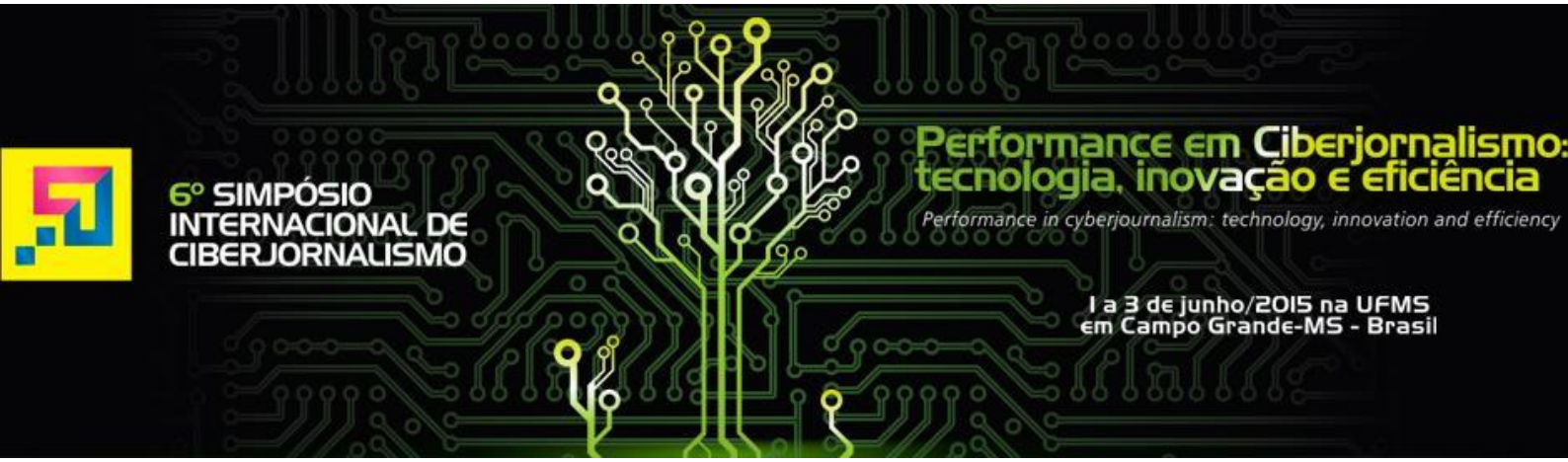

Interoperabilidade da informação jornalística em bases de dados: considerações sobre a adoção de metadados[1]

Walter Teixeira Lima Júnior[2]
André Rosa de Oliveira[3]

Resumo: Valorizada como espaço para reaproveitamento da memória jornalística, a utilização de bases de dados como repositório de informações pode ser incrementada a partir de uma estruturação por meio de metadados, incluindo a adoção de vocabulários controlados, ontologias formais e outras ferramentas semânticas. Com a ausência destas estruturas, informações tornam-se descontextualizadas com facilidade e perdem relevância. O artigo defende que o aproveitamento das bases de dados na web deve ir além de seu uso como repositório e recuperação por meio de palavras-chave: o impulso promovido pelos metadados estabelece a notícia como parte de um sistema elaborado, capaz de relacionar suas bases de dados, promover a interoperabilidade - ou seja, seu múltiplo uso em diversas plataformas - e transforma-se em um sistema de suporte à decisão para o jornalista a partir das relações e correlações construídas por máquinas.

Palavras-chave: Jornalismo Digital em Bases de Dados, Metadados, Interoperabilidade, Sistemas de Suporte à Decisão, Multidisciplinaridade.

[1] Artigo enviado na modalidade "Estudos de Jornalismo".

[2] Docente do Programa de Pós-Graduação da Universidade Metodista de São Paulo e pós-doutor em Comunicação e Tecnologias Digitais. E-mail: digital@walterlima.jor.br

[3] Jornalista. Docente nos cursos de Comunicação Social das Faculdades Integradas Rio Branco e doutorando pela Universidade Metodista de São Paulo. E-mail: andrerosa.jor@gmail.com

1 INTRODUÇÃO

Considere um programa jornalístico televisivo com boletins de trânsito. O apresentador utiliza-se, para apoiar sua informação, de uma visualização do Waze, sistema que faz o mapeamento e localização de ruas e avenidas em cidades com o intuito de indicar os melhores trajetos com base em coleta de informações sobre tráfego[4]. Ao final do boletim, o âncora faz um convite ao telespectador: "Informações do Waze, que você vê em detalhes no site de nossa emissora".

Independente da estranheza em chamar um software por meio de sua página, posicionando-se como um intermediário (questões comerciais não estão em discussão neste artigo), deve-se observar como um sistema, apoiado por aplicativos instalados em dispositivos móveis conectados à Internet durante a circulação de automóveis fez com que boletins tradicionais, com repórteres em carros ou helicópteros, pareçam anacrônicos, obsoletos. Mais do que isso: as informações do aplicativo minimizam variáveis determinantes no processo de construção da notícia em redações que exigem produção ágil e constante ao mesmo tempo em que os erros de informação, cometidos de maneira recorrente há décadas, sejam evitados.

Em outra situação, um profissional de comunicação preocupado com a credibilidade das informações que compartilha em seu perfil pessoal do Twitter decide experimentar uma aplicação externa, que dialoga com seu navegador, permitindo avaliação das mensagens, denominado TweetCred[5]. Ao analisar os dados da mensagem levando em conta 45 variáveis distintas, o sistema atribui nota de 1 a 7 para a mensagem, avaliação aperfeiçoada pelas características de aprendizagem semi-supervisionada do algoritmo (GUPTA et al., 2014).

Aplicações como Waze ou TweetCred, apropriadas pelo Jornalismo, transformam-se em Sistemas de Suporte à Decisão para esse profissional. Trata-se de soluções tecnológicas baseadas em sistemas computacionais que podem ser usados no suporte a complexos processos de tomada de decisão e resolução de problemas (LIMA JUNIOR, 2009). O conceito vem sendo delineado desde os anos 1970, quando pesquisadores na área de tecnologia da informação buscam melhorar a eficiência como usuários podem tomar decisões melhores e mais eficazes a partir de dados estruturados. Esta preocupação aumenta diante dos fluxos

[4] Disponível em <http://www.waze.com>. Acesso em 12.mai.2015.
[5] Disponível em <http://twitdigest.iiitd.edu.in/TweetCred/>. Acesso em 12.mai.2015.

informacionais em plataformas de mídia social conectada: desinformação, *astroturfings*[6], spams e fraudes se confundem com conteúdo relevante em um ecossistema informativo complexo (CIAMPAGLIA et al., 2015).

> Nos últimos anos, a computação ubíqua transformou a paisagem do jornalismo. Ele tem minado modelos de negócios, reequilibrado o poder relativo dos repórteres e do público e acelerando a entrega de informações em todo o mundo. Ainda que as formas mais antigas do jornalismo estejam começando a escapar de nossos olhos, acreditamos no entanto que a computação também começou a apresentar aos repórteres uma série de novas técnicas com o intuito de perseguir a antiga missão de interesse público do jornalismo. (TURNER; HAMILTON, 2009, p. 4, tradução nossa[7])

Este artigo pretende reforçar a importância da produção jornalística nesse ecossistema digital complexo, capaz de produzir um acervo de grande amplitude, porém com aproveitamento computacional limitado. A Web, como estamos acostumados, é composta basicamente por páginas com marcação HTML, contendo vários tipos de conteúdos estruturados – ou não. Em uma visão elaborada por Tim Berners-Lee, é possível estruturar dados ao invés de documentos, dotando-os de caráter semântico e tornando-os legíveis tanto por usuários quanto por máquinas (BERNERS-LEE; HENDLER; LASSILA, 2001).

Não bastam, portanto, práticas como a escolha de palavras-chaves, expressões e vínculos que potencializam sua indexação por meio de *tags*, bem como o uso de marcações HTML adequadas à estrutura dos documentos. Mesmo com estas ações, informações jornalísticas são facilmente descontextualizadas, perdendo relevância. O encadeamento de células informativas na Web – que constitui uma "memória múltipla, instantânea e cumulativa" (PALACIOS, 2003) não deve ser entendido apenas como um repositório. Ela assume caráter estruturante, procurando aperfeiçoar o processo de recuperação das informações e o relacionamento entre os conteúdos, reforçando o paradigma do Jornalismo Digital em Bases de Dados.

> O JDBD é o modelo que tem as bases de dados como definidoras da estrutura e da organização, bem como da apresentação dos conteúdos de

[6] Termo utilizado para designar ações políticas ou publicitárias que tentam criar a impressão de que são movimentos espontâneos e populares.

[7] Versão original: "In recent years, ubiquitous computation has transformed the landscape of journalism. It has undermined business models, rebalanced the relative power of reporters and audiences, and accelerated the delivery of information worldwide. Even as older modes of journalism are beginning to slip from view however, we believe that computation has also begun to present reporters with a series of new techniques with which to pursue journalism's long-standing public interest mission."

> 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; TORRES, 2013)

Para que a informação jornalística possa ser, de fato, interoperável – isto é, capaz de permitir a criação de produtos jornalísticos digitais dinâmicos – é necessário entendê-lo como um sistema, subordinado a uma costura computacional solta de dados, metadados e formatos realizada por atores humanos e não-humanos, exigindo novas experimentações e oportunidades (BERTOCCHI, 2014). Neste cenário, o Jornalismo será capaz de construir ferramentas consistentes, capazes de dialogar com a produção informativa e de apoiar a tomada de decisão do profissional, melhorando de maneira qualitativa o trabalho do jornalista.

2 METADADOS, ONTOLOGIAS E ALGORITMOS

Pesquisadores e jornalistas encontram, a partir de uma lógica computacional, oportunidades para encontrar métodos, fontes e caminhos para descobrir, apresentar, agregar, rentabilizar e arquivar histórias, conectando comunidades com a informação que elas precisam para governarem a si próprios (COHEN; HAMILTON; TURNER, 2011). Tanto o jornalismo quanto a ciência da informação se debruçam diante da relação do homem com a informação. Não à toa, as perspectivas do jornalismo computacional aproximam ainda mais os profissionais da comunicação e da tecnologia – isso inclui entender o jornalismo como um "código-fonte aberto", possível de ser "hackeado" (MANCINI, 2011; USHER; LEWIS, 2012).

O conceito de Web Semântica, elaborado por Tim Berners-Lee, potencializa as práticas de jornalismo computacional na Internet. Em um ambiente onde agentes de software possam identificar padrões de informação em páginas e executem tarefas complexas em bases de dados, estes podem ser capazes de serem utilizados e relacionados em aplicações variadas – definimos esta possibilidade como *linked data* (BIZER; HEATH; BERNERS-LEE, 2009). Enquanto a Web conecta documentos por meio de suas URLs, a Web Semântica estabelece conexões entre dados, que também devem ter localizações únicas, tornando possível a interoperabilidade da informação a partir de técnicas de integração de dados oriundos de fontes diferentes.

Esse contexto pode ser interpretado tanto como a expectativa de que sistemas poderão "compreender" dados e documentos; de outro, os obstáculos técnicos, econômicos ou mesmo das próprias organizações, que tornam sua possibilidade uma falsa promessa (LAMMEL; MIELNICZUK, 2012; RIBAS, 2007). Independente do posicionamento dos veículos, espera-se ainda a participação dos atuais consumidores de informação, atualmente espectadores, diante de uma arquitetura de conteúdos que permite sua adaptação e distribuição em novos canais por meio de APIs (GARCÍA; PERDRIX; GIL, 2006; PIETOSO, 2009). De toda forma, a primeira etapa do jornalismo na Web para fazer parte deste processo e ser entendido como um software se resume a um elemento: metadados.

Metadados são informações que permitem a descrição, organização, atualização, reutilização, validação, recuperação, preservação e recontextualização de objetos de informação, estruturando-os de modo a serem compreendidos tanto por humanos quanto por máquinas. Não se trata apenas de um acréscimo do código HTML, comuns em processos de otimização de documentos na Web, mas sim da descrição de objetos e suas relações com outros conceitos.

Esta representação é importante: de que forma é possível expressar dados e regras em uma linguagem capaz de permitir relacionamento e integração entre os dados? Metadados são percebidos como ingrediente para a competitividade da informação. Artigo do estrategista em conteúdo Michael Andrews[8], publicado na área de inteligência do *Content Marketing Institute*, como "algo invisível, como um tempero que funciona mesmo sem saber que ele está ali"[9]. Tal discurso, no entanto, não revela os obstáculos diante da exigência permanente por agilidade.

> Atualmente, metadados para notícias são bastante heterogêneos e difíceis de serem enriquecidos ou detalhados suficientemente para cobrir todo o conhecimento que estes documentos contêm. Anotações manuais são impraticáveis e infindáveis. Ferramentas de marcação automáticas permanecem muito pouco desenvolvidas. Portanto, serviços informativos especializados exigem ferramentas que podem pesquisar e extrair informação específica diretamente de textos não estruturados na Web. Estas ferramentas podem ser guiadas por uma ontologia que determinaria qual tipo

[8] Artigos e perfil pessoal disponível em <http://storyneedle.com>

[9] Robust Metadata: The Secret Sauce of Relevance. Disponível em <http://contentmarketinginstitute.com/intelligent-content/blog/metadata-secret-sauce-relevance/>. Acesso em 8.mai.2015.

> de informação seria extraído. (KALLIPOLITIS; KARPIS; KARALI, 2012, tradução nossa[10])

No contexto computacional, as representações do conhecimento expressas por linguagens de marcação representam camadas de base. A partir dos metadados temos ontologias e vocabulários controlados, abrindo caminhos para o desenvolvimento de aplicações que abrem possibilidade para comparar, relacionar e combinar bases distintas de dados. Ontologias são infraestruturas de representação formal do conhecimento em algum domínio de interesse, entendido como um conjunto de conceitos, relações e funções dentro de um vocabulário comum, com contexto definido e sem ambiguidades. Constitui um tipo muito específico de metadados, direcionados para lógicas formais de máquina (SICILIA; LYTRAS, 2009). Pesquisadores ligados à Web reduziram o termo, definindo-a como os bancos de metadados que definem formalmente as relações entre os termos, dialogando com vocabulários controlados e técnicas como processamento de linguagem natural.

Com informações estruturadas por metadados disponíveis, somadas a uma trilha de softwares – isto é, uma ou mais listas de instruções específicas direcionadas a uma pergunta específica, temos os componentes elementares de um algoritmo. O poder destes agentes inteligentes, capazes de orientar o fluxo de trânsito diante de alternativas melhores sem a intervenção humana, desperta inegável fascínio.

> Algoritmos já escreveram sinfonias que se movem como as compostas por Beethoven; percorreram termos legais com a destreza de um advogado sênior; diagnosticam pacientes com mais precisão do que um médico, escrevem artigos noticiosos com a mão suave de um repórter experiente; dirigem veículos em vias urbanas com melhor controle do que um ser humano. (STEINER, 2012, tradução nossa[11])

Em outras palavras, o autor destaca: programas que varrem sites em busca de palavras-chave requer alguma habilidade, já dominada por milhões de pessoas. Mas para criar algo de fato inovador, por meio de um algoritmo elegante capaz de resolver problemas humanos, é

[10] Versão original: "Metadata for news items are currently quite heterogeneous and it is difficult to be rich or detailed enough to cover all the knowledge that these documents contain. Manual annotation is impractical and unscalable and automatic annotation tools remain largely undeveloped. Therefore, specialized knowledge services require tools that can search and extract specific knowledge directly from unstructured text on the Web. These tools could be guided by an ontology that would determine what type of knowledge to harvest."

[11] Versão original: "Algorithms have already written symphonies as moving as those composed by Beethoven, picked through legalese with the deftness of a senior law partner, diagnosed patients with more accuracy than a doctor, written news articles with the smooth hand of a seasoned reporter, and driven vehicles on urban highways with far better control than a human."

preciso talento. Isso inclui Sistemas de Suporte à Decisão – ou mesmo softwares que prometem a automação na produção de notícias.

3 SISTEMAS AMEAÇAM OU ABREM NOVAS OPORTUNIDADES?

O aumento na disponibilidade de dados estruturados na Web despertou atenção das organizações de mídia. formatos de metadados voltados para sistematizar processos de arquivamento e digitalização de informações jornalísticas. Destaque para o NITF (*News Industry Text Format*[12]), uma especificação para marcações de conteúdo e estrutura em XML publicada pela *International Press Telecommunications Council* (IPTC). Os recursos disponibilizados por este conselho permitem a adoção de metadados e ontologias a objetos como textos, fotografias, áudios e vídeos, maximizando a interoperabilidade de informação e produzindo conexões significativas (TRONCY, 2008). Em 2010, o boletim do IPTC (MIRROR, 2010) foi além e repercutiu a seguinte questão entre seus leitores: "a mídia consegue utilizar *linked data* por um futuro mais forte"? Um olhar mais detalhado nos modelos econômicos em redações revela os maiores obstáculos entre a discussão acadêmica e a prática, além da ausência de uma "cultura de metadados" interna.

> A experiência mostra que, devido a aversão ao risco, falta de recursos financeiros e atores experientes, a indústria da mídia tende a se comportar com muita cautela quando se trata da adoção de novas tecnologias e metodologias de criação de conteúdo e reutilização, especialmente quando eles carregam um forte potencial disruptivo e afetam seu core business, a competência ou a cultura corporativa. (PELLEGRINI, 2012, tradução nossa[13])

Paralelamente, outras ferramentas abertas que permitem anotações semânticas manuais, capazes de associar metadados ao conteúdo jornalístico de forma amigável. PETASIS (2012) recorda que, na última década, uma variedade de ferramentas para anotações semânticas foram desenvolvidas (incluindo a popular *GATE*[14]), além de apresentar a desenvolvida pelo autor, o *SYNC3*. A lista de possibilidades para qualquer usuário criar estrutura de dados semânticos em conteúdos Web continua: *PundIt*, *Hermes* e *Loomp*

[12]Disponível em <http://www.nitf.org>. Acesso em, 16.jan.2015

[13] Versão original: "Experience shows that due to risk aversion, lack of financial resources and expertise actors in the media industry tend to behave very cautiously when it comes to the adoption of new technologies and methodologies of content creation and reuse, especially when they carry a strong disruptive potential and affect their core business, competencies or corporate culture."

[14] Disponível em <http://gate.ac.uk>, acesso em 8.mai.2015

(FRASINCAR; BORSJE; LEVERING, 2009; GRASSI et al., 2013; LUCZAK-RÖSCH; HEESE, 2009).

Processos de codificação manual, no entanto, representