conteúdos relacionados ao tema da página.

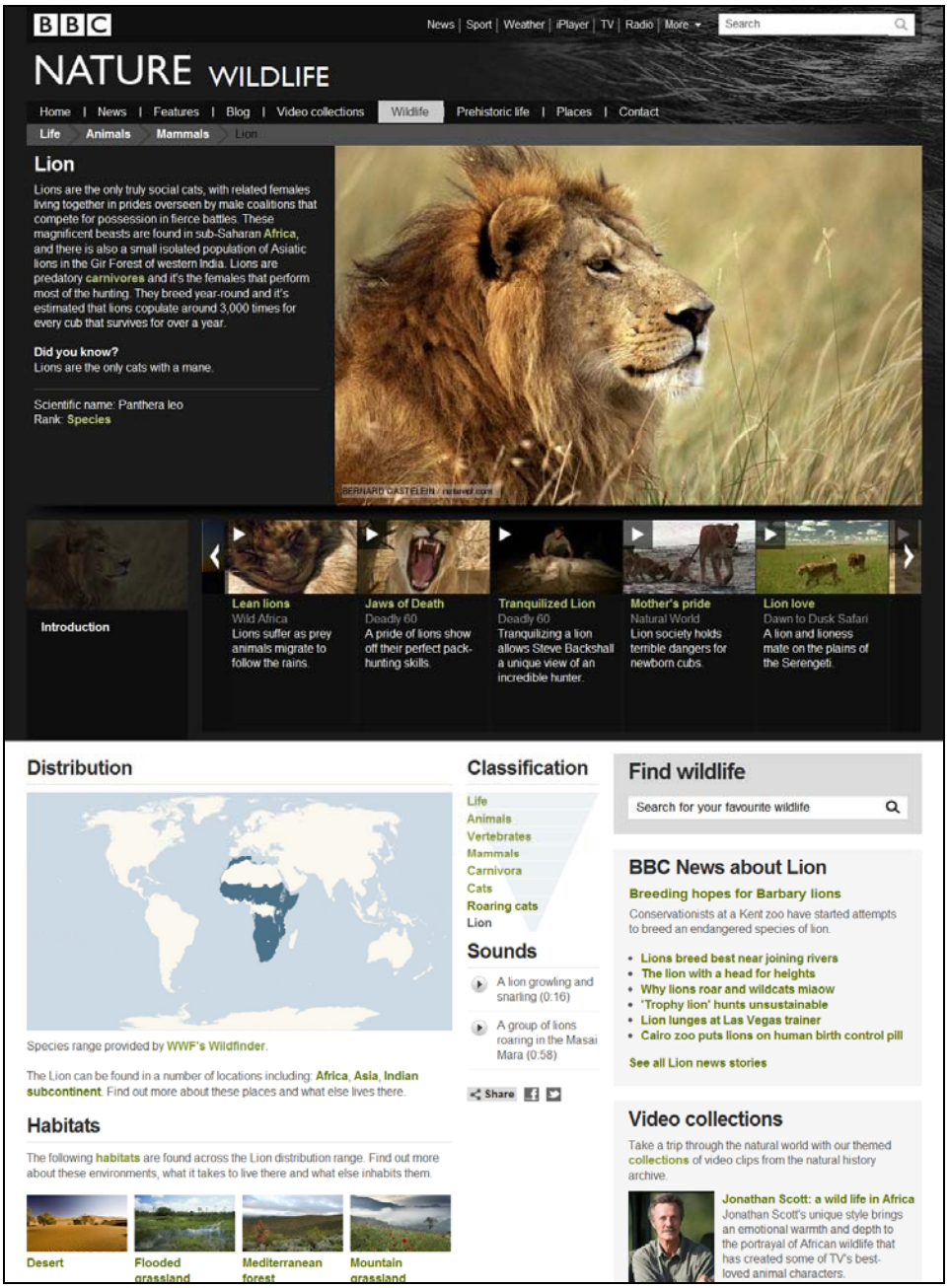


Figura 37 – Visão parcial da página das espécies[111]

A estrutura da parte superior da página dinâmica segue sempre o mesmo padrão: título, texto e galeria de mídias. Já na parte inferior, os elementos da página mudam de acordo com o tipo de conteúdo tratado (se é sobre uma espécie, mostra habitats e comportamentos da espécie; se é sobre um habitat, mostra as espécies que fazem parte do habitat; e assim por diante). Para uma descrição mais detalhada, apresentamos novamente a página das espécies na Figura 38, porém com indicações dos elementos que formam o *layout*.

[111] Disponível em: <http://www.bbc.co.uk/nature/life/Lion/>. Acesso em: 22 dez. 2011.

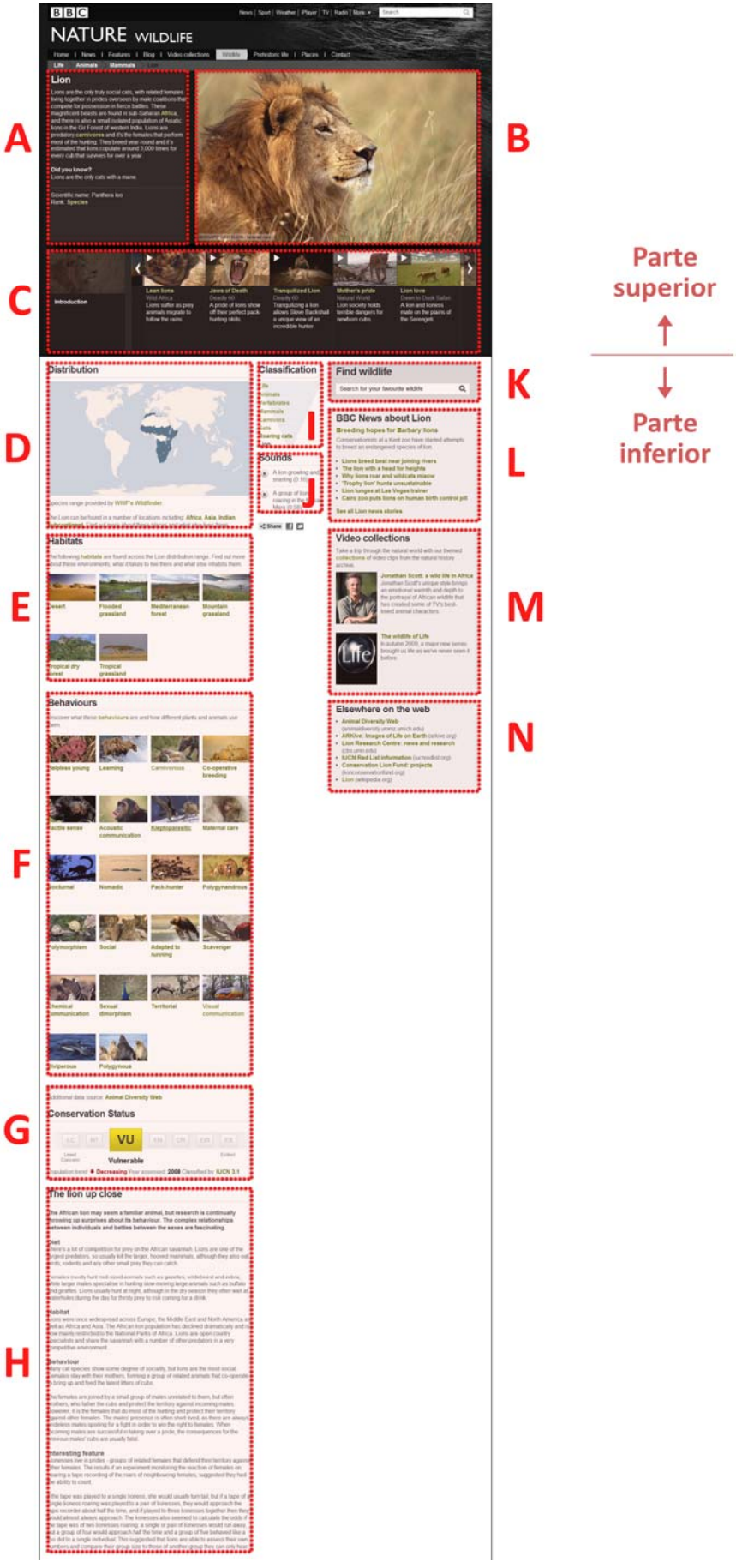


Figura 38 – Página das espécies, com marcações indicativas

Na parte superior da Figura 38, são apresentados os seguintes conteúdos:

A) O título e a descrição do conteúdo mostrado no momento. No caso do leão, como mostrado na figura, o texto da descrição apresenta links para algumas características da espécie, como uma região onde o animal é encontrado (África) e um comportamento (carnívoro), além do nível da classificação biológica a qual o leão se refere (espécie). Cada um destes links remete o usuário a uma página com estrutura semelhante, porém com conteúdos relacionados ao respectivo link.
B) Espaço para a reprodução das mídias (clipes de vídeo produzidos pelas dezenas de programas de televisão da BBC).
C) Lista horizontal dos vídeos que a página agrega. Ao clicar em uma das miniaturas, o vídeo é aberto no espaço demarcado com a letra B, e é mostrada a descrição do vídeo no espaço A (título do clipe, texto descritivo e nome do programa de origem da BBC em que o vídeo foi produzido, lincado para o site do respectivo programa). A lista de miniaturas pode ser deslizada para o lado, para se revelar mais miniaturas (a página da espécie leão continha 32 miniaturas no momento da pesquisa).

Na parte inferior da Figura 38, são apresentados os seguintes conteúdos:

D) Mapa com destaque às biorregiões onde a espécie é encontrada.
E) Lista de habitats em que a espécie em questão pode ser encontrada. Este espaço serve como um menu para outros conteúdos do site. Cada imagem é um link que remete o usuário para uma página semelhante a esta, porém que agrega animais encontrados no habitat em questão.
F) Lista de comportamentos (e adaptações) que a espécie demonstra, tais como “nômade”, “noturno”, “social” ou “territorial”. Mais uma vez, o espaço serve como um menu para o conteúdo do site, neste caso para páginas que agregam vídeos de animais com estes mesmos comportamentos.
G) Status da conservação da espécie (ex.: vulnerável, ameaçado de extinção, extinto).
H) Texto linear com informações mais detalhadas sobre o assunto em questão.
I) Nível da classificação biológica em que o assunto da página se localiza. Como no caso o leão é uma espécie, então também são mostrados os outros níveis anteriores, como gênero, família etc. A lista da classificação se torna um menu para conteúdos do site.
J) Mídias sonoras relacionadas ao tema, oriundos de programas da BBC. No caso da Figura 38, são disponibilizadas gravações de rugidos de leões.

K) Campo para pesquisa no site.
L) Lista de notícias da BBC sobre o tema da página. Os links podem remeter o usuário para diferentes páginas da BBC.
M) Lista com coleções especiais de vídeos em que o assunto em questão é mencionado.
N) Links para páginas externas à BBC que tratem sobre o assunto em questão.

A espécie é o nível mínimo na organização do conteúdo. Todas as outras páginas agregam conteúdos de várias espécies. Por isso, ao invés de mostrarem links para habitats e comportamentos relacionados a uma espécie (como ocorre na página do leão), as páginas dos outros níveis da classificação biológica apresentam links para grupos de animais que fazem parte do referido nível. Ou seja: a parte inferior da página mostra links para a exploração dos níveis que derivam daquele grupo. Na Figura 39, há uma comparação entre três páginas de níveis biológicos diferentes: na parte inferior da página da espécie leão (primeira tela), as pequenas imagens são links para habitats e comportamentos dos leões (indicação A). Nas duas outras telas (classe mamíferos e filo vertebrados), as pequenas imagens são para grupos de animais que fazem parte dos referidos níveis (indicações B e C). Assim, ao se começar a exploração do site pela página do primeiro nível da classificação (reino), a navegação pode levar o usuário de um nível a outro, até que chegue à página de qualquer espécie tratada pelo site.

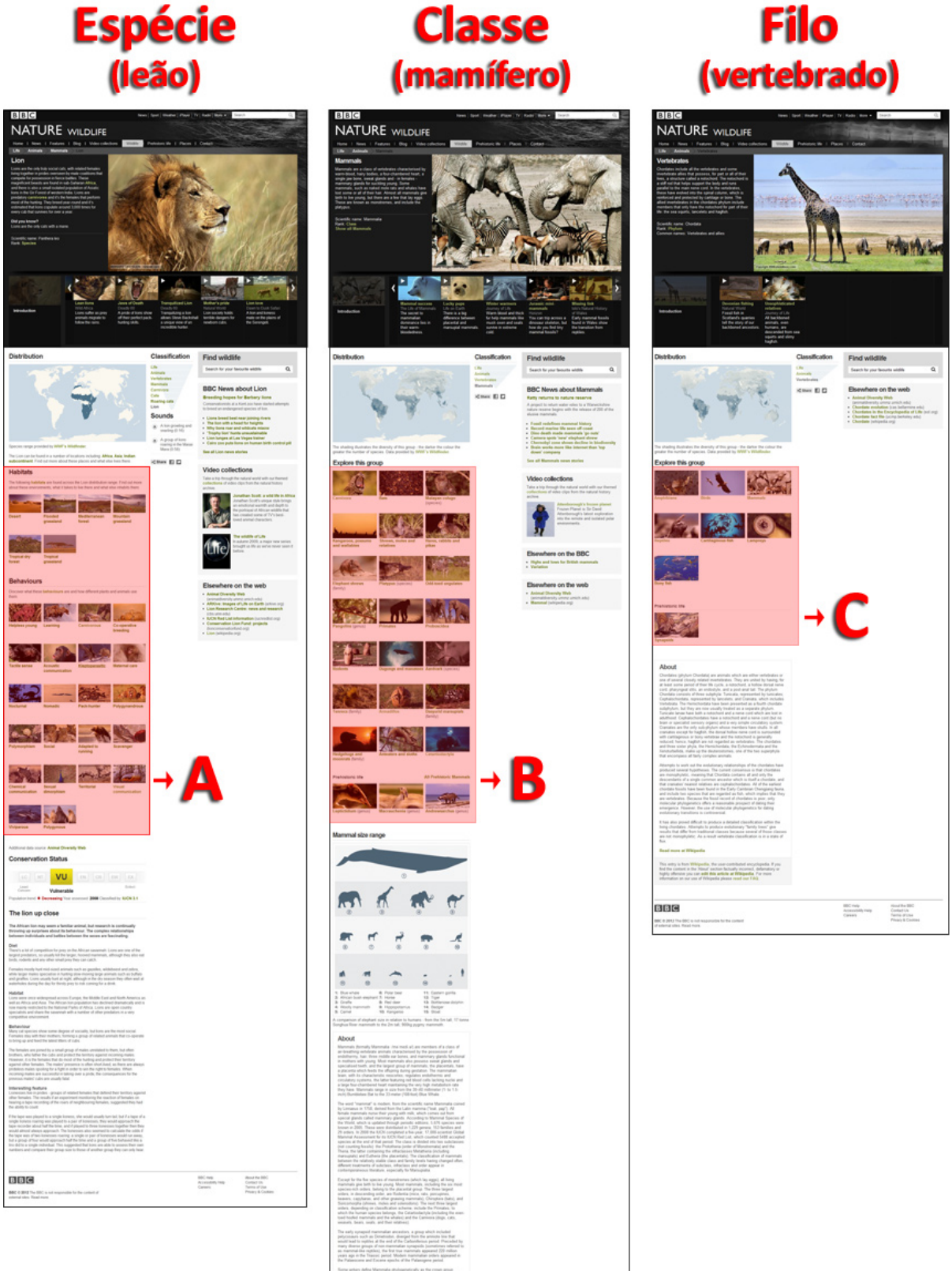


Figura 39 – Comparação entre as páginas de espécie (leão), classe (mamíferos) e filo (vertebrados)

Cada página monta a sua estrutura de navegação automaticamente, de acordo com os tipos de relacionamentos que possuem com os conteúdos. Por exemplo: assim como a página

da espécie lista links para comportamentos e para habitats, as páginas dos comportamentos e dos habitats listam todas as espécies que deles fazem parte, como ocorre na Figura 40. Dessa forma, o sistema cria automaticamente uma malha de páginas interlincadas, rica em relacionamentos entre conceitos. Na Figura 40, é possível perceber a capacidade de agregação das páginas. Por exemplo, na página de comportamento/adaptação, há uma grande lista de espécies que possuem tal característica, e todas estão organizadas de acordo com a classe que pertencem. Ainda, cada página divide as coleções entre plantas e animais.

As diferentes maneiras de se categorizar os conteúdos permitem que o site formule e distribua pelas páginas internas várias listas de links, que convidam o usuário a continuar a navegação pelo site de acordo com o contexto, como se as próprias páginas internas fosse um grande menu de navegação.

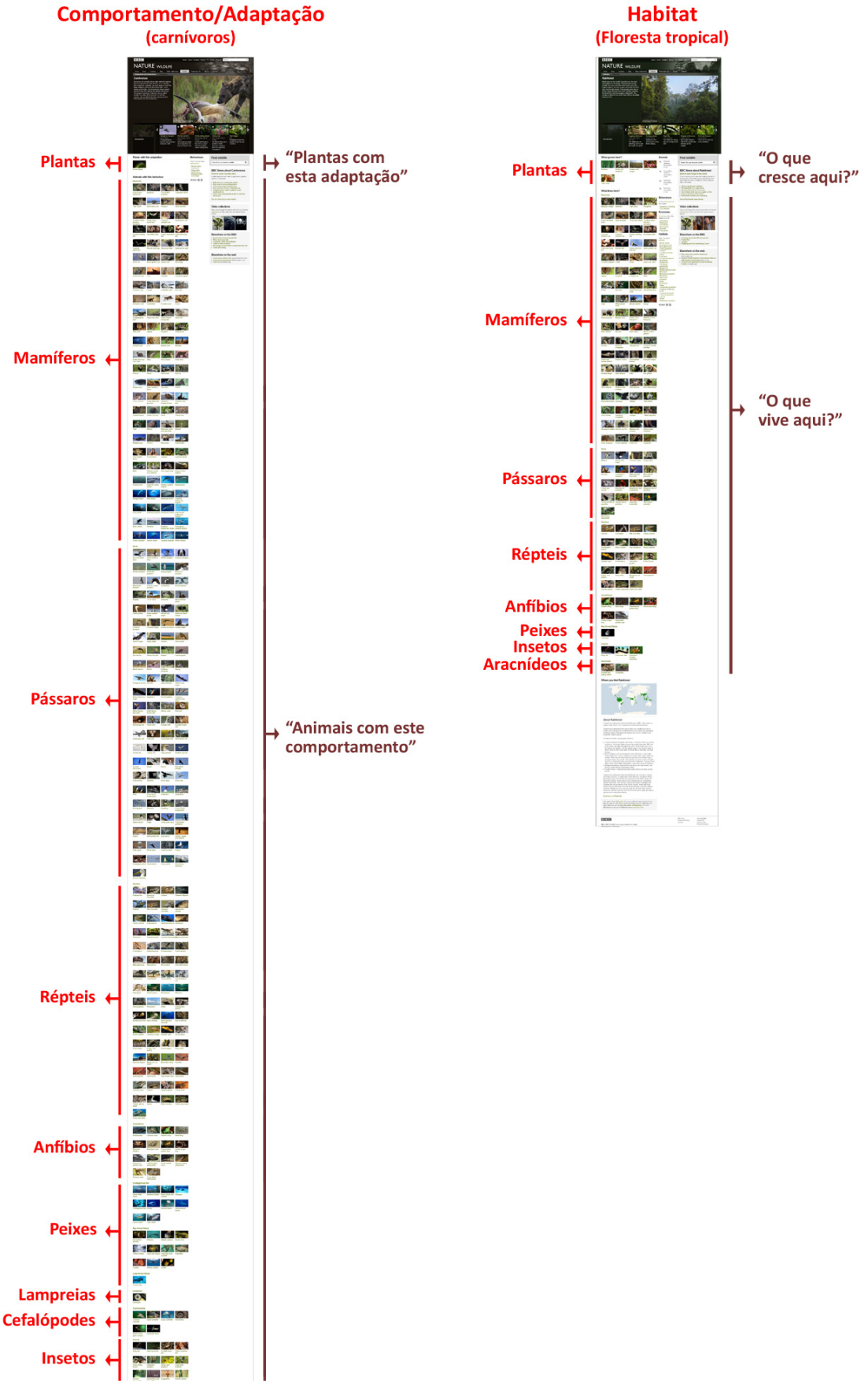


Figura 40 – Página de comportamento/adaptação (esquerda) e da página de habitat (direita)

As páginas apresentam informações em diferentes formatos, como textos, clipes de áudio, estatísticas e mapas. Destes, o tipo de conteúdo que se destaca são os clipes de vídeos, que, no Wildlife, são trechos curtos obtidos de documentários televisivos (aproximadamente 3 minutos), mas que apresentam um sentido completo. Embora sejam construções de narrativas que buscam relatar o real, tais vídeos não apresentam a urgência do relato sobre um acontecimento recente, característica que faz parte da definição de notícia. Por essa razão, acreditamos que o conteúdo do site não seja classificado como produção jornalística, pois até mesmo no jornalismo especializado, como no ambiental e no científico, há uma busca pela publicação de fatos relacionados a acontecimentos recentes. Porém, percebemos que a BBC aproveita essa base de conhecimento para enriquecer seus próprios conteúdos jornalísticos, tanto nos seus sites especializados sobre as questões ambientais e científicas quanto no seu site principal de notícias. Em outras palavras, o enorme conjunto de páginas e de suas interligações formuladas de acordo com os tipos de relacionamentos entre conceitos tornam o Wildlife uma base de conhecimento sobre um domínio específico do mundo (a vida natural) que servem de **complemento aos produtos jornalísticos** da BBC.

A partir desse repositório, a BBC cria conexões entre o conhecimento ali organizado e os textos jornalísticos dos outros sites da emissora. Essa conexão pode ser apresentada pelos dois lados: se há links para notícias nas páginas do Wildlife, pode haver links do Wildlife nas respectivas páginas de notícias. Como exemplo, tomamos novamente a página da espécie leão, que mostra uma caixa de notícias da própria BBC que mencionam a espécie (Figura 41).

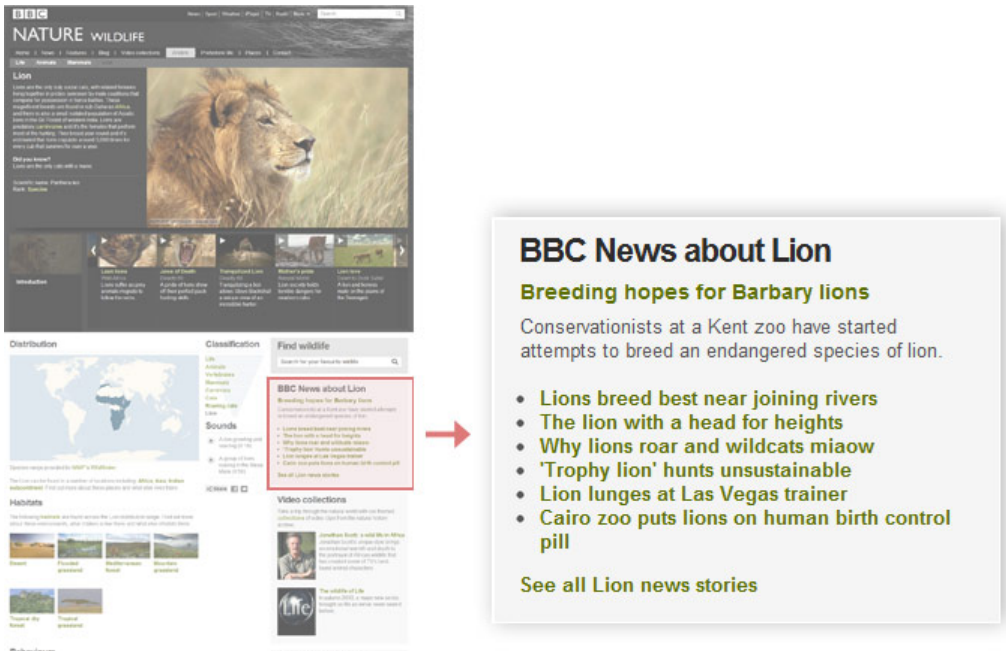


Figura 41 – Caixa de links para notícias relacionadas ao conceito de "leão"

Na Figura 41, são listadas as seis últimas notícias indexadas. A seguir, citamos a origem de cada uma delas:

- *Breeding hopes for Barbary lions at Port Lympne*: BBC News Kent[112]
- *Lions breed best near joining rivers*: BBC News Science & Environment[113]
- *The lion with a head for Heights*: BBC Wiltshire[114]
- *Why lions roar and wildcats miaow*: BBC Earth News[115]
- *'Trophy lion' hunts unsustainable*: BBC Earth News[116]
- *Lion lunges at Las Vegas trainer:* BBC News US & Canada[117]
- *Cairo zoo puts lions on human birth control pill:* BBC News Middle East[118]

A Figura 42 mostra a página de uma das notícias listadas acima, do site BBC Earth News, especializado em jornalismo ambiental. A notícia, sobre a crescente ameaça da caça indiscriminada de leões e leopardos, apresenta uma caixa com links da BBC relacionados ao tema. Nesta caixa, há dois links para o Wildlife: na marcação A, um link para a página da espécie leão e, na marcação B, para a página da espécie leopardo. Para cada link, o título apresenta o nome da espécie e a descrição "vídeos, arquivos de áudio, fatos, fotos e matérias", ou seja, indica que o Wildlife é um complemento que contextualiza as informações da notícia.

[112] http://www.bbc.co.uk/news/uk-england-kent-15862433
[113] http://www.bbc.co.uk/news/science-environment-12806519
[114] http://news.bbc.co.uk/local/wiltshire/hi/people_and_places/nature/newsid_9135000/9135050.stm
[115] http://news.bbc.co.uk/earth/hi/earth_news/newsid_9028000/9028491.stm
[116] http://news.bbc.co.uk/earth/hi/earth_news/newsid_8993000/8993557.stm
[117] http://www.bbc.co.uk/news/world-us-canada-11236560
[118] http://www.bbc.co.uk/news/world-middle-east-11099756

Figura 42 – Página de notícia no site BBC Earth News[119]

Embora os clipes de vídeos armazenados no Wildlife não sejam relatos de acontecimentos recentes, ainda assim eles contam histórias reais sobre o mundo natural, e muitas vezes são histórias que apontam para questões atuais, como no caso de um vídeo na página sobre leões que trata sobre americanos que pagam milhares de dólares para viajarem à África com o objetivo de caçar os felinos. Embora o conteúdo que alimente o site seja proveniente de mais de 30 anos de produção de documentários da BBC, muitas das informações são referentes a questões atuais, e isso pode servir para pautar reportagens e artigos opinativos. Ao perceber este potencial e também o potencial de contextualização da notícia, que já estava sendo explorado por diversos sites da BBC, a emissora decidiu, em março de 2011, mesclar o site Wildlife com um site de informações jornalísticas especializado na temática natureza, chamado BBC Nature (SCOTT, 2011).

O site BBC Nature possui oito seções temáticas Home, News, Features, Blog, Video Collections, Wildlife, Prehistoric life e Places. Cada uma das oito seções possui uma página inicial, com chamadas para seus respectivos conteúdos. O acesso a cada seção ocorre nos links do menu principal (Figura 43).

[119] Disponível em: <http://news.bbc.co.uk/earth/hi/earth_news/newsid_8993000/8993557.stm>. Acesso em: 17 jan. 2012.

Figura 43 – Menu principal do site BBC Nature, com links para as seções do site[120]

Embora cada link apresente uma página inicial com uma configuração visual própria, estas seções compartilham basicamente da mesma base de conteúdos: os conteúdos editoriais do Nature (notícias e reportagens) e as páginas do Wildlife. A diferença entre elas é a prioridade em mostrar determinados tipos de conteúdos. A seguir, detalhamos as particularidades de cada seção.

- As três primeiras seções (**Home**, **News** e **Features**) são páginas estruturalmente bastante semelhantes (ver ANEXO B, ANEXO C e ANEXO D), em que são apresentadas chamadas para os conteúdos jornalísticos e galerias de mídias. A diferença entre as três seções é a prioridade dada no destaque aos links: na Home, os links são um apanhado geral da produção de todo o site; no News, as chamadas dão preferência para matérias mais factuais; e em Features, as chamadas dão preferência a reportagens ou conteúdos mais elaborados tecnicamente, como as galerias de mídias.
- O **Blog** (ver **ANEXO E**), intitulado *Wonder Monkey*, é uma área para publicação de postagens do editor do site, Matt Walker, que busca inserir em seus artigos links para as páginas do Wildlife.
- A seção **Video collections** (ver **ANEXO F**) lista coleções de vídeos com um tema específico. Por exemplo: no dia dos pais, o site pode criar uma lista com vídeos sobre espécies em que o macho é o responsável pela proteção das crias. Enquanto as seções Home, News e Feature geralmente apresentam uma narrativa jornalística, no Video Collection o produto tem um com formato diferenciado (semelhante às páginas do Wildlife). De acordo com Raimond et al. (2010a), embora a seção utilize layout semelhantes às páginas do Wildlife, ela não apresenta a mesma automação, ou seja, as coleções de vídeo são planejadas e estruturadas manualmente. Os autores afirmam que essa situação é proposital, pois dessa forma a coleção se distingue como um produto editorial, e isso evita que o site apresente uma caracterização enciclopédica.

[120] Disponível em: <http://www.bbc.co.uk/nature/>. Acesso em: 17 jan. 2012.

- O link **Wildlife** (ver **ANEXO G**) leva para uma página inicial com chamadas para as páginas das espécies, dos habitats, dos comportamentos, dos filos, etc. Assim como as seções Home, News e Features são *hubs* para o conteúdo jornalístico, o Wildlife é o principal *hub* para as páginas de animais, plantas e outros seres.
- A seção **Prehistoric life** (ver **ANEXO H**) é a versão da página inicial do Wildlife, porém para a natureza pré-histórica.
- O link **Places** (ver **ANEXO I**) também mostra links para as páginas do Wildlife, porém os apresenta sobre um mapa do planeta terra.

Em outras palavras, de forma resumida: as quatro primeiras seções (Home, News, Features e Blog) servem principalmente como entrada para conteúdos jornalísticos; as seções Wildlife, Prehistoric e Places servem como entrada, principalmente, para páginas do Wildlife; e a seção Video Collection reúne os vídeos do Wildlife em coleções montadas manualmente. Estes conteúdos de interconectam entre si através de links. A página inicial de cada uma das seções apresenta uma interface própria, com uma organização particular dos links e de outros elementos do *layout*. É como se cada seção fosse um site independente, mas que compartilha a mesma base de conteúdos.

Por fim, um último exemplo sobre o aproveitamento da base de conhecimento do Wildlife no próprio BBC Nature é através dos artigos escritos pelo editor do site, Matt Walker, publicados no blog disponível como uma das oito seções do Nature. Na Figura 44, observamos uma reprodução parcial de um artigo escrito por Walker, em que aparecem vários links dentro do texto, inseridos manualmente pelo próprio editor. Cada um dos links na tela representa uma espécie do Wildlife. No artigo em questão, foi possível contar 11 links para páginas da própria BBC Nature (para páginas de espécies do Wildlife ou para coleções temáticas de vídeos da seção Video Collections), além de dois links para outros sites da BBC e um link para um site externo.

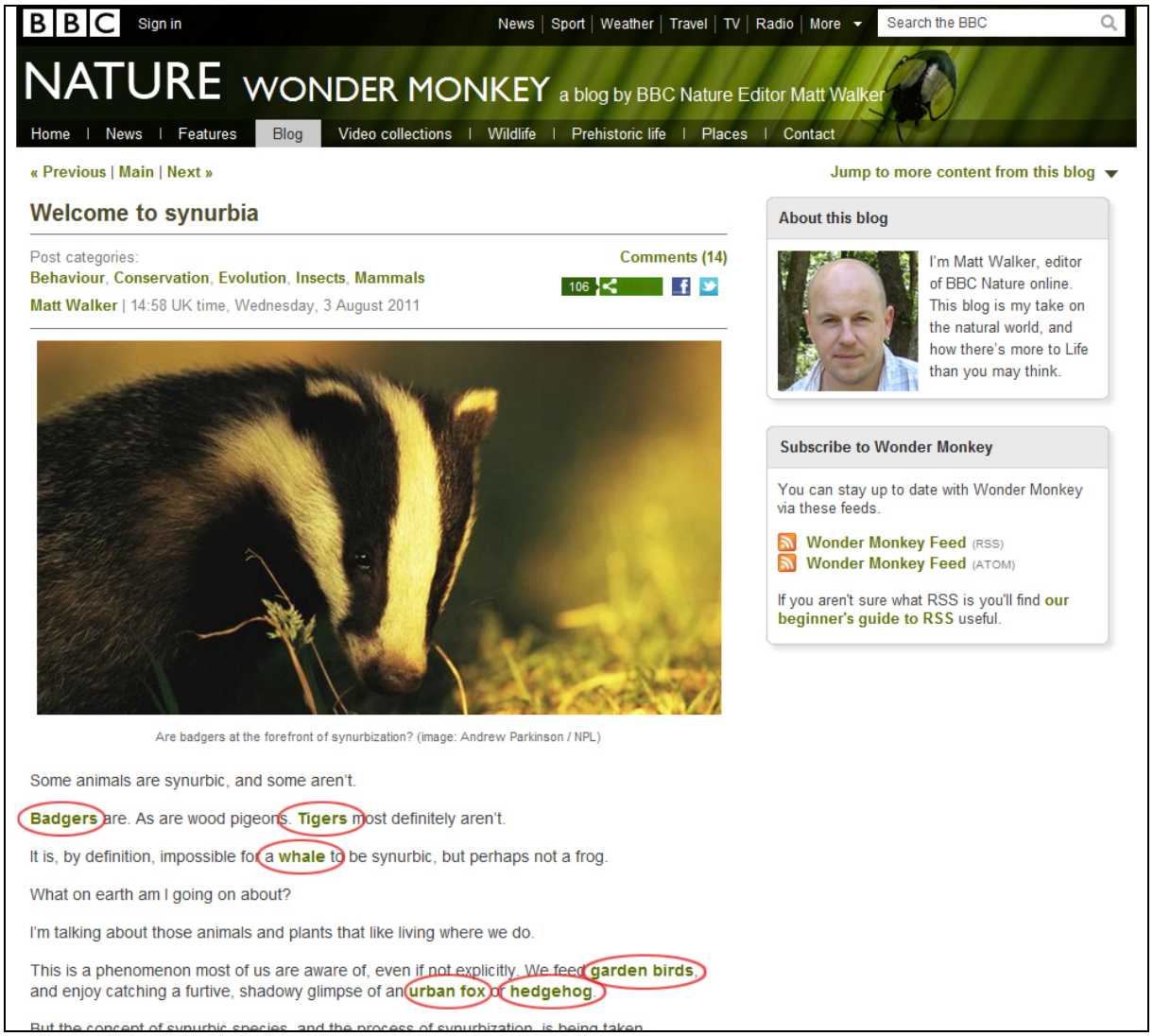


Figura 44 – Reprodução parcial de artigo em blog do site BBC Nature. Marcações nossas[121]

O BBC Nature é um site com uma grande e complexa estrutura, por isso apresenta outros detalhes que não abrangemos nesta descrição, pois não contribuem para a compreensão sobre o funcionamento do sistema semântico na organização do conteúdo do site. São detalhes como galerias de fotos que mudam de lugar de acordo com a página, links para compartilhamento das páginas em redes sociais ou rankings de notícias e clipes mais populares.

Ao analisarmos o BBC Nature sobre uma outra perspectiva, podemos considera-lo como um portal que disponibiliza notícias sobre a temática natureza e, ao mesmo tempo, agrega diferentes produtos da BBC sobre esta temática, entre eles o Wildlife. Podemos fazer uma analogia ao site BBC Sports, que publica notícias sobre esportes e, também, agrega os sites da Copa do Mundo e das Olimpíadas. Para a nossa análise, consideramos apenas o BBC

[121] Disponível em: <http://www.bbc.co.uk/blogs/wondermonkey/>. Acesso em: 18 jan. 2012.

Wildlife, que é de fato o espaço com sistema automatizado para a publicação dinâmica e semântica dos conteúdos multimídia da BBC.

3.3.2 Contexto e justificativa para uso das tecnologias semânticas

Em relação ao Wildlife, a equipe encontrou um grande desafio na proposta de desenvolver um site que distribuísse milhares de clipes de vídeos entre mais de mil páginas possíveis de serem criadas, em um sistema de publicação automatizado que não deveria apresentar ambiguidades.

Além disso, os desenvolvedores da BBC procuraram conceber uma navegação imersiva no conteúdo, sem o ordenamento de menus centralizados, de maneira que o usuário pudesse trilhar seus caminhos de acordo com seus interesses, como se fosse uma jornada sem mapa pelo mundo natural. Segundo um dos desenvolvedores,

> No passado, você sentaria em frente à TV e assistiria um documentário de uma hora sobre a vida selvagem.
> Isso não funciona muito bem na web – pessoas geralmente criam suas jornadas e assistem a clipes de vídeo com menor duração.
> Mas no site Nature, nós estamos permitindo que os usuários criem seus próprios documentários – eles podem começar [o acesso] em um animal, assistir a um clipe, seguir um link para outro animal, ler sobre aquele animal e por aí vai... [122] (SINCLAIR, 2009, online)

O desafio de se criar uma navegação sem ambiguidade por uma rede com centenas de espécies e outras centenas de páginas agregadoras demandava um sistema sólido de identificadores únicos. Outra demanda era o desenvolvimento de um modelo de relacionamentos flexível, pois as divisões da taxonomia biológica se expandem para uma grande quantidade de terminações, e cada uma delas poderia se relacionar não apenas com os níveis anteriores da taxonomia, mas também com determinadas características (comportamento/adaptação, habitat, biorregião) em comum com outras espécies. Em um momento futuro do projeto, poderia surgir a necessidade de se criar o conceito de uma nova característica. O projeto ainda exigia o reaproveitamento automático de conteúdos existentes na web, pois não seria prático para a BBC a produção de tantas descrições e dados sobre tantas espécies.

---

[122] *In the past, you'd sit down in front of the TV and watch an hour long wild life documentary.*
This doesn't work so well on the web - people are used to making their own journeys, and watching smaller length clips.
But on the /nature site, we're letting users make their own documentary - they can start on an animal, watch a clip, follow a link to another related animal, read about that animal an so on..

Diante dos desafios, a equipe da BBC identificou que a melhor solução seria evitar a abordagem tradicional de um site como um conjunto de documentos, e pensá-lo como uma rede de unidades conceituais do mundo real e de suas relações. As páginas seriam apenas uma decorrência destas relações, ou seja, seriam espaços criados dinamicamente para apresentar os resultados das associações entre conceitos do mundo natural. Para isso, adotaram como melhor solução para o Wildlife a ideia da Web Semântica.

3.3.3 Identificação de recursos e tecnologias semânticas utilizadas

Segundo os dados coletados a partir dos depoimentos dos desenvolvedores da BBC e de outros documentos, as principais tecnologias semânticas utilizadas no site foram as seguintes:

- Triplas em RDF, para relacionar recursos a objetos.
- Uma versão serializada das páginas das espécies (RDF/XML), para permitir a interoperabilidade de seus próprios dados com projetos de terceiros.
- Ontologia própria em RDF, para modelar os relacionamentos entre conceitos do domínio natural.
- URIs baseados nos identificadores da DBpedia (ou seja, da Wikipedia).
- Coleta de informações da Linked Data (reaproveitamento de conteúdo da Wikipedia através do projeto DBpedia).
- Processo de *tagging* com uso de vocabulário controlado (DBpedia).

3.3.4 Descrição do funcionamento das tecnologias semânticas

Antes do desenvolvimento de uma solução semântica para o site Wildlife, primeiro foi necessário assegurar que houvesse um ambiente com as condições necessárias para que o sistema funcionasse. Uma destas condições era a de um sistema sólido de identificadores para os conteúdos da BBC (URIs) para permitir a troca de dados entre sites da BBC. A falta de um sistema sólido de identificação de recursos era um empecilho para a interoperabilidade entre sites da BBC. Segundo Raimond et al (2010a), a falta de integração de dados entre os sites da BBC limitou algumas operações, como a de extrair dados de um contexto e apresenta-los de maneira diferentes em outro local.

Raimond et al. (2010a) afirmam que haveria a possibilidade de integrar conteúdos entre sites diferentes através de *feed* RSS. O problema desta solução é que as listas RSS não permitem segmentar os dados de acordo com o contexto. Por exemplo: como fazer com que

um *feed* de notícias sobre várias espécies mostre apenas informações sobre os elefantes em determinado contexto? Outra limitação do RSS é a impossibilidade de se realizar pesquisas (*queries*) nos *feeds*.

Os problemas citados até o momento ganham proporções ainda maiores se considerarmos que todos os canais de TV e rádio da BBC veiculam de 1000 a 1500 programas por dia. Até meados da década de 2000, os sites destes programas ainda eram produzidos da forma tradicional: desenvolvimento manual de um *layout* específico para o programa com XHTML e CSS. Essa lógica resultava na produção de sites apenas para os grandes programas da emissora. Segundo Raimond et al. (2010a), a BBC deixava de aproveitar a cauda longa[123] da imensa quantidade de conteúdos produzidos de forma distribuída em centenas de programas que não estavam presentes na web.

A partir destes pressupostos, em 2007, foi lançado o site BBC Programmes, que reúne os sites dos programas da emissora. Nele, cada programa possui uma URI que o identifica na web. Também foi desenvolvida uma ontologia para o Programmes, que definem um modelo de conceitos, como, por exemplo, uma *Brand* (traduzido por nós como franquia) possui *Series* (traduzido por nós como seriados) que possui *Episodes* (episódios). A ontologia ainda tem outros conceitos com vários tipos de relacionamentos entre eles, que formam um modelo de organização do conteúdo e que informa a projetos externos o que eles representam. Além de possuir uma ontologia, o BBC Programmes ainda associa metadados aos programas através da técnica de *tagging*. Essas *tags* são baseadas em um vocabulário controlado e compartilhado (o Dublin Core), que possui predicados apropriados para produções editoriais, tais como "autor", "formato", "gênero", "licença" e "direitos".

Segundo Raimond et al. (2010a, 2010b), o site Wildlife provê um **identificador único** da web **(URI)** para cada espécie (e outros níveis da taxonomia), cada habitat e cada comportamento/adaptação. Desta maneira, o site mantém a lógica de utilizar URIs para identificar conceitos do mudo real, ao invés da lógica tradicional de identificar páginas (idem, 2010a). Em outras palavras: as URIs identificam recursos ao invés de identificar apenas uma página HTML, e esses recursos podem inseridos como URIs nas triplas RDF.

Com uma URI para cada recurso, o site utiliza o sistema de triplas RDF para relacionar um conceito a outro. Assim, a URI do conceito "leão" é associado ao conceito de

[123] "Cauda longa" é o termo utilizado para a situação em que a soma dos produtos menos consumidos em um determinado mercado pode acumular valor aproximado ou comparável ao valor dos produtos mais vendidos. O fenômeno ocorre porque a diversidade de produtos com consumo baixo é muito maior do que a dos produtos mais consumidos (os *hits*) (ANDERSON, 2006). O fenômeno pode ser representado por um gráfico em um plano cartesiano, que toma a forma semelhante a uma cauda comprida, justificando assim o nome "cauda longa".

“vertebrados”, através do uso de um predicado apropriado, definido pela ontologia. Como os programas da BBC possuem URI (pelo site Programmes), então é possível utilizar triplas RDF para associar a produção destes programas às URIs do Wildlife.

Seguindo a lógica da Web Semântica, de tornar as informações compreensíveis tanto por humanos quanto por máquinas, o site Wildlife oferece as páginas em dois formatos: em HTML (para leitura humana) e em RDF (para as máquinas). Para que isso seja possível, basta que um desenvolvedor ou uma máquina (agente) acesse o endereço de uma página do Wildlife utilizando a extensão **.rdf** no final da URL. Este processo faz com que o servidor envie para o cliente um arquivo RDF, ao invés do arquivo HTML (ver Figura 45). O arquivo é serializado no formato RDF/XML, ou seja, as triplas RDF são escritas com a sintaxe do XML. Este processo é chamado de *content-negotiation*: um mecanismo do protocolo HTTP que permite ao cliente solicitar ao servidor o envio de outros tipos de arquivos a partir de um único URI. Devido a essa possibilidade, os desenvolvedores afirmam que o Wildlife não necessita de uma API, pois o próprio site é uma API (RAIMOND et al., 2010b).

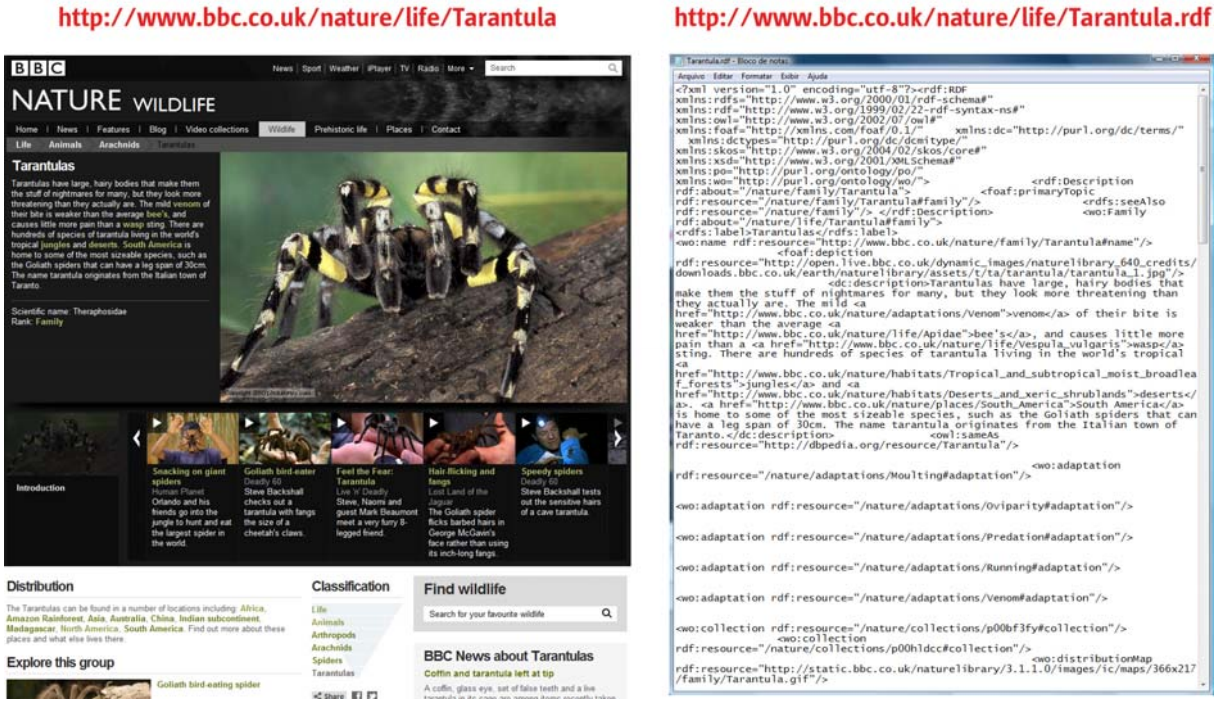


Figura 45 – À esquerda, a página da espécie Tarântula. À direita, a página serializada em RDF/XML

Ao analisarmos o arquivo RDF da espécie tarântula (Figura 45), foi possível encontrar linhas de código que relacionam a URI da espécie a URIs de vídeos disponibilizados no BBC Programmes. Na Figura 46, é mostrado um trecho deste arquivo RDF/XML com duas triplas combinadas para um mesmo sujeito (que acaba formando um grafo). O sujeito é representado

pela linha 1 (URI de um clipe de vídeo do site Programmes) e forma uma tripla com a URI da linha 2 (que indica o título do vídeo) e outra tripla com a URI da linha 3 (que indica a relação do vídeo com a URI da tarântula).

```
1 | <po:Clip rdf:about="http://www.bbc.co.uk/programmes/p00dlr0s#programme">
2 |     <dc:title>Snacking on giant spiders</dc:title>
3 |     <po:subject rdf:resource="/nature/family/Tarantula#family"/>
4 | </po:Clip>
```

Figura 46 – Triplas RDF que descrevem um vídeo do site BBC Programmes

Para facilitar a compreensão do código apresentado na Figura 46, poderíamos dividir o código XML em duas triplas RDF, como mostra a Figura 47:

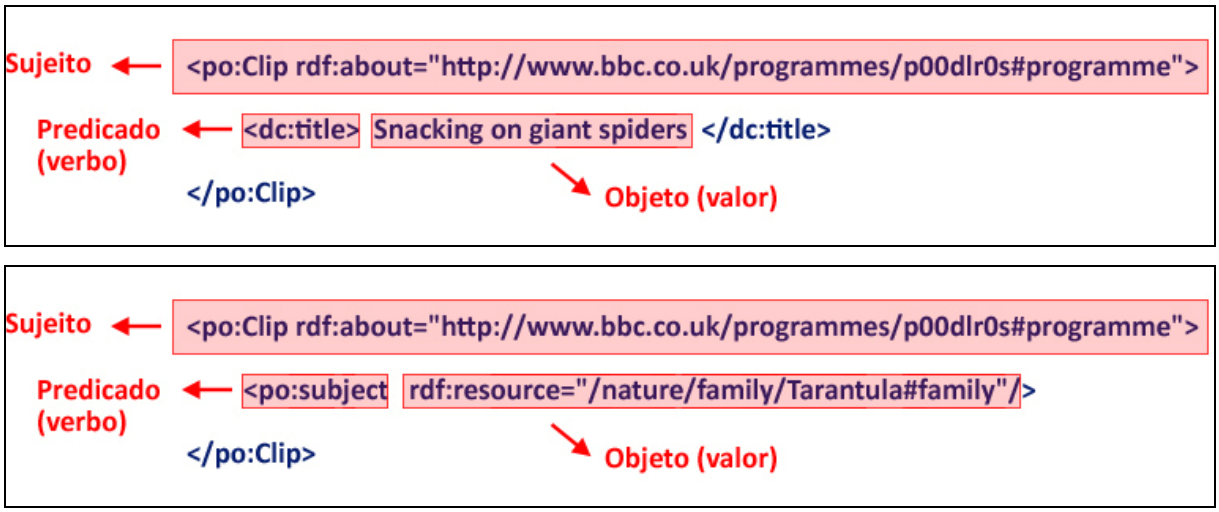


Figura 47 – Triplas RDF que descrevem um vídeo do site BBC Programmes

Uma forma ainda mais fácil de compreender essas relações é a partir de um grafo. Desenvolvemos na Figura 48 um grafo que representa as relações entre as triplas RDF:

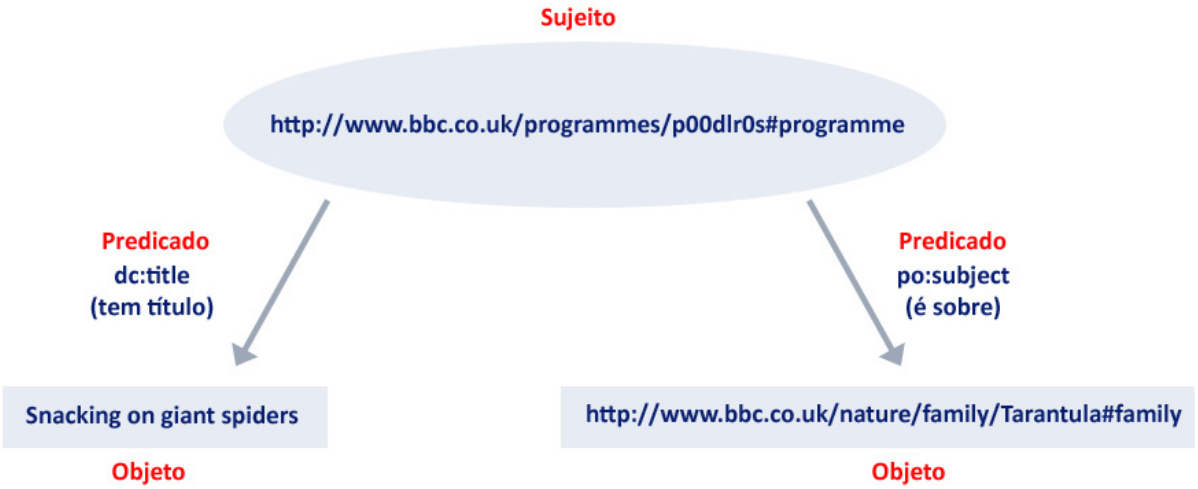


Figura 48 – Grafo das triplas que descrevem um vídeo do site BBC Programmes

Ainda sobre os triplos representados no código XML da Figura 46, a combinação **po:** é uma abreviação que identifica o endereço da ontologia do site Programmes (Programmes Ontology[124]), utilizada para definir termos (e seus relacionamentos) do universo dos programas da BBC, como o que é um clipe, um seriado, um episódio etc. Então, <po:Clip> significa que o sujeito em questão faz parte da classe "Clip" da ontologia do Programmes (de acordo com a ontologia, a classe Clip define clipes multimídia que fazem parte de episódios). Ou seja: o recurso identificado pela URI é um clipe. Já a combinação **dc:** é uma abreviação para o endereço do vocabulário Dublin Core[125], utilizado para definir termos editoriais, como autoria, título, ano de publicação etc. Então, o código <dc:title> é um predicado definido pelo vocabulário Dublin Core, que indica o título do sujeito. Por fim, o código <po:subject> é um predicado da ontologia do site Programmes que relaciona um produto a um assunto.

Na Figura 49, é possível observar que, na versão em HTML da página, o clipe de vídeo descrito pela tripla RDF é disponibilizado no site de seu programa de origem (à direita da figura, no BBC Programmes) e também na página do Wildlife (à esquerda da figura).

[124] Disponível em: <http://www.bbc.co.uk/ontologies/programmes/2009-09-07.shtml>. Acesso em: 12 jan. 2012.
[125] Disponível em: <http://purl.org/dc/elements/1.1/>. Acesso em: 12 jan. 2012.

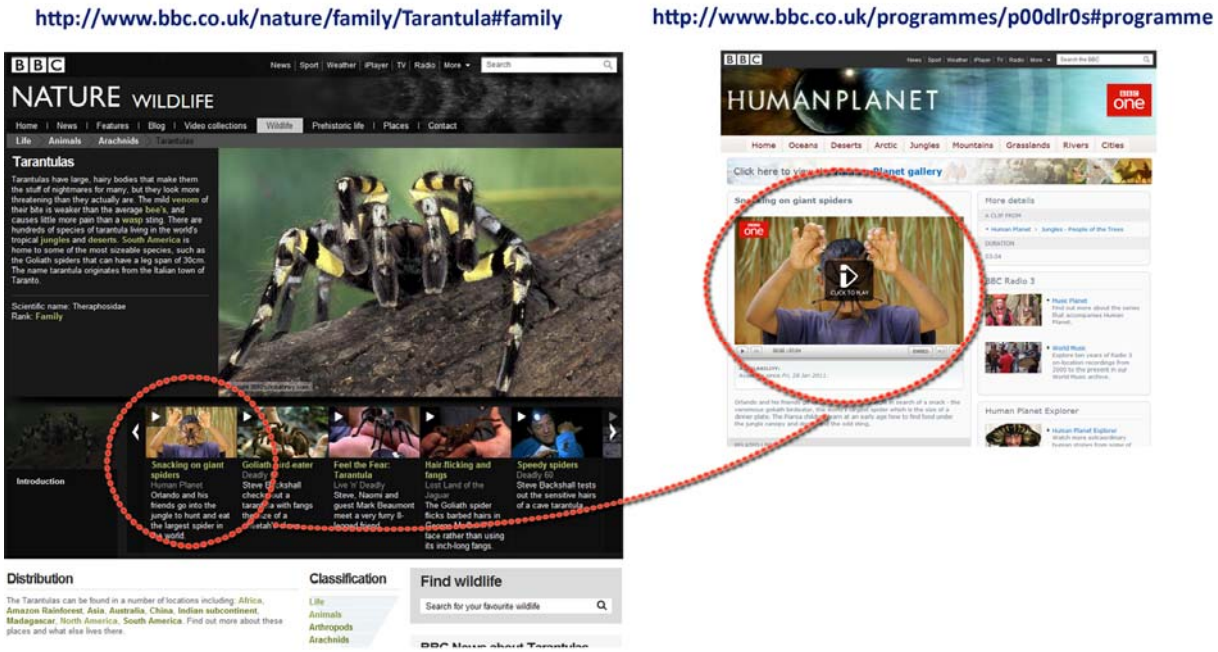


Figura 49 – Clipe de vídeo do BBC Programmes agregado à página do Wildlife

As triplas RDF indicam às máquinas o tipo de relacionamento que existe entre o recurso (o conceito presente na página da Wildlife) e os clipes de vídeos do BBC Programmes. Esse relacionamento ocorre com a associação de URIs via triplas RDF, como visto na Figura 46. Porém, antes de expressar esse relacionamento em RDF, é necessário extrair conceitos do clipe de vídeo. Afinal, um clipe de vídeo pode ser agregado a outras páginas além da página da espécie. Para a extração destes conceitos, é realizado o **processo de *tagging*** nestes clipes de vídeos (RAIMOND et al., 2010a).

Os conteúdos do Wildlife não são constituídos apenas por vídeos do BBC Programmes. A maior parte das mais de mil páginas do Wildlife possui descrições textuais sobre o assunto tratado. Para que isso pudesse ser possível, o sistema foi projetado para **reaproveitar conteúdos da web** de forma automatizada, mais especificamente da Wikipedia. Para Oliver (2010), esse reaproveitamento é benéfico para ambos os lados, pois se a BBC tem a vantagem de reaproveitar conteúdos moderados por uma comunidade com milhões de usuários, a Wikipedia, por sua vez, recebe em troca o constante enriquecimento de seus conteúdos por parte de profissionais da BBC, que se preocupam em manter a qualidade das informações que reaproveitam.

Para facilitar a integração automatizada entre os conteúdos do Wildlife e da Wikipedia, os desenvolvedores decidiram padronizar as URIs do site de acordo com os

mesmos identificadores utilizados pela enciclopédia (RAIMOND et al., 2010b). Tomamos como exemplo a espécie “leão”, que possui os seguintes identificadores nos respectivos sites:

- Endereço utilizado pela Wikipedia: http://en.wikipedia.org/wiki/Lion
- Endereço utilizado pelo Wildlife: http://www.bbc.co.uk/nature/life/Lion

Essa sincronização de identificadores com a Wikipedia é realizada com o apoio do projeto DBpedia, que recupera os dados estruturados que estão armazenados na Wikipedia e os publicam em RDF. Em outras palavras, o Wildlife adota a DBpedia como um vocabulário controlado de termos, o que facilita a identificação de recursos e a interoperabilidade com outros projetos da Linked Data (SCOTT, 2010).

Os identificadores da DBpedia também são utilizados como vocabulário padrão para o **processo de *tagging*** aplicado aos clipes de vídeos oriundos do BBC Programmes (SCOTT, 2009; RAIMOND et al., 2010a). Os nomes de espécies definidos pela Wikipedia acabam por descrever o significado dos clipes de vídeos. Desta maneira, é possível agregar automaticamente em uma página do Wildlife tanto as informações da Wikipedia quanto os vídeos do BBC Programmes sobre uma determinada espécie, pois ambos utilizam o mesmo identificador (SCOTT, 2009).

Além da Wikipedia, o site também reaproveita conteúdos de outras fontes, como os dados sobre conservação de animais da ONG WWF (*World Wild Life*)[126], as classificações sobre comportamentos e habitats do site "*Animal Diversity Web*"[127] do Museu de Zoologia da Universidade de Michigan, e, por fim, da “lista vermelha” de animais que correm perigo de extinção organizada pela ONG IUCN (*International Union for Conservation of Nature*)[128] (SCOTT, 2009; RAIMOND et al., 2010a).

Raimond et al. (2010b) ainda explicam que parte do conteúdo editorial da BBC continua sendo produzido sem seguir os princípios da Web Semântica. Por isso, para que seja possível aproveitar estes conteúdos, a emissora também aplica *tags* a este conteúdo, baseados no vocabulário da DBpedia. Desta forma, as páginas do Wildlife passam a agregar notícias e reportagens da BBC, além dos vídeos do Programmes e dos conteúdos de outros sites da web.

A **ontologia Wildlife** foi escrita em RDF e é disponibilizada na web para acesso público[129]. Nela, foi estruturado um modelo para representar conceitos e relacionamentos entre as espécies e os outros níveis da taxonomia biológica, além dos conceitos de habitats,

[126] http://www.worldwildlife.org/science/data/item1872.html
[127] http://animaldiversity.ummz.umich.edu/site/index.html
[128] http://www.iucnredlist.org/
[129] http://www.bbc.co.uk/ontologies/wildlife/

comportamentos/adaptações, biorregiões e status de conservação (DODDS; SCOTT, 2010). Dessa maneira, a ontologia se torna o modelo de estruturação do site, pois a criação dinâmica de páginas respeitam esse modelo. Oliver (2010) apresenta um gráfico (Figura 50) que simplifica o funcionamento do sistema de publicação dinâmico e semântico do Wildlife.

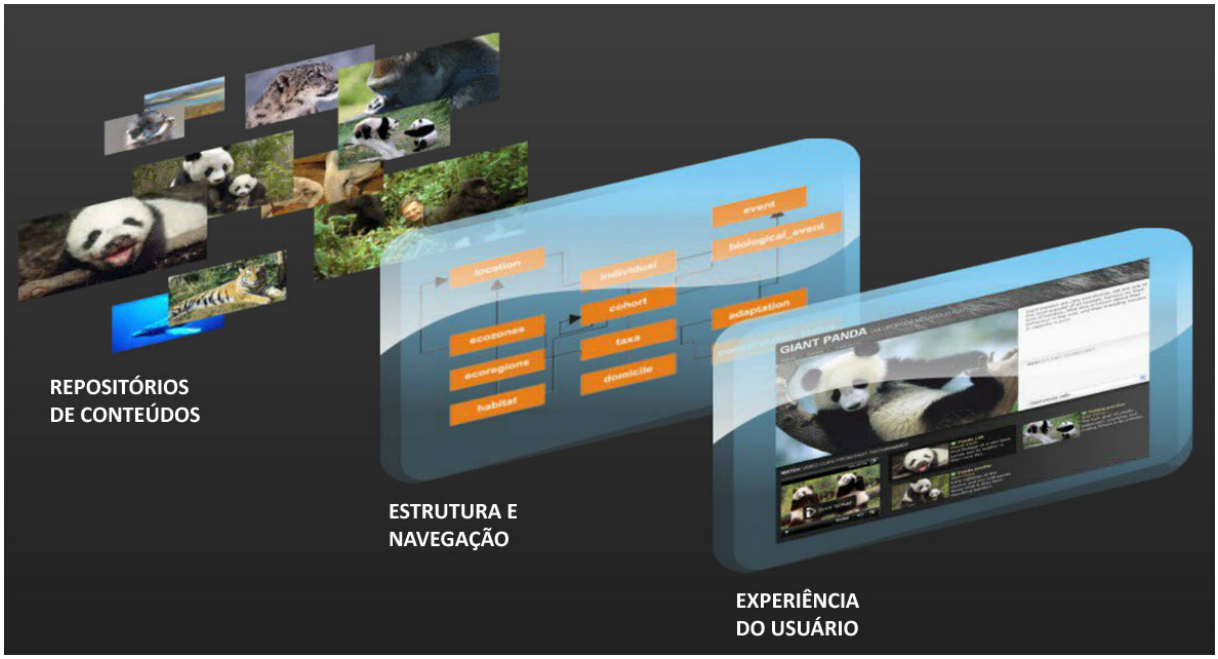


Figura 50 – Camadas que fazem o fluxo de publicação dinâmica e semântica do BBC Wildlife (OLIVER, 2010b, tradução nossa)

O gráfico é muito semelhante a outro que explica o funcionamento do site BBC World Cup 2010. Nele, é possível observar três camadas: a primeira, representada por fotografias de animais, é a camada de conteúdos armazenados em repositórios, produzidos por programas da BBC e descritos com metadados pelo processo de *tagging*. Na segunda camada, o modelo de conceitos e de seus relacionamentos definidos na ontologia, que determina a estrutura e a navegação do site. Após a associação entre as *tags* dos conteúdos e o modelo da ontologia, são criadas páginas que recebem apenas os conteúdos selecionados pelo modelo da ontologia e, finalmente, apresentadas para os usuários. Além das páginas HTML, também são criados os arquivos em RDF/XML.

Segundo Dodds e Scott (2010), autores da ontologia, o desenvolvimento da mesma teve o cuidado de manter condições para que no futuro ela possa se inter-relacionar a outras ontologias especializadas, tais como as especializadas em ecologia, bioinformática ou outras ciências, permitindo a interoperabilidade de dados, que, de certa maneira, funcionam como uma expansão do modelo Wildlife.

O site Wildlife é um projeto de grande envergadura, tanto em relação à quantidade de informações quanto à complexidade das tecnologias empregadas. Por isso, para que possamos melhor compreender o funcionamento das tecnologias semânticas, recapitulamos os principais pontos da explicação desenvolvida até o momento e listamos logo abaixo, de forma resumida:

- O site possui um vocabulário (lista de termos) para as espécies e para os outros níveis da taxonomia. Tais termos são oriundos do esquema de nomes utilizados pela Wikipedia, através do projeto DBpedia.
- Os clipes de vídeos são oriundos dos documentários produzidos pelos programas da BBC, e passam por processo de tagging, no tipo "vocabulário controlado", pois são utilizados os termos da DBpedia.
- O modelo da ontologia do Wildlife define conceitos (e relações entre estes conceitos) para o domínio natural: níveis da classificação biológica, habitats, comportamentos/adaptações e biorregiões. Este modelo se torna a estrutura de organização e navegação do site.
- O Wildlife monta páginas dinâmicas de acordo com o modelo da ontologia. Para as espécies, os conceitos são delimitados pelos termos da DBpedia. Já os conceitos de habitats, comportamentos/adaptações e biorregiões são definidos de acordo com os dados de outros sites da web (ex.: WWF).
- As páginas dinâmicas agregam diversos conteúdos relacionados ao conceito, de modo automático. A relação entre conteúdos (textos, áudios, vídeos) e conceitos ocorre pela comparação entre os metadados associados aos conteúdos e o modelo da ontologia.
- O site Wildlife foi inserido no escopo do BBC Nature, um portal de conteúdos editoriais sobre a temática natureza. Notícias, reportagens e blogs do portal e de outros sites da BBC reaproveitam os conteúdos do Wildlife para complementar e contextualizar as informações jornalísticas, assim como o material jornalístico também enriquece as páginas do Wildlife.
- Por fim, o mecanismo de content-negotiation permite às máquinas solicitarem ao servidor do Wildlife que, ao invés de uma página HTML, seja enviado um arquivo serializado RDF/XML, e isso possibilita o compartilhamento destes conteúdos com as páginas da BBC e com outras iniciativas da Linked Data.

Finalizamos aqui a descrição do segundo caso estudado. No próximo tópico, passamos para a análise das contribuições das tecnologias identificadas, baseando-nos nas categorias do JDBD propostas por Barbosa (2007, 2008a).

3.3.5 Contribuições das tecnologias semânticas ao atual paradigma do JDBD

Ao associarmos os depoimentos dos desenvolvedores da BBC com as categorias de análise elencadas por Barbosa (2007, 2008a), o sistema de publicação semântico apresentou possibilidades de potencialização em grande parte das categorias, principalmente na de automatização. A seguir, realizamos uma análise em cada categoria do JDBD baseados nos dois casos estudados.

3.3.5.1 Dinamicidade

Assim como ocorreu no primeiro caso estudado, o site BBC Wildlife apresentou alto nível de dinamicidade, embora atualmente grande parte dos sites que utilizam BDs já possam ser considerados sistemas bastante dinâmicos, devido à lógica da separação entre conteúdo e apresentação, pois tal separação exige um sistema dinâmico de publicação. A dinamicidade, no caso do Wildlife, é potencializada devido à autonomia que o sistema semântico tem em decidir como as entidades devem ser relacionadas entre si e, em consequência, como os menus devem ser criados. Assim, da mesma forma como ocorreu no caso estudado anteriormente (BBC World Cup 2010), a dinamicidade no sistema semântico se aplica não apenas nas operações mecânicas de publicação, mas também nas operações mais complexas de tomada de decisão. A potencialização da dinamicidade está diretamente relacionada à próxima categoria: a da automatização.

3.3.5.2 Automatização

A automatização é total na publicação do conteúdo do site. Assim como o primeiro caso estudado, o sistema do BBC Wildlife ainda exige a operação manual de associação dos conteúdos aos metadados (*tags*), mas, após essa operação, a publicação e a organização das páginas das espécies e das páginas agregadoras apresentam automatização total, tanto nos conteúdos (mídias, títulos, descrições, estatísticas etc.) quanto na estrutura de navegação (listas dinâmicas de links).

Nos sistemas tradicionais, a análise das *tags* é realizada com uma estratégia de comparação sintática (ex.: semelhança ou igualdade das sintaxes). A comparação sintática pode ser ambígua, pois compara igualdade de palavras, não de significados. No caso do site Wildlife, é utilizada uma abordagem semântica no processo de comparação entre as *tags* e o modelo de conceitos (ontologia), resultando em inferências que maximizam a autonomia das máquinas no processo de publicação.

3.3.5.3 Flexibilidade

O caso estudado demonstrou que as produções de equipes diferentes e dispersas podem ser reunidas de forma automática em um mesmo produto, o que torna o processo produtivo mais flexível do que uma produção centralizada. Os produtos atuais do JDBD que não utilizam tecnologias semânticas também permitem a produção descentralizada com o uso de sistemas gerenciadores de conteúdo (CMS); porém, geralmente exigem o emprego de um mesmo CMS entre as equipes. No sistema semântico, os conteúdos podem ser armazenados em diferentes bases de dados, e ainda assim serem integrados, porém, desde que apresentem certas condições para a integração, como o fornecimento de metadados (*tags*) ou de versões serializadas do RDF.

Outra contribuição à flexibilidade é o fato de o site deixar de utilizar o tradicional menu centralizado e imutável, e passar a adotar as próprias páginas como recurso de navegação, ou seja, o site possui uma navegação contextual, que pode ser reorganizada com a mudança do modelo da ontologia. Consideramos essa característica como um enriquecimento da flexibilidade na estrutura e na navegação.

3.3.5.4 Inter-relacionamento/Hiperlinkagem

Na Web Semântica, o inter-relacionamento automatizado entre conteúdos é baseado em significados, e não apenas da igualdade de sintaxes entre palavras-chaves, o que maximiza a qualidade desses relacionamentos. As inferências realizadas com o inter-relacionamento baseado em ontologias permitem que o sistema origine coleções de conceitos relacionados ao assunto da página, gerados no formato de listas de links contextualizados, que funcionam como menus de navegação para outras páginas, maximizando assim a hiperlikagem.

3.3.5.5 Densidade informativa

Consideramos que as listas contextualizadas de links, que funcionam como menus nas páginas, aumentam a densidade informativa da matéria, pois além de servirem como recurso de navegação, também informam ao usuário que o determinado conceito possui certas características, como no caso da página da espécie leão, mostrada na Figura 37, em que as listas de links da parte inferior da página, ao mesmo tempo em que servem de menu para navegação, também informam que o leão pode viver em cinco habitats além da savana africana, tais como o deserto, a floresta mediterrânea ou as pastagens alagadas. Então, neste caso, a qualidade da categoria de inter-relaiconamento/hiperlinkagem contribui para a densidade informativa do produto.

Além disso, a densidade informativa foi enriquecida com a convergência de conteúdos agregados de diferentes sites internos e externos à BBC; afinal, a densidade informativa não diz respeito apenas à quantidade de informações, mas também a diversidade das mesmas.

3.3.5.6 Diversidade temática

Assim como no primeiro caso estudado, no site Wildlife há a predominância de um tema: o mundo natural. Entretanto, as páginas dedicadas às espécies demonstraram capacidade de agregação de diversos conteúdos relacionados a um tema, como no caso da caixa de notícias para a espécie leão, ou ainda na formulação automática de coleções de vídeos, que podem tratar sobre temas distintos que mencionam o mesmo animal.

Em relação às listas dinâmicas de links contextualizados ao conceito tratado na página, que funcionam como menus, podemos considerá-las uma maneira de aumentar a diversidade temática, já que a ontologia auxilia o sistema a "descobrir" tópicos diversos em relação ao conceito tratado na página, como, por exemplo, as características do animal.

3.3.5.7 Visualização

Consideramos que na categoria de visualização não houve contribuição relevante em relação ao que já é praticado em produtos da web sintática. As páginas seguem a estética base de dados: são *layouts* formados por imagens, textos, caixas e links com dimensões delimitadas pelos dados das BDs.

3.3.5.8 Convergência

O Wildlife agrega conteúdos de diferentes formatos, como textos, áudios e vídeos. Em um primeiro momento, tal situação poderia caracterizar uma contribuição da Web Semântica à categoria da convergência**.** Porém acreditamos que a convergência se destaca no produto estudado devido a outra questão que vai além da convergência de mídias: a capacidade do sistema semântico de convergir conteúdos oriundos de diferentes fontes da web. As páginas são espaços agregadores de conteúdos externos: vídeos do BBC Programmes, descrições e identificadores da Wikipédia via DBpedia, notícias e reportagens da BBC News. Essa característica demonstra que a união entre identificadores únicos e consistentes (URI) com um modelo que define conceitos e relacionamentos (ontologia) na Web Semântica é uma combinação que potencializa a capacidade de compartilhamento, pois facilita a interoperabilidade e evita as ambiguidades.

3.4 Avaliação geral sobre o uso das tecnologias semânticas no jornalismo digital

Acreditamos que as funções das tecnologias semânticas, apresentadas nos casos BBC World Cup 2010 e BBC Wildlife, fazem parte do conceito do Jornalismo Digital em Base de Dados. Afinal, os sistemas apresentados em ambos os casos também tiveram como função a organização de conteúdos que já estavam previamente armazenados em bases de dados tradicionais. As principais operações das tecnologias semânticas ocorreram em uma camada acima dos conteúdos armazenados: a dos metadados. Por isso, no jornalismo digital, Web Semântica e bases de dados relacionais podem ser complementares, da mesma forma que os documentos hipertextuais em HTML continuaram existindo com o surgimento das BDs.

A partir dos casos estudados, observamos que as tecnologias semânticas podem contribuir com alguns avanços em determinadas funções desempenhadas pelos atuais sistemas em bases de dados. Para fins de comparação, recuperamos uma afirmação de Palacios (2003), de que as características do jornalismo digital não são necessariamente rupturas em relação às práticas tradicionais do impresso, da TV e do rádio, pois são, na maioria, continuidades e potencializações. Como exemplo, ele cita que a característica da multimidialidade no suporte digital é de certa forma uma continuidade, já que a televisão já fazia a convergência entre imagem, som e texto. Da mesma forma, a característica hipertextualidade já ocorria antes da web, em produtos armazenados em CD-ROM. O que a internet e a web fazem é potencializar tais características, devido ao aproveitamento de recursos técnicos que as redes digitais

oferecem. Para Palacios, a especificidade do jornalismo digital está nestas potencializações das características, mas não apenas de forma isolada: a especificidade está, principalmente, na combinação das características potencializadas.

Concluímos que, neste caso apresentado, as características do JDBD podem ser potencializadas em determinados contextos, devido, principalmente, à combinação das mesmas com a eficiente automatização do sistema semântico. Sabemos que os atuais produtos digitais em bases de dados relacionais podem ser automatizados e muitas vezes dispensam as operações manuais (automatização total). Porém, a Web Semântica se coloca como solução vantajosa em relação ao atual cenário, principalmente devido ao uso de ontologias, que enriquecem a qualidade da automatização no gerenciamento de informações.

Como exemplo, ilustramos o caso do *site* Google News[130], que apresenta processo de automatização total (BARBOSA, 2007). O *site* apresenta notícias procedentes de diversas fontes e as organiza em listas de acordo com determinadas editorias. O sistema utiliza algoritmo próprio do Google para associar palavras-chaves às notícias publicadas nos últimos 30 dias (DONG, SMITH e BUCHANAN, 2011). Por mais que seja um sistema automatizado, e por mais que seja um serviço eficiente ao apresentar resultados relevantes em relação à pesquisa feita pelo usuário, o Google News ainda apresenta falhas na identificação de significados. Para ilustração, realizamos um teste (ANEXO A) na versão norte-americana do *site*: clicamos no link "Rio Grande do Sul" (opção oferecida dinamicamente no menu do Google News norte-americano) para listar notícias relativas ao estado gaúcho. O site retornou diversas notícias que não tinham relação com a palavra-chave da pesquisa, devido à falha no reconhecimento de conceitos. Entre as notícias listadas, encontramos os títulos "*Reading mayor chooses Lenin Agudo for community-development director*" e "*Garibaldi wins Obispo concession at Sonora Lottery*", o que demonstra que o site falha ao considerar que os termos *Agudo* (um sobrenome na primeira chamada) e *Garibaldi* (um nome de empresa na segunda chamada) sejam nomes de cidades do estado do Rio Grande do Sul. Os algoritmos do Google foram eficientes para buscar em uma BD nomes de cidades do Estado do Rio Grande do Sul, mas falhou na identificação de significados dentro dos conteúdos das notícias, pois realizou apenas uma comparação sintática entre palavras-chaves. Uma abordagem semântica neste sistema poderia evitar tais ambiguidades.

Tratamos neste texto que os avanços da Web Semântica são continuidades e potencialidades do que já é encontrado nos atuais produtos do Jornalismo Digital em Base de

[130] http://news.google.com.

Dados. Porém, acreditamos que seja possível indicar uma possível ruptura que a Web Semântica traz ao Jornalismo Digital em Base de Dados: a da interoperabilidade automatizada entre diferentes sites e serviços. Tal característica ganha importância com o massivo crescimento da quantidade de dados publicados no ciberespaço, que resulta em duplicidades nos processos de produção e reprodução da informação. A interoperabilidade automatizada, que permite o reaproveitamento de conteúdos em um ambiente que produz dados de forma massiva, pode ser vantajosa para as empresas jornalísticas, pois poupa recursos na produção, e para os jornalistas, pois poupa esforços na produção de algo já existente. Sabemos que, na lógica do mercado capitalista, seria utópico esperar que empresas jornalísticas compartilhassem os seus esforços na produção conjunta e complementar das mesmas notícias. Porém, como o caso estudado demonstrou, é possível reaproveitar informações originadas em diferentes projetos da web que sejam abertas ao compartilhamento e que possam ser confiáveis, tais como os outros sites da mesma empresa, relatórios e estatísticas de ONGs e fontes de dados oficiais, como no caso das páginas das espécies no BBC Wildlife, que reaproveitava automaticamente conteúdos oriundos da Wikipédia e da ONG World Wildlife Fund (WWF). Além do reaproveitamento de conteúdos de terceiros, os casos demonstraram que o reaproveitamento pode ocorrer entre diferentes produtos da mesma empresa, como no caso das notícias do principal site de notícias da BBC e dos artigos de diferentes blogs, que alimentavam (e enriqueciam) as páginas dos times e dos jogadores no BBC World Cup 2010.

Para que essa ruptura venha a se consolidar na prática jornalística, concordamos com a visão de Berners-Lee (2006): é necessário que surjam mais iniciativas em que produtores de conteúdos se adaptem aos padrões da Web Semântica, pois só assim é possível uma interoperabilidade eficiente. Sem padrão, não há convenções; sem convenções, não há comunicação entre os sites e serviços independentes. Além da adoção de padrões, outra prática recomendada por Berners-Lee (2006) é a de se manter a cultura da abertura de dados e de criar interconexões entre repositórios, como ocorre no projeto Linked Data, que cresce significativamente a cada ano. Outra condição (bastante lógica) para a consolidação desta ruptura é a do jornalismo começar a explorar as tecnologias semânticas com o desenvolvimento de produtos compatíveis com esta proposta. Para isso, seria necessária uma aproximação maior dos campos do Jornalismo, da Ciência da Informação e da Ciência da Computação.

CONSIDERAÇÕES FINAIS

A proposta do presente trabalho foi a de analisar a aplicação da Web Semântica em dois produtos jornalísticos, o BBC World Cup 2010 e o BBC Nature, a fim de se compreender como esta tecnologia pode contribuir com o jornalismo digital, principalmente na organização e no gerenciamento das informações jornalísticas.

No decorrer do referencial teórico, vimos que a base de dados é a tecnologia estruturante dos produtos digitais informacionais de nosso tempo. A necessidade de se armazenar grandes quantidades de dados não é a única razão pela adoção das BDs como lógica estruturante. Além de estrutura, elas são recursos técnicos que potencializam o gerenciamento dos dados até então realizados pelos humanos. Encurtam o tempo, maximizam a eficácia de operações, enriquecem as possibilidades de combinações entre dados e informações. Com o desenvolvimento do jornalismo digital, essa prática profissional passou a adotar as bases de dados como estrutura dos produtos jornalísticos.

Na nossa análise, percebemos que as tecnologias semânticas potencializam algumas das funções atualmente desempenhadas por bases de dados relacionais no jornalismo digital. Concluímos que duas categorias do JDBD se mostraram mais propícias a serem potencializadas: a automatização e a convergência.

Em relação à automatização, destacamos as ontologias, que garantem às máquinas a capacidade de identificar conceitos, de relacioná-los eficientemente e de gerar inferências. Esta última implicação, a da geração de inferências, conferem aos sistemas semânticos a vantagem da autonomia às máquinas na tomada de decisões, como, por exemplo, em como criar automaticamente novos menus para determinados contextos.

Na categoria da convergência, que tomamos aqui como um conceito maior do que a simples convergência de mídias em um suporte, a Web Semântica apresenta uma importante contribuição, pois, graças à URI, que identifica recursos da web sem ambiguidades, e aos vocabulários, que padronizam termos e conceitos, é possível convergir em um mesmo produto conteúdos de diferentes formatos, oriundos de diferentes fontes, mas que tratam do mesmo conceito.

Por fim, consideramos que a Web Semântica pode vir a representar um salto ainda maior do que uma potencialização de características até então exploradas. Esta possível ruptura seria a interoperabilidade automatizada. Ela permite que diferentes sites (que estejam formatados na lógica da Web Semântica) troquem entre si dados e informações de

maneira automatizada, a partir de associações de conceitos definidos por vocabulários ou ontologias compartilhados. Acreditamos que esse é um salto significativo porque resulta em diversas potencializações:

a) A diversidade de fontes de dados (tanto em quantidade quanto em tipo) pode **enriquecer o produto informacional** em diferentes categorias, como no inter-relacionamento/hiperlinkagem, na diversidade informativa e na diversidade temática. Foi o que ocorreu nas paginas das espécies do Wildlife: para cada espécie, uma combinação de conteúdos provenientes de diferentes sites da web formava uma página mais rica em cada uma das três categorias citadas.
b) A convergência de dados oriundos de fontes diversificadas aumenta em grande proporção a vantagem do **reaproveitamento** de dados e informações produzidas por terceiros[131]. O reaproveitamento pode resultar em três benefícios evidentes: no enriquecimento do produto, como no caso do Wildlife, que teve as mais de mil páginas alimentadas por informações especializadas que eram constantemente atualizadas pelas fontes de dados; na rotina produtiva dos jornalistas, já que o reaproveitamento libera tempo de produção e permite aos profissionais se dedicarem a outros projetos, e na integração de equipes, já que o reaproveitamento pode ocorrer com informações factuais recém publicadas por outras equipes da mesma organização.
c) O constante compartilhamento entre vários projetos complementares entre si podem formar uma **base de conhecimento compartilhada** que cresce de forma colaborativa. É o que ocorre hoje no Linked Data, em que diferentes projetos, geralmente especializado em determinados temas, permitem a consulta a seus dados com pesquisas *query* via SPARQL. Em outras palavras: é como se os projetos formassem uma grande base de dados distribuída e compartilhada.

A interoperabilidade ganha destaque na Web Semântica porque, ao contrário da maioria dos sistemas em base de dados relacionais, nela há uma premência pelo uso de padrões abertos, o que facilita a comunicação entre sites e serviços que utilizam os mesmos padrões. Outra razão é o modelo utilizado na Web Semântica de se compartilhar conceitos de predicados entre os sites, pois essa lógica evita o problema das conceituações conflitantes entre bases de dados relacionais, que são projetadas com seus predicados

[131] Quanto à questão da confiabilidade da fonte, é um tema que merece discussões, mas ao refletirmos nos casos estudados, concluímos que o reaproveitamento pode ser aplicado em fontes seguras, como outros sites da mesma organização, de organizações parceiras, de ONGs consolidadas e de fontes oficiais.

próprios e arbitrários. Sem uma convenção de conceitos e relacionamentos, a interoperabilidade é dificultada.

Tratamos a interoperabilidade automatizada como uma potencial ruptura no jornalismo digital. Porém, reforçamos que ainda potencial. Para que ela de fato seja uma ruptura, é preciso que certas condições sejam satisfeitas. A mais básica delas é a popularização de produtos jornalísticos que de fato aproveitam as tecnologias semânticas. Outra condição é a da proliferação de repositórios de dados abertos e apropriados para a lógica da Web Semântica, como ocorre com o projeto da Linked Data. Por fim, uma condição necessária para o desencadeamento desta ruptura seria o desenvolvimento de produtos jornalísticos criativos, que saibam como explorar o reaproveitamento de dados.

Defendemos que os dois casos estudados nesta pesquisa são exemplos que satisfazem estas condições. O resultado é evidente: além dos números e estatísticas apresentados pelos desenvolvedores em seus depoimentos publicados na web, também temos como prova uma experiência relativamente simples, porém reveladora, realizada por nós: em uma busca pelo termo "lion" no site de busca do Google (versão em inglês), o resultado indicou a existência de mais de 78 milhões de sites, e a página da espécie leão no BBC Wildlife apareceu em sétimo lugar (ANEXO J). Realizamos outro experimento, com resultado ainda mais significativo: buscamos pelos termos "World Cup 2010", que indicou mais de 325 milhões de sites, e o site BBC World Cup 2010 aparece em quarto lugar, perdendo apenas para as duas páginas oficiais da FIFA e a página da Wikipedia (ANEXO K).

Se a Web Semântica apresenta tantas contribuições, se as tecnologias semânticas já existem há aproximadamente uma década e se já há exemplos de sucesso na web, então não teríamos como fugir da inevitável questão: por que ela não é explorada mais intensamente pelo jornalismo digital? Sabemos que as empresas jornalísticas não decidem adotar as novas tecnologias de forma sincronizada. A adoção é gradativa, e algumas das empresas nem sequer aproveitam características das primeiras gerações do jornalismo digital, como a hipertextualidade em narrativas. No caso da Web Semântica, temos um agravante: a adoção destas tecnologias pode ser um processo difícil e demorado por parte dos desenvolvedores, devido a diversas razões relacionadas a um sistema complexo e ainda em processo de maturação. Para Kashyap et al. (2008), a Web Semântica já apresenta na prática várias vantagens e qualidade, mas também apresenta problemas, que podem se tornar obstáculos para o seu progresso, como, por exemplo: a curva de aprendizagem sobre o funcionamento do RDF e da OWL; problemas de integração entre serviços; a dificuldade em obter acordos sobre os conceitos dos termos definidos em ontologias; iniciativas privadas que se recusam a

compartilhar seus conhecimentos (ontologias); a predominância de conteúdos textuais e não estruturados; os interesses comercias, entre outros possíveis empecilhos.

Não queremos incorrer aqui em um "futurismo" superficial. Não temos certezas sobre qual será o futuro da Web Semântica, pois, como afirmam Kashyap et al., apenas o tempo dirá se a proposta terá sucesso ou não. Entretanto, ao mirarmos para o passado e refletirmos sobre o futuro, podemos presumir que este sistema poderá vir a apresentar processos de construção semânticas mais simples, assim como ocorreu com a própria World Wide Web, que cresceu em quantidade de conteúdos produzidos com o aparecimento de sites com sistemas publicadores de conteúdos, como os blogs e os wikis. Algumas propostas de facilitar a anotação semântica de conteúdos já existem, como no caso dos Microformats e do RDFa, que são duas formas de inserir pequenos códigos dentro do HTML, a fim de se indicar às máquinas os significados de determinadas partes do texto. Como exemplo, podemos citar um fato recente no jornalismo: a *International Press Telecommunications Council* (IPTC), influente consórcio internacional de agências de notícias e empresas jornalísticas, tais como a Agence France-Presse (AFP), a The Associated Press (AP) e o The New York Times, lançou oficialmente em outubro de 2011 uma linguagem de marcação baseada em RDFa, chamada de **rNews** (IPTC, 2011), que deverá permitir aos jornalistas estruturarem minimamente os significados presentes em seus conteúdos jornalísticos, de acordo com a lógica da Web Semântica.

Nestes últimos anos, as máquinas vêm desempenhando um papel substancial no gerenciamento da informação. Tomamos como comprovação dessa afirmativa a proliferação de sistemas estruturados em bases de dados. As máquinas se tornam ferramentas que liberam o potencial criativo do humano, pois assumem em nosso lugar as operações mecânicas e repetitivas, dignas de uma máquina. A proposta da Web Semântica é de se firmar como uma solução neste sentido: deixar para as máquinas a tarefa tediosa de buscar e organizar grandes quantidades de dados e informações, e deixar para os humanos as funções dignas de um ser racional e criativo: as de análise, reflexão e criação.

REFERÊNCIAS BIBLIOGRÁFICAS

ANDERSON, Chris. **A cauda longa**: do mercado de massa para o mercado de nicho. Rio de Janeiro: Elsevier, 2006.

AKERKAR, Rajendra. **Foundations of the Semantic Web**: XML, RDF & Ontology. Nova Déli, Índia: Narosa, 2009.

ALVES, R. C. V. **Web Semântica**: uma análise focada no uso de metadados. 2005. 180 f. Dissertação (Mestrado em Ciência da Informação) – Faculdade de Filosofia Ciências, Universidade Estadual Paulista, Marília, 2005.

ANTONIOU, Grigoris; HARMELEN, Frank van. **A Semantic Web Primer**. 1 ed. EUA: MIT Press, 2004.

AUDY, Jorge Luis Nicolas; ALEXANDRE, Gilberto Keller de Andrade e. **Fundamentos de Sistemas de Informação**. Porto Alegre: Bookman, 2005.

BARBOSA, Suzana. **"Ainda há muito o que se explorar na apuração dos bancos de dados", diz professora.** Entrevista concedida a Amanda Lopez para o blog Jornalismo Digital, *online*, 2011. Disponível em: <http://www.jornalismodigital.org/2011/08/ainda-ha-muito-o-que-se-explorar-na-apuracao-dos-bancos-de-dados-diz-professora/>. Acesso em: 12 nov. 2011.

BARBOSA, Suzana. **Jornalismo Digital em Base de Dados (JDBD) -** Um paradigma para produtos jornalísticos digitais dinâmicos. Tese de doutorado. Facom/Ufba, Salvador, 2007.

BARBOSA, Suzana. Modelo JDBD e o ciberjornalismo de quarta geração. In: **Congreso Internacional de Periodismo en la Red**, 3., Madrid: Facultad de Periodismo da Universidad Complutense de Madrid, 2008a. Disponível em: <http://grupojol.wordpress.com/2011/05/07/barbosa-2008/>. Acesso em: 12 jun. 2011.

BARBOSA, Suzana. **As bases de dados no curso da convergência jornalística**: uma análise preliminar a partir do modelo JDBD. 2008b. Disponível em: <http://grupojol.files.wordpress.com/2011/05/2008_barbosa_base_de_dados.pdf>. Acesso em: 14 jan. 2012.

BARBOSA, Suzana. Jornalismo digital e bases de dados: mapeando conceitos e funcionalidades. In: FIDALGO, A.; RAMOS, F.; OLIVEIRA, J. P.; Mealha, Ó. (Orgs.). **Livro de Actas** – 4º Congresso da Associação Portuguesa de Ciências da Comunicação (SOPCOM). 2005. Disponível em: <http://www.bocc.ubi.pt/pag/barbosa-suzana-jornalismo-digital-bases-dados.pdf>. Acesso em: 16 nov. 2011.

BERNERS-LEE, Tim. **The World Wide Web: Past, Present and Future**. W3C, 1996. Disponível em: <http://www.w3.org/People/Berners-Lee/1996/ppf.html>. Acesso em: 21 nov. 2011.

BERNERS-LEE, Tim; HENDLER, James; LASSILA, Ora. **The Semantic Web**. Scientific American Special Online Issue, abril de 2002, p. 24-30. Disponível em: <http://cms.brookes.ac.uk/modules/notes/112_SemWeb.pdf>. Acesso em: 28 nov. 2010.

BERNERS-LEE, Tim. **Linked Data**. Design Issues, W3C. 2006. Disponível em: <http://www.w3.org/DesignIssues/LinkedData.html>. Acesso em: 11 jan. 2012.

BERNERS-LEE, Tim. **Giant Global Graph**. Decentralized Information Group (DIG), 2007. Disponível em: <http://dig.csail.mit.edu/breadcrumbs/node/215>. Acesso em: 21 nov. 2011.

BERNERS-LEE. **Tim Berners-Lee on the next Web.** Palestra no TED, 2009. Disponível em: <http://www.ted.com/talks/tim_berners_lee_on_the_next_web.html>. Acesso em: 29 nov. 2011.

BERTOCCHI, Daniela. **Ciberjornalismo e Web Semântica**: Considerações sobre o uso de tags em narrativas jornalísticas digitais. In: 7o. SBPJor - Encontro Nacional de Pesquisadores em Jornalismo, 2009, São Paulo. Anais do 7o. SBPJor - Encontro Nacional de Pesquisadores em Jornalismo.

BERTOCCHI, Daniela. **Narrativas jornalísticas no contexto da web semântica**. 2010. Anais do II Seminário de Ciberjornalismo do Mato Grosso do Sul.

BIZER, Christian; HEATH, Tom; BERNERS-LEE, Tim. **Linked Data - The Story So Far**. 2009. Disponível em: <http://tomheath.com/papers/bizer-heath-berners-lee-ijswis-linked-data.pdf>. Acesso em: 12 jan. 2012.

BRANDÃO, Anarosa Alves Franco; LUCENA, Carlos José Pereira de. **Uma Introdução à Engenharia de Ontologias no contexto da Web Semântica**. PUC-Rio. 2002.

BRADSHAW, Paul. **How to be a data journalist**. Datablog, 2010. Disponível em: <http://www.guardian.co.uk/news/datablog/2010/oct/01/data-journalism-how-to-guide>. Acesso em: 21 nov. 2011.

BREITMAN, Karin. **Web Semântica**: A internet do futuro. Rio de Janeiro: LTC, 2005.

BUENO, Francisco da Silveira. **Minidicionário da língua portuguesa**. São Paulo: FTD: LISA, 1996.

CAIRO, Alberto. **Interactividad en infografía de prensa.** Artigo publicado no Malofiej 15. University of North Caroline, 2008. Disponível em: <http://www.albertocairo.com/imagenes/2008/articulos/articulomalofiej.pdf>. Acesso em: 18 nov. 2011.

CANAVILHAS, João. **Webjornalismo:** Considerações gerais sobre jornalismo na web. Comunicação apresentada no I Congresso Ibérico de Comunicação. Universidade da Beira Interior - Portugal. 2001. Disponível em: <http://www.bocc.ubi.pt/pag/_texto.php?html2=canavilhas-joao-webjornal.html>. Acesso em: 09 nov. 2011.

CANTAIS, Jaime; DOMINGUEZ, David; GIGANTE, Valeria; LAERA, Loredana; TAMMA, Valentina. An example of food ontology for diabetes control. In: WELTY, C.; GANGEMI, A. "**Working notes of the ISWC 2005 workshop on Ontology Patterns for the Semantic Web**", Galway, Irlanda, 2005. Disponível em: <http://www.csc.liv.ac.uk/~floriana/PIPS/papers/FoodOntology.pdf>. Acesso em: 9 out. 2011.

CECCONI, Carlos. **W3C, o futuro da Web, HTML5**. Palestra. 2010. Disponível em: <http://www.youtube.com/watch?v=aeubheKRqj8>. Acesso em: 10 abr. 2011.

CHANG, F.; DEAN, J.; GHEMAWAT, S.; HSIEH, W.; WALLACH, D.; BURROWS, M.; CHANDRA, T.; FIKES, A.; GRUBER, R. **Bigtable:** A Distributed Storage System for Structured Data. Disponível em: <http://static.googleusercontent.com/external_content/untrusted_dlcp/research.google.com/en//archive/bigtable-osdi06.pdf>. Acesso em: 7 fev. 2012.

CODINA, Lluís. **Web 2.0, 3.0 y Web Semántica: Impacto en los sistemas de información**. 2011. Disponível em: <http://www.lluiscodina.com/>. Acesso em: dez. 2011.

DANTAS, Mario. **Tecnologias de Redes de Comunicação e Computadores**. Rio de Janeiro: Axcel Books, 2002.

DIAZ NOCI, Javier; SALAVERRÍA, Ramón. **Manual de Redación Ciberperiodística.** Barcelona: Ariel, 2003.

DIMITROV, Marin. **Metadata management for the BBC's 2010 World Cup site using OWLIM**. Apresentação no European Semantic Technology Conference 2010. Video Lectures, 2010. Disponível em: <http://videolectures.net/estc2010_dimitrov_utopwc/>. Acesso em: 27 jan. 2012.

DODDS, Leigh; SCOTT, Tom. **Wildlife Ontology**. BBC, 2010. Disponível em: <http://www.bbc.co.uk/ontologies/wildlife/2010-11-04.shtml>. Acesso em: 22 jan. 2012.

DONG, L.; SMITH, R.; BUCHANAN, Bruce. **NewsFinder: Automating na Artificial Intelligence News Service**. Artigo apresentado na Twenty-Third IAAI Conference, 2011. Disponível em: <http://www.aaai.org/ocs/index.php/IAAI/IAAI-11/paper/view/3446>. Acesso em: 12 fev. 2012.

ESCOBAR, Maurício; LEMKE, Ana Paula; RIBEIRO, Marcelo Blois. **SemantiCore 2006 – Permitindo o Desenvolvimento de Aplicações baseadas em Agentes na Web Semântica**. Estudo desenvolvido pelo Intelligent Systems Engineering Group da PUCRS, financiado pela Dell Computadores do Brasil Ltda. 2006. Disponível em: <http://www.les.inf.puc-rio.br/seas2006/papers/X072.pdf>. Acessado em: 20 mar. 2011.

FARBIAZ, A.; BARBOSA, Suzana. A estética base de dados e os modos diferenciados para visualização da informação jornalística. In: **III Simpósio Nacional da ABCiber - Associação Brasileira de Pesquisadores em Cibercultura**. São Paulo: ESPM, 2009. Disponível em: <http://www.abciber.com.br/simposio2009/trabalhos/anais/pdf/artigos/5_jornalismo/eixo5_art1.pdf >. Acesso em: 12 nov 2011.

FIDALGO, António. Sintaxe e Semântica das Notícias Online: Para um Jornalismo Assente em Base de Dados. In.: LEMOS, A. L. M. (Org.); SILVA, J. M. (Org.); SÁ, S. P. (Org.); PRYSTON, A. (Org.). **Mídia.br**. Livro da XII Compós - 2003. Porto Alegre: Sulina, 2004.

GARTNER. **Gartner Says Sales of Mobile Devices Grew 5.6 Percent in Third Quarter of 2011; Smartphone Sales Increased 42 Percent**. Disponível em: <http://www.gartner.com/it/page.jsp?id=1848514>. Acesso em: 18 nov 2011.

GRUBER, Thomas R. A Translation Approach to Portable Ontology Specifications. **Knowledge Acquisition**. V. 5, n. 2, 1993, p. 199-220. Disponível em: <http://tomgruber.org/writing/ontolingua-kaj-1993.pdf>. Acesso em: 10 out. 2011.

HEBELER, John; FISHER, Matthew; Ryan, Blace; PEREZ-LOPEZ, Andrew; DEAN, Mike. **Semantic Web Programming**. Wiley Publishing: Indianapolis (EUA), 2009.

JOHNSON, Steven. **Cultura da interface**: como o computador transforma nossa maneira de criar e comunicar. Rio de Janeiro: Jorge Zahar, 2001.

KASHYAP, Vipul; BUSSLER, Christoph; MORAN, Matthew. **Semantic Web:** Semantics for Data and Services on the Web. Berlin, Heidelberg: Springer-Verlag Berlin Heidelberg, 2008.

KIRYAKOV, A.; BISHOP, B.; OGNYANOFF, D.; PEIKOV, I.; TASHEV, Z.; VELKOV, R. **The Features of BigOWLIM that Enabled the BBC's World Cup Website**. Workshop em Semantic Data Management SemData@VLDB. 17 de set. de 2010. Disponível em: <http://ceur-ws.org/Vol-637/paper6.pdf>. Acesso em: 27 jan. 2012.

LAMMEL, Iuri. **Padrão HTML5**: possíveis efeitos no Jornalismo Digital. Anais do XXXIII Congresso Brasileiro de Ciências da Comunicação (Intercom). Set. de 2010. Disponível em: <http://www.intercom.org.br/papers/nacionais/2010/resumos/R5-2173-1.pdf>. Acesso em: 07 fev. 2012.

LARRONDO, Ainarra; MIELNICZUK, Luciana; BARBOSA, Suzana. **Narrativa jornalística e base de dados: discussão preliminar sobre gêneros textuais no ciberjornalismo de quarta geração**. Artigo apresentado no VI Encontro Nacional de Pesquisadores em Jornalismo. São Paulo, 2008. Disponível em: <http://sbpjor.kamotini.kinghost.net/sbpjor/admjor/arquivos/coordenada8lucianamielniczuk.pdf >. Acesso em: 11 nov. 2011.

LEÃO, Lucia. **O Labirinto da Hipermídia:** arquitetura e navegação no ciberespaço. São Paulo: Iluminuras, 2001.

LEMOS, André; LÉVY, Pierre. **O futuro da internet**: Em direção a uma ciberdemocracia planetária. São Paulo: Paulus, 2010.

LOUKIDES, Mike. Data Science and data tools. In.: O'Reilly Radar Team. **Big Data Now**: Current Perspectives from O'Reilly Radar. E-book, edição para Kindle. EUA: O'Reilly, 2012.

MACHADO, Elias. **O ciberespaço como fonte para os jornalistas.** 2002. Disponível em: <http://www.bocc.ubi.pt/pag/machado-elias-ciberespaco-jornalistas.pdf>. Acesso em: 5 jun. 2011.

MACHADO, Elias. **O jornalismo digital em base de dados**. Florianópolis: Calandra, 2006.

MACHADO, E.; PALACIOS, M. Um modelo híbrido de pesquisa: a metodologia aplicada pelo GJOL. In.: Lago, Claudia e Benetti, Marcia. (Org.). **Metodologia de pesquisa em jornalismo**. Petrópolis: Vozes, 2007, p. 199-222.

MANOVICH, Lev. **Metadata, Mon Amour**. 2002. Disponível em: <http://www.manovich.net/TEXTS_07.HTM>. Acessado em: 15 jul. 2010.

MANOVICH, Lev. **Software takes command**. 2008. Disponível em: <http://softwarestudies.com/softbook/manovich_softbook_11_20_2008.pdf>. Acesso em: 18 nov. 2011.

MANOVICH, Lev. **The Language of New Media**. 2001. Disponível em: <http://ucsd.academia.edu/LevManovich/Papers>. Acesso em: 14 nov. 2011.

MARTINS, Gilberto de Andrade. **Estudo de Caso**: uma estratégia de pesquisa. São Paulo: Atlas, 2006.

MIELNICZUK, Luciana. **Considerações sobre interatividade no contexto das novas mídias**. 2001. Disponível em: <http://www.facom.ufba.br/jol/pdf/2001_mielniczuck_linkparatextual.pdf>. Acesso em: 12 out. 2005.

MIELNICZUK, Luciana. **Jornalismo na Web**: uma contribuição para o estudo do formato da notícia na escrita hipertextual. Tese de doutorado Facom/Ufba, Salvador, 2003.

MOREIRA, Carla Barbosa. **Princípio de ligação Sintaxe/Semântica:** Construções estativas. Dissertação (mestrado) apresentada ao Programa de Pós-Graduação em Letras da Universidade Federal de Minas Gerais. Belo Horizonte, 2000.

NOCI, Javier Diaz (Org.) ; PALACIOS, Marcos (Org.) . **Online journalism: research methods.** A multidisciplinary approach in comparative perspective. Bilbao: Servicio Editorial de la Universidad del País Vasco., 2009. Disponível em: <http://www.argitalpenak.ehu.es/p291-content/es/contenidos/libro/se_indice_ciencinfo/es_ciencinf/adjuntos/journalism.pdf>. Acesso em: 03 dez 2011.

O'DONOVAN, J. **The World Cup and a call to action around Linked Data**. BBC Blogs, 2010. Disponível em: <http://www.bbc.co.uk/blogs/bbcinternet/2010/07/the_world_cup_and_a_call_to_ac.html>. Acesso em: 26 jan. 2012.

OLAVSRUD, Thor. **Berners-Lee Talks Up Semantic Web**. InternetNews.com. Disponível em: <http://www.internetnews.com/dev-news/article.php/3081191>. Acesso em: 07 fev 2012.

OLIVER, Silver. **News Rewired**. Youtube, 2010a. Disponível em: <http://www.youtube.com/watch?v=bY5kONXROCY>. Acesso em: 14 dez. 2011.

OLIVER, Silver. **How the emergence of the semantic web changes our approach to information architecture**. SlideShare, 2010b. Disponível em: <http://www.slideshare.net/silveroliver/how-the-emergence-of-the-semantic-web-changes-our-approach-to-information-architecture>. Acesso em: 11 jan. 2012.

OLIVER, Silver. **Mining the oil shale of journalism with semantic web technologies**. 2011. Disponível em: <http://blockslabpillar.com/2011/02/20/mining-the-oil-shale-of-journalism-with-semantic-web-technologies/>. Acesso em: 14 jan. 2012.

OLIVIERO, Carlos A. J. **Faça um aplicativo: Banco de dados cliente/servidor com Delphi 6 – Orientado a projeto.** São Paulo: Érica, 2002.

OPEN CALAIS. **About**. Disponível em: <http://www.opencalais.com/about>. Acessado em: 29 set. 2011.

PALACIOS, Marcos. **Jornalismo online, informação e memória:** apontamentos para o debate. (2002b). Disponível em: <http://www.facom.uba.br/jol/pdf/2002_palacios_informacaomemoria.pdf>. Acesso em: 08 out. 2006.

PALACIOS, Marcos. Ruptura, Continuidade e Potencialização no Jornalismo Online: o Lugar da Memória. In: MACHADO, Elias & PALACIOS, Marcos (Orgs). **Modelos do Jornalismo Digital**, Salvador: Calandra, 2003.

PALACIOS, Marcos; MIELNICZUK, Luciana; BARBOSA, Suzana; RIBAS, Beatriz; NARITA, Sandra. **Um mapeamento de características e tendências no jornalismo online brasileiro e português**. Trabalho apresentado no XXV Intercom. Salvador, 2002.

PAUL, Christiane. **The Database as System and Cultural Form**: Anatomies of Cultural Narratives. *Online*. Disponível em: <www.cityarts.com/paulc/RISD/Paul_Database.doc>. Acesso em: 12 nov. 2011.

PAVLIK, J. **The Impact of Technology on Journalism**. Journalism Studies, V. 1, Nº 2, 2000, p. 229–237. Disponível em: <http://www.tandfonline.com/doi/abs/10.1080/14616700050028226>. Acesso em: 8 fev. 2012.

PIETROFORTE; LOPES. Semântica Lexical. In: FIORIN, José Luiz (org.). **Introdução à Linguística.** São Paulo: Contexto, 2003, p. 114

PRIMO, Alex. **Quão interativo é o hipertexto? :** Da interface potencial à escrita coletiva. Fronteiras: Estudos Midiáticos, São Leopoldo, v. 5, n. 2, p. 125-142, 2003.

RAMALHO, Rogério Aparecido Sá. **Web Semântica**: aspectos interdisciplinares da gestão de recursos informacionais no âmbito da Ciência da Informação. Dissertação (Mestrado em Ciência da Informação) – Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, 2006.

RAIMOND, Yves; SCOTT, Tom; OLIVER, Silver; SINCLAIR, Patrick; SMETHURST, Michael. Use of Semantic Web technologies on the BBC Web Sites. In.: WOOD, David (ORG.). **Linking Enterprise Data**. EUA: Springer, 2010a.

RAIMOND, Yves; SCOTT, Tom; SINCLAIR, Patrick; MILLER, Libby; BETTS, Stephen; Mcnamara, Frances. Case Study: Use of Semantic Web Technologies on the BBC Web Sites. In.: W3C. **Semantic Web Use Cases and Case Studies**. 2010b. Disponível em: <http://www.w3.org/2001/sw/sweo/public/UseCases/BBC/>. Acesso em: 17 jan. 2011.

RAYFIELD, J. **BBC World Cup 2010 dynamic semantic publishing**. 2010. Disponível em: <http://www.bbc.co.uk/blogs/bbcinternet/2010/07/bbc_world_cup_2010_dynamic_ sem.htm>. Acesso em: 26 jan. 2012.

RAYFIELD, J. **BBC Dynamic Semantic Publishing [DSP].** 2012. Disponível em: <http://www.slideshare.net/JemRayfield/dsp-bbcjem-rayfieldsemtech2011>. Acesso em: 11 jan. 2012.

REESE, George. **Database Programming with JDBC and Java**. 2. ed. EUA: O'Reilly, 2000.

RIBAS, Beatriz. **Características da notícia na Web - considerações sobre modelos narrativos**. (Comunicação individual). II Encontro Nacional de Pesquisadores em Jornalismo - SBPJor, 2004. Disponível em: <http://www.facom.ufba.br/jol/pdf/2004_ribas_caracteristicas_noticia_web.pdf>. Acesso em: 13 nov. 2011.

RIBAS, Beatriz. **Web Semântica e produção de notícias:** anotações para o estudo da aplicação da tecnologia ao campo do Jornalismo. 5º Encontro Nacional de Pesquisadores em Jornalismo - SBPJor. 2007.Disponível em: <http://sbpjor.kamotini.kinghost.net/sbpjor/admjor/arquivos/coordenada_8_._beatriz_ribas.pdf>. Acessado em: 27 jun. 2010.

RODRIGUES, Adriana Alves. **Infografia interativa em base de dados no jornalismo digital**. Dissertação de mestrado. Universidade Federal da Bahia, Salvador, 2009.

ROGERS, Simon. **Facts are sacred: the power of data**. E-book, edição Kindle. _____: Guardian Books, 2011.

SALAVERRÍA, Ramon. **Redacción periodística en internet.** Barcelona: EUNSA, 2005.

SCHWINGEL, C. Ferramentas de publicação de conteúdos na internet no contexto do ciberjornalismo. In: CD ROM do **XI Encontro de Professores de Jornalismo**. São Paulo, 2008. Disponível em: <http://www.facom.ufba.br/jol/pdf/Schwingel_2008_ENPJ.pdf>. Acesso em: 11 nov. 2011.

SCHWINGEL, C. **Jornalismo Digital de Quarta Geração:** a emergência de sistemas automatizados para o processo de produção industrial no Jornalismo Digital. In: Compós, 2005, Niterói. CD-ROM Compós, 2005. Disponível em: <http://www.facom.ufba.br/jol/pdf/Schwingel_2005_Compos.pdf>. Acesso em: 14 nov. 2011.

SCHWINGEL, Carla. **A produção de conteúdos no ciberespaço:** sistemas de gerenciamento de conteúdos. Artigo apresentado no VII Encontro Nacional de Pesquisadores em Jornalismo. São Paulo, 2009.

SCHWINGEL, Carla**. Os sistemas de publicação como fator da terceira fase do Jornalismo Digital**. 2004. Disponível em: <http://www.facom.ufba.br/jol/pdf/2004_schwingel_sistemas_publicacao.PDF>. Acessado em: 27 jun. 2010.

SCOTT, Tom. **Opening up the BBC's natural history archive**. Blog Derivadow.com, 2009. Disponível em: <http://derivadow.com/2009/07/28/opening-up-the-bbcs-natural-history-archive/>. Acesso em: 13 dez. 2011.

SCOTT, Tom. **Apis and APIS a wildlife ontology**. Blog Derivadow, 2010. Disponível em: <http://derivadow.com/2010/03/02/apis-and-apis-a-wildlife-ontology/>. Acesso em: 13 dez. 2011.

SCOTT, Tom. **One BBC nature**. Blog Derivadow, 2011. Disponível em: <http://derivadow.com/2011/05/13/one-bbc-nature/>. Acesso em: 13 dez. 2011.

SHADBOLT, N.; BERNERS-LEE, T.; HALL, W. **The Semantic Web Revisited**. IEEE Intelligent Systems, vol. 21, n. 3, maio/junho de 2006, p. 96-101. Disponível em: <http://eprints.ecs.soton.ac.uk/12614/1/Semantic_Web_Revisted.pdf>. Acesso em: 26 jan 2012.

SEGARAN, Toby; EVANS, Colin; TAYLOR, Jamie. **Programming the Semantic Web**. EUA: O'Reilly Media, 2009.

SIEGEL, David. **Pull**: The Power of the Semantic Web to Transform Your Business. EUA: Portfolio, 2009.

SINCLAIR, Patrick. **Linked Data on the BBC**. 2009. Disponível em: <http://www.slideshare.net/metade/linked-data-on-the-bbc>. Acesso em: 9 fev. 2012.

SILVA FILHO, Antonio Mendes da. **Programando com XML**. Rio de Janeiro: Elsevier, 2004.

SOUZA, Renato Rocha; ALVARENGA, Lídia. **A Web Semântica e suas contribuições para a ciência da informação.** Ci. Inf., Brasília, v. 33, n. 1, abril de 2004. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-19652004000100016&lng=en&nrm=iso>. Acesso em: 27 mar. 2011.

TAKAI, O.; ITALIANO, I.; FERREIRA, J. **Introdução a Banco de Dados.** [Online]. Disponível em: <http://www.ime.usp.br/~jef/apostila.pdf>. Acesso em: 15 nov. 2011.

THIS WE KNOW. **About.** [Página da web]. *Online*. Disponível em: <http://www.thisweknow.org/about>. Acesso em: 18 set. 2011.

TUCHMAN, Gaye. **Making news: a study in the construction of reality**. Michigan: Free Press, 1978.

VIÉGAS, Fernanda. **Journalism in the Age of Data.** Entrevista concedida a Geoff McGhee em documentário online sobre o jornalismo na era dos dados, *online*, 2010. Disponível em: <http://datajournalism.stanford.edu/>. Acesso em: 10 nov. 2011.

WANGLON, Paolla. **Aplicativos jornalísticos em mídias móveis**: o formato para *smartphones*. Monografia de graduação. Universidade Federal de Santa Maria, Santa Maria/RS, 2010.

W3C. **HTML 4.01 Specification**. 1999. Disponível em: <http://www.w3.org/TR/1999/REC-html401-19991224/>. Acesso em: 20 nov. 2011.

W3C. **W3C Semantic Web Frequently Asked Questions**. 2001a. Disponível em: <http://www.w3.org/2001/sw/SW-FAQ>. Acessado em: 22 jun. 2010.

W3C. **W3C Semantic Web Activity**. 2001b. Disponível em: <http://www.w3.org/2001/sw/>. Acesso em: 28 nov 2010.

W3C. **RDF Primer**. W3C Recommendation. 2004a. Disponível em: <http://www.w3.org/TR/rdf-primer/>. Acesso em: 20 nov. 2011.

W3C. **RDF Vocabulary Description Language 1.0: RDF Schema**. W3C Recommendation. 2004b. Disponível em: <http://www.w3.org/TR/rdf-schema/>. Acesso em: 20 jan. 2012.

W3C. **OWL Web Ontology Language Guide**. W3C Recommendation. 2004c. Disponível em: <http://www.w3.org/TR/owl-guide/>. Acesso em: 21 jan. 2012.

APÊNDICE A – Roteiro para observação e análise dos casos estudados

PARTE I – IDENTIFICAÇÃO E DESCRIÇÃO DO PRODUTO

1. Identificação do produto

1.1. Nome:
1.2. URL:
1.3. Empresa/Instituição/Organização:
1.4. Localidade de origem:
1.5. Data da pesquisa:

2. Descrição do produto observado (especificidades):

- Tipo/função (é um portal? É uma reportagem? É uma infografia? etc.)
- Forma (arquitetura da informação, interface, navegação etc.)
- Conteúdo (tipo de conteúdo, formato da narrativa jornalística etc.)

PARTE II – ANÁLISE DO FUNCIONAMENTO E DAS VANTAGENS DAS TECNOLOGIAS SEMÂNTICAS

1. Contexto e justificativa para uso das tecnologias semânticas

2. Identificação de recursos e tecnologias semânticas utilizadas

- Utiliza modelo de dados para descrição em triplas? Ex.: RDF, serializações do RDF (RDF/XML, Notation-3 (N3), Turtle, N-Triples, RDFa, RDF/JSON)
- Utiliza metadados compartilhados? Ex.: Dublin Core, FOAF, CC etc.
- Utiliza ontologias?
- Utiliza técnica de *tagging* ou *software* para extração de conceitos?
- Acessa dados estruturados de *datasets* compartilhados? Ex.: Freebase, DBpedia.

3. Descrição do funcionamento das tecnologias semânticas

4. Identificação das vantagens do uso das tecnologias semânticas identificadas

Obs.: analisar como a semântica influencia/altera as categorias do JDBD:

- Dinamicidade
- Automatização
- Inter-relacionamento/hiperlinkagem
- Flexibilidade
- Densidade informativa
- Diversidade temática
- Visualização
- Convergência

APÊNDICE B – Lista de fonte para análise do site BBC World Cup 2010

Abaixo, listamos os principas profissionais que serviram como fontes de dados secundários para a identificação e descrição das tecnologias semânticas no caso BBC World Cup 2010.

Jem Rayfield, arquiteto técnico sênior do departamento BBC Future Media & Technology. Rayfield participou diretamente no desenvolvimento da solução semântica para o site BBC World Cup 2010. Os dados foram coletados de duas fontes: de um depoimento seu sobre este desenvolvimento, publicado em um blog da própria BBC (RAYFIELD, 2010), e de uma apresentação em slides disponibilizada pelo próprio Rayfield (2011), em que apresenta breves informações sobre a semântica em quatro diferentes projetos da BBC. Rayfield contribui com informações detalhadas sobre o funcionamento das tecnologias.

John O'Donovan, arquiteto técnico chefe do departamento BBC Future Media & Technology. O'Donovan participou diretamente no desenvolvimento da solução semântica para o site BBC World Cup 2010. Os dados foram coletados de duas fontes: de um depoimento seu sobre este desenvolvimento, publicado em um blog da própria BBC (O'DONOVAN, 2010). Suas contribuições se dão mais na parte conceitual do sistema.

Silver Oliver, arquiteto da informação da BBC. Oliver participou no desenvolvimento de diferentes iniciativas da BBC além do World Cup 2010. Seus dados foram coletados de uma apresentação, gravada em vídeo, em que apresenta questões técnicas relacionadas à plataforma de publicação semântica da BBC, onde comenta sobre o projeto World Cup 2010 (OLIVER, 2010a). Também foi encontrada uma apresentação sua sobre como a emergência da Web Semântica modifica práticas na arquitetura da informação, em que também cita informações úteis sobre o projeto BBC World Cup 2010 (OLIVER, 2010b).

Marin Dimitrov, desenvolvedor da empresa Ontotex, a responsável pelo triple store BigOWLIN (utilizado pelo site da BBC). Embora seja funcionário da empresa Ontotex, Dimitrov colaborou na implantação do triple store junto aos desenvolvedores das BBC. As contribuições de Dimitrov foram obtidas de uma apresentação de trabalho na European Technology Conference 2010, gravada em vídeo e disponibilizada na web (DIMITROV, 2010).

APÊNDICE C – Lista de fonte para análise do site BBC Wildlife

Abaixo, listamos os principais profissionais que serviram como fontes de dados secundários para a identificação e descrição das tecnologias semânticas no caso BBC World Cup 2010.

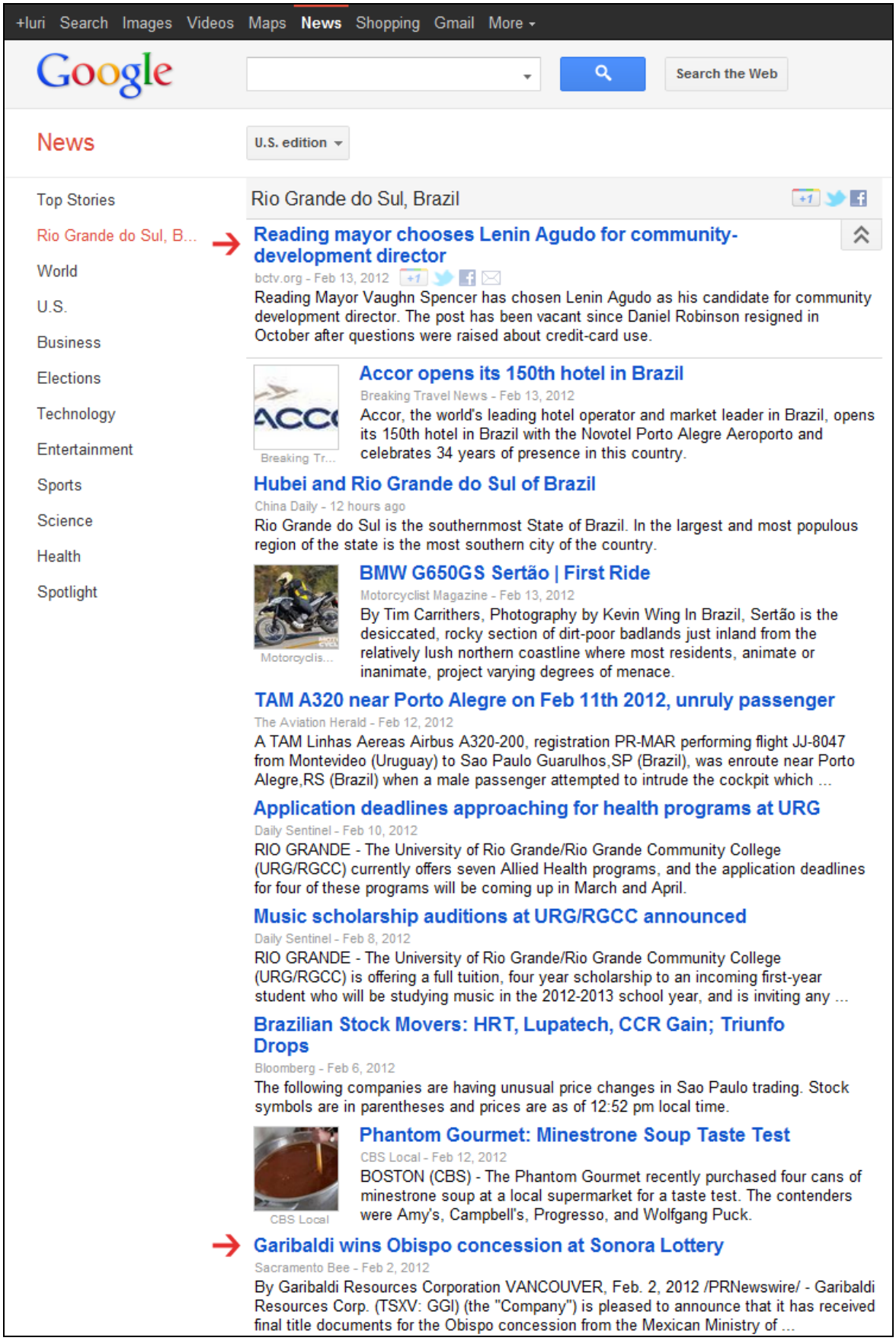
Yves Raimond, tecnologista sênior da equipe de Pesquisa e Desenvolvimento da BBC. Raimond foi um dos responsáveis pelo desenvolvimento e manutenção do site BBC Programmes, um repositório semântico dos programas da BBC. Encontramos um artigo em que ele e outros desenvolvedores da BBC explanam sobre alguns dos projetos semânticos da BBC (RAIMOND et al., 2010a). Também escreveu um breve relato sobre o projeto Wildlife no site da W3C, junto com outros profissionais da BBC (RAIMOND et al., 2010b).

Silver Oliver, arquiteto da informação da BBC. Oliver participou no desenvolvimento de diferentes iniciativas da BBC além do Wildlife. Seus dados foram coletados de uma apresentação, gravada em vídeo, em que apresenta questões técnicas relacionadas à plataforma de publicação semântica da BBC, onde comenta sobre o projeto Wildlife (OLIVER, 2010a). Também foi encontrada uma apresentação sua sobre como a emergência da Web Semântica modifica práticas na arquitetura da informação, em que também cita informações úteis sobre o projeto BBC Wildlife (OLIVER, 2010b).

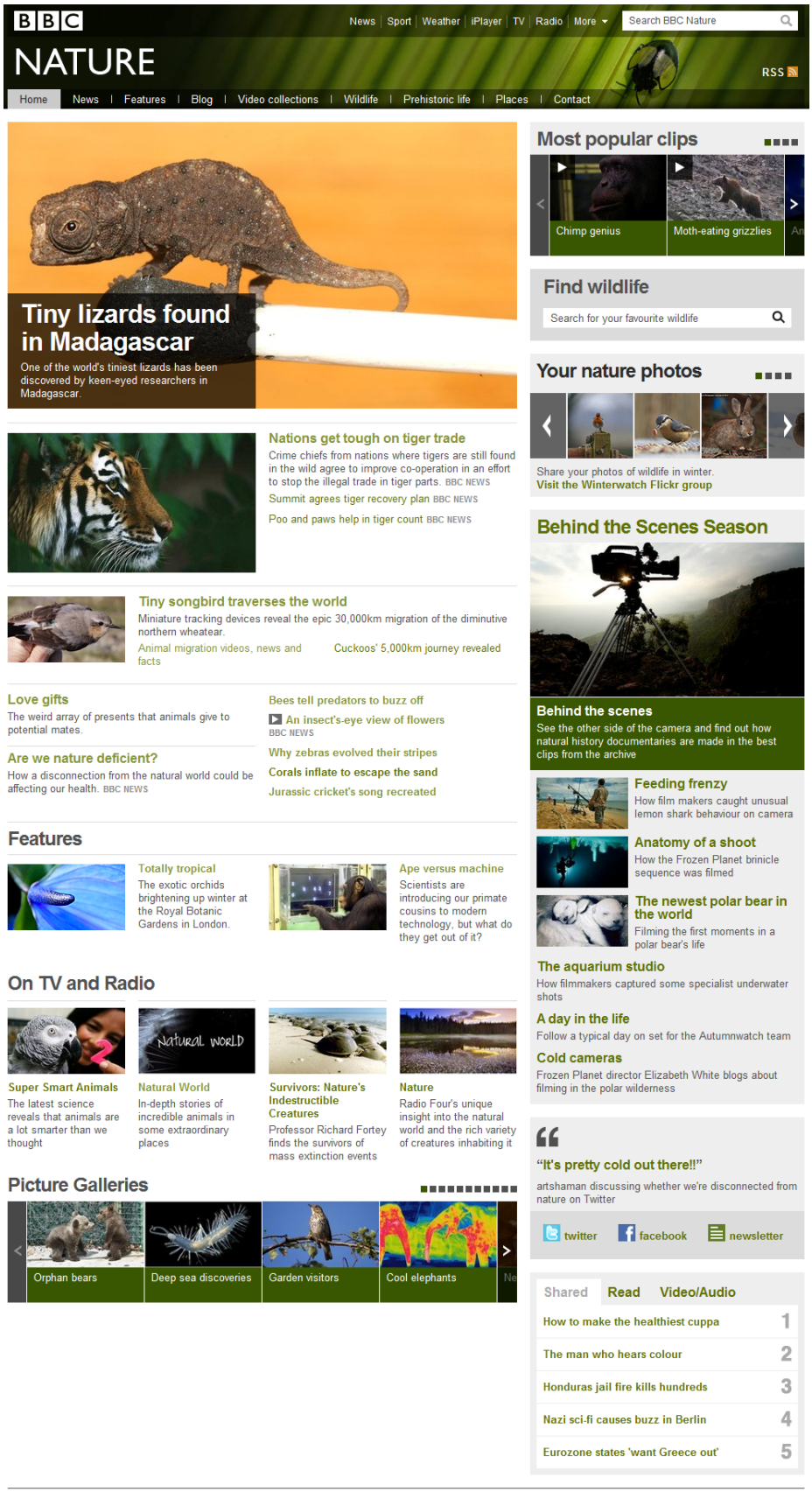
Tom Scott, que até 2011 era profissional da área de tecnologia da BBC, mas atualmente é responsável por projeto semânticos na Nature.com. Scott foi um dos dois autores da ontologia do Wildlife, junto com **Leigh Dodds**. Além de suas colaborações nas informações presentes na documentação oficial da ontologia, uma outra fonte de dados foi o seu blog oficial. Nos primeiros anos logo após o lançamento do Wildlife (entre 2009 e 2011), Scott publicou diversos posts em que esclarecia várias questões relacionadas ao desenvolvimento do site (SCOTT, 2009, 2010, 2011).

Patrick Sinclair, web developer e trabalha como engenheiro de software da BBC. Trabalhou em projetos da BBC relacionados a tecnologias semânticas, como o BBC Music. Sinclair foi o autor de uma apresentação sobre projetos da BBC no âmbito da Web Semântica, apresentado em evento no Brasil e disponibilizado na web (SINCLAIR, 2009).

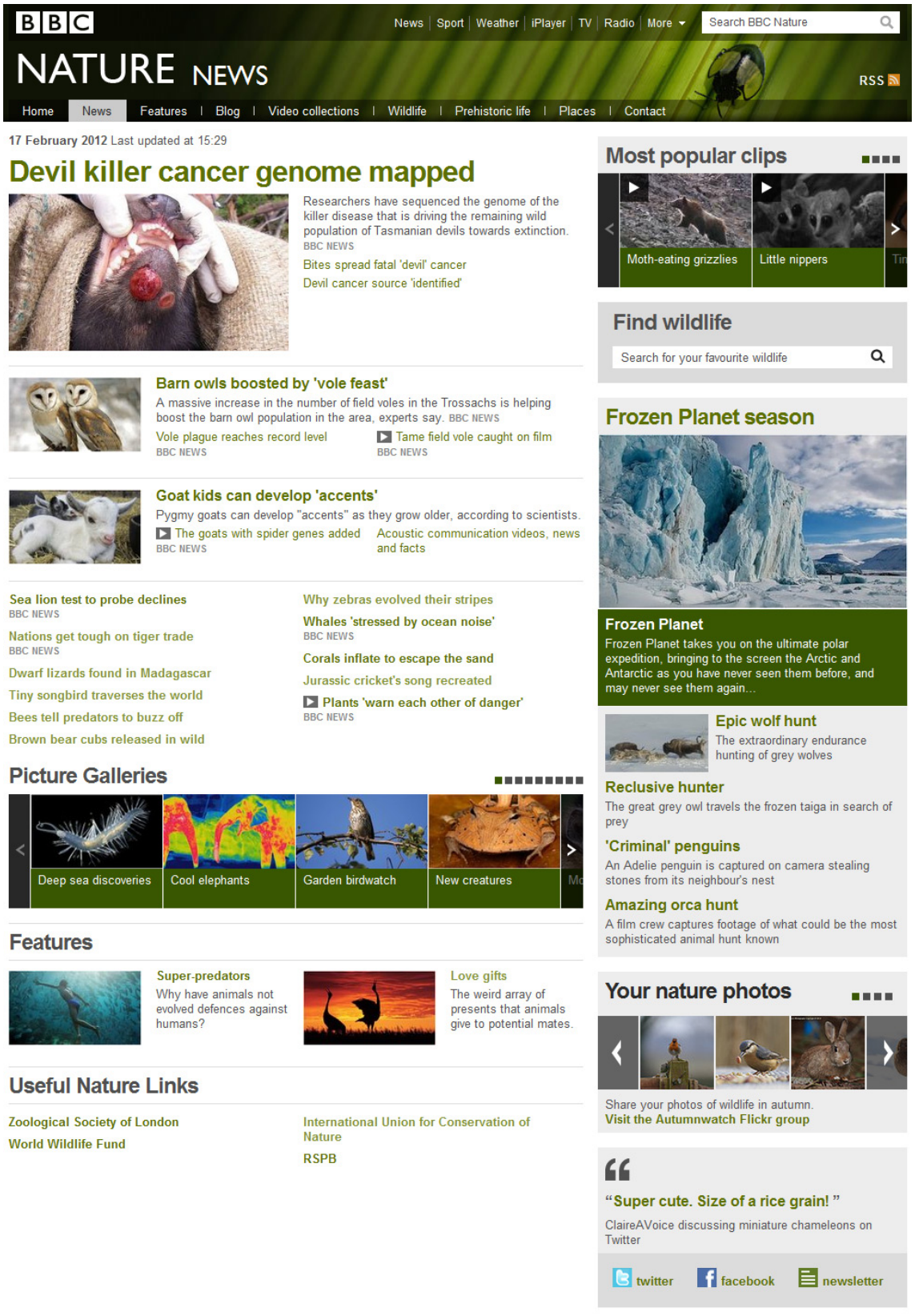
ANEXO A – Tela da página do Google News



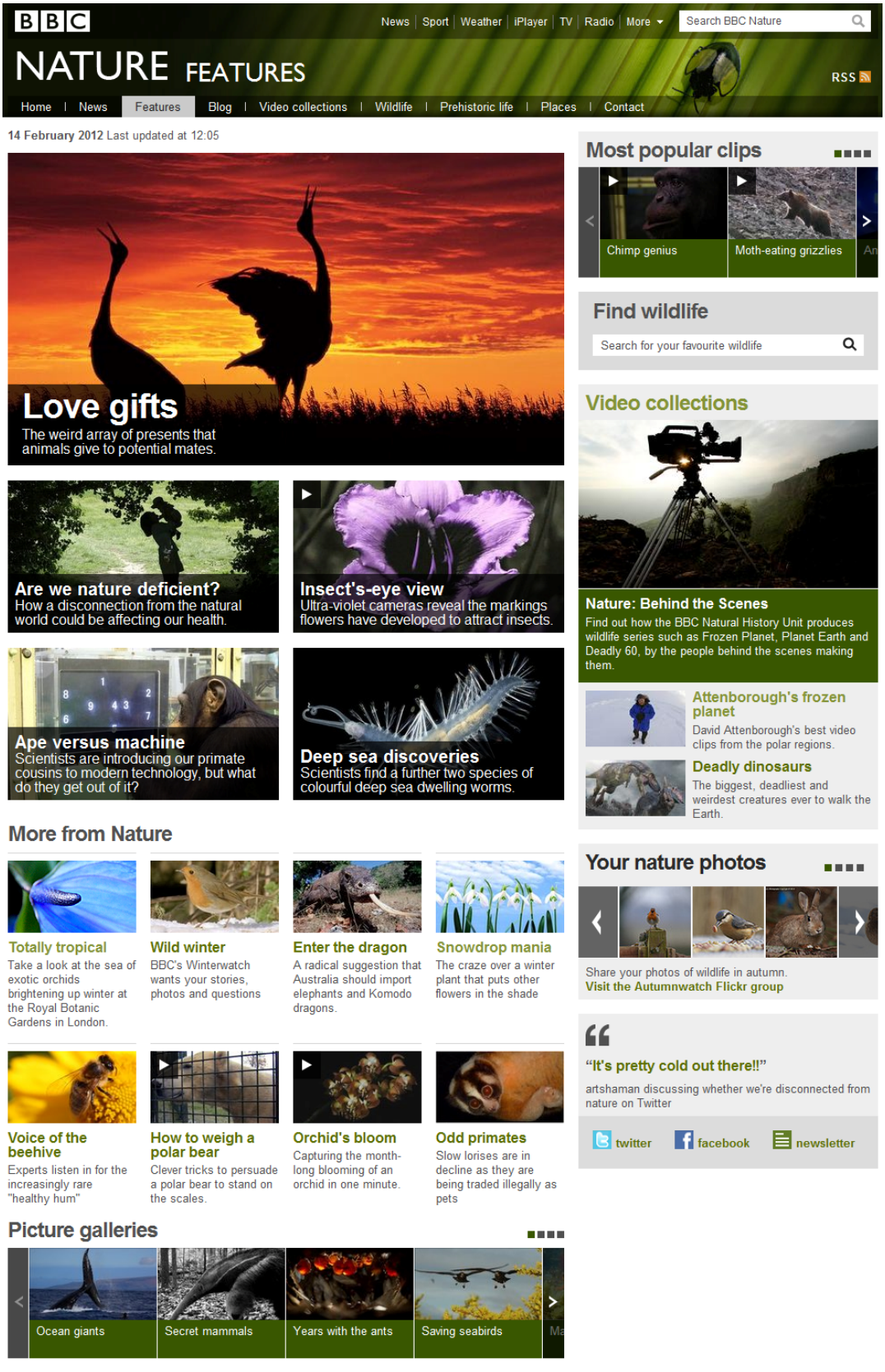
ANEXO B – Tela inicial da seção *Home* do site BBC Nature



ANEXO C – Tela inicial da seção *News* do site BBC Nature



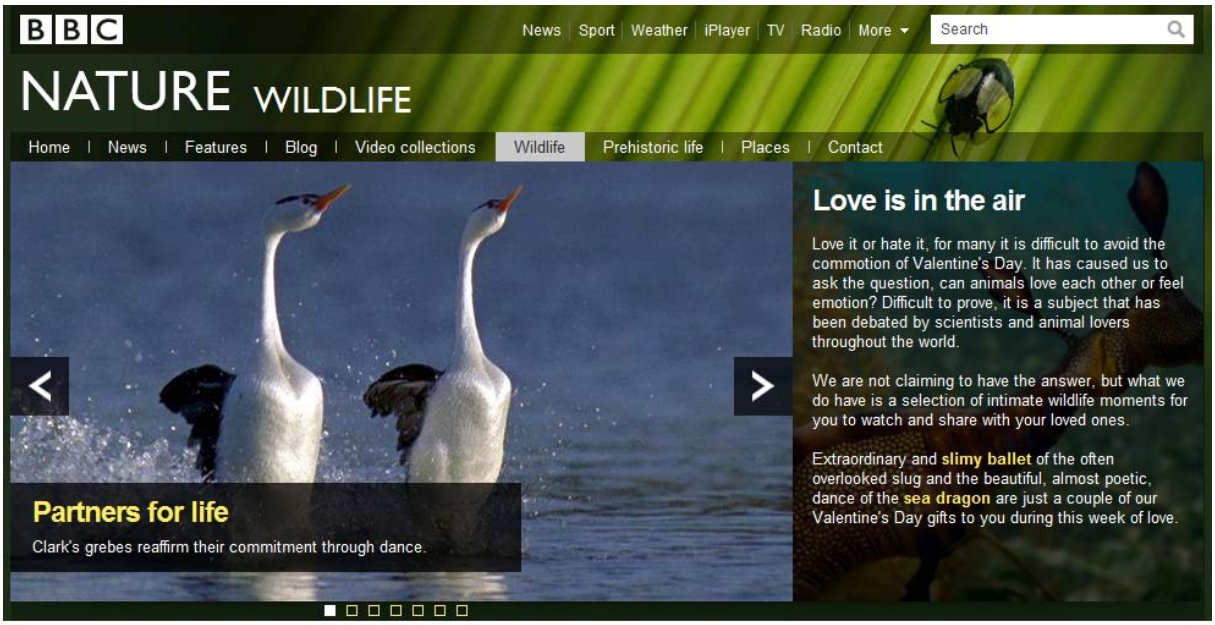
ANEXO D – Tela inicial da seção *Features* do site BBC Nature



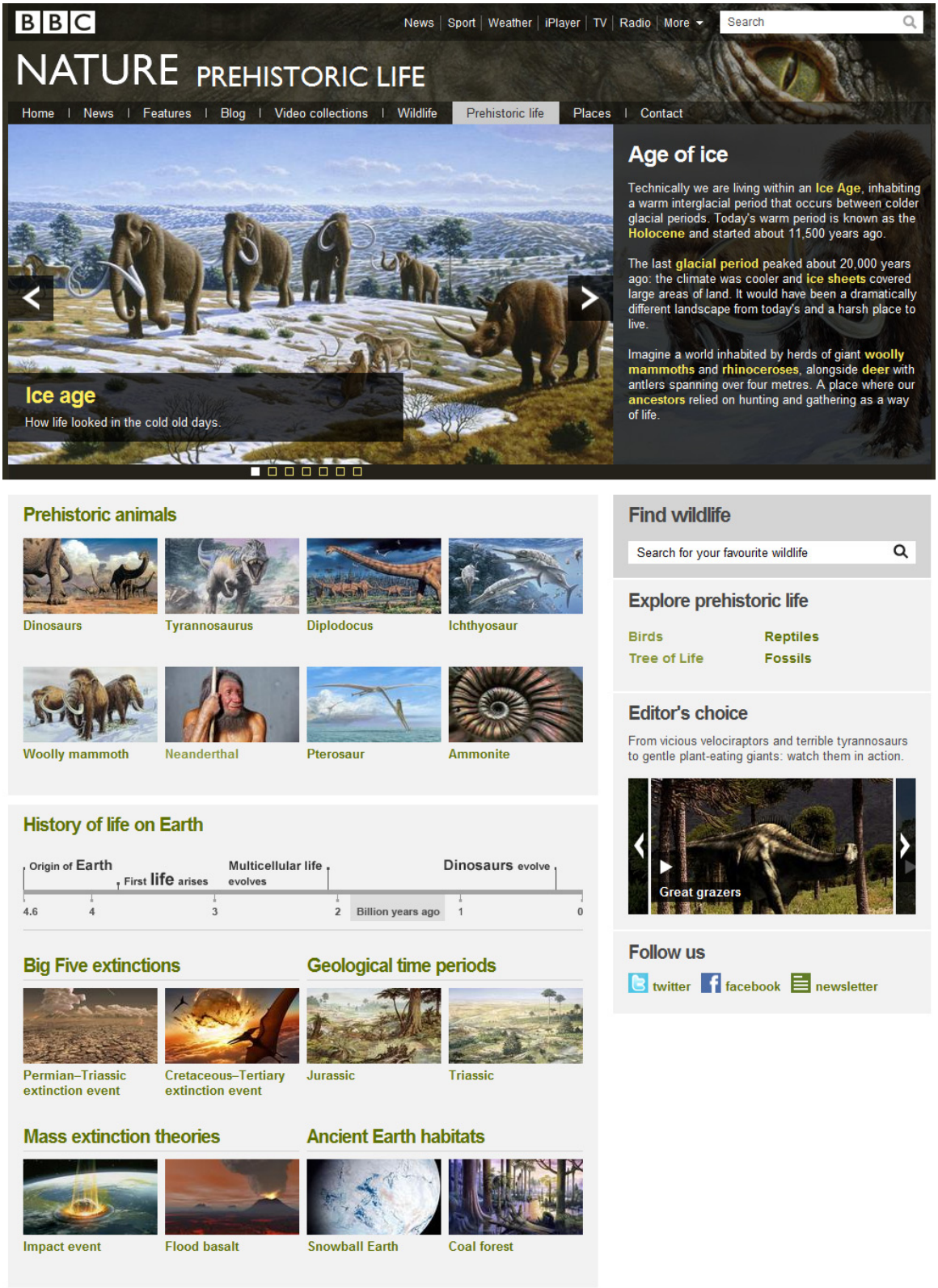
ANEXO E – Tela inicial da seção *Blog* do site BBC Nature

ANEXO F – Tela inicial da seção *Video Collections* do site BBC Nature

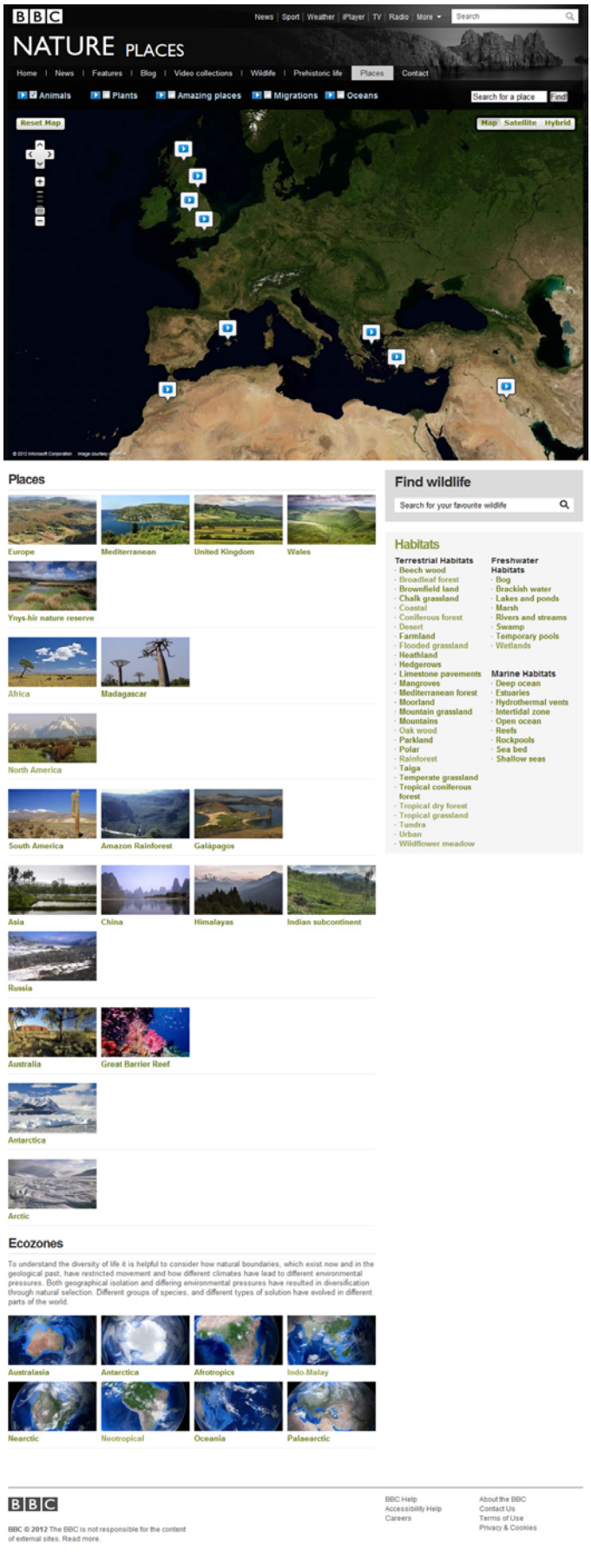
ANEXO G – Tela inicial da seção *Wildlife* do site BBC Nature



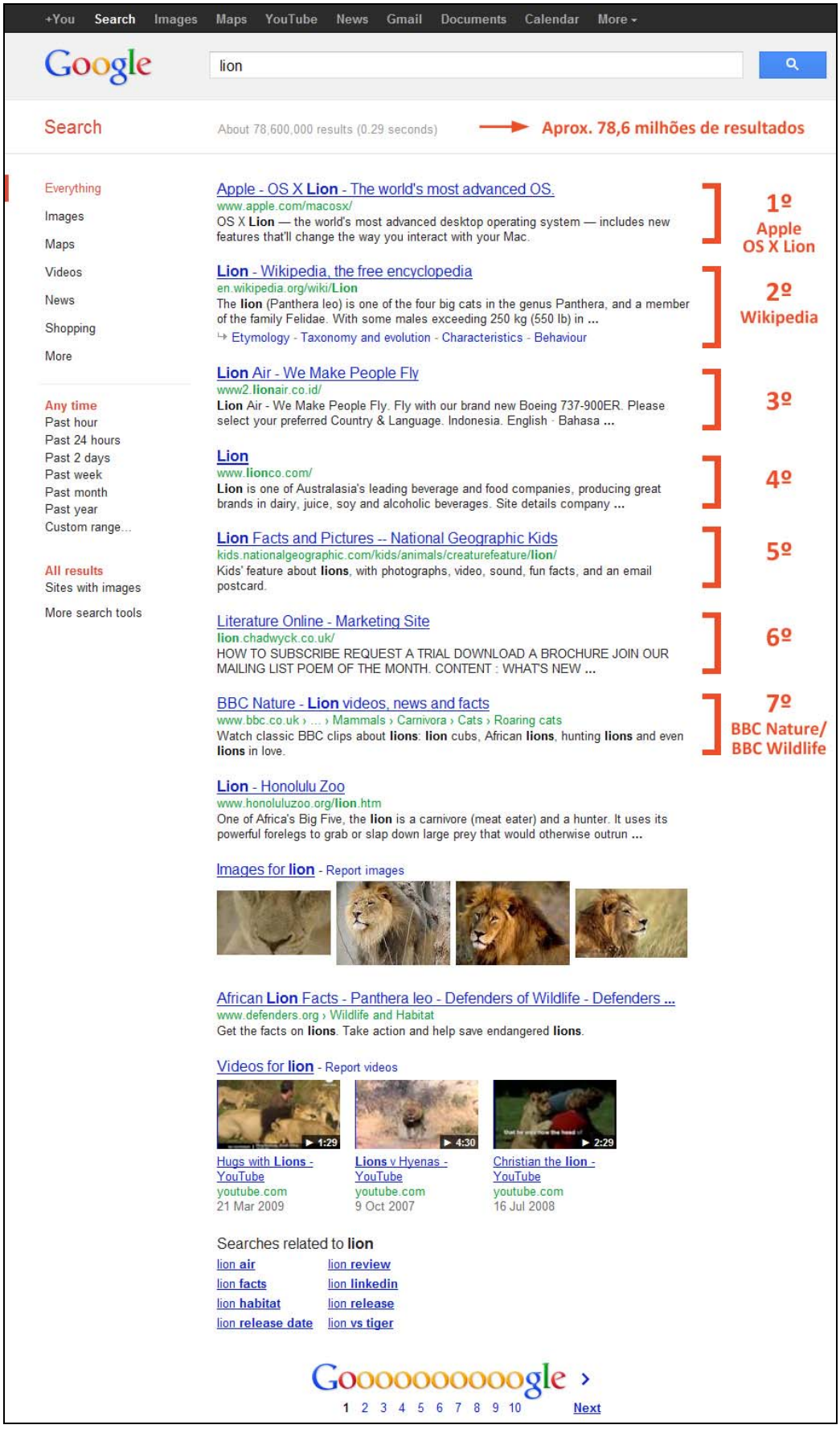
ANEXO H – Tela inicial da seção *Prehistoric Life* do site BBC Nature



ANEXO I – Tela inicial da seção *Places* do site BBC Nature



ANEXO J – Resultado de busca no Google pelo termo "lion"



ANEXO K – Resultado de busca no Google pelos termos "world cup 2010"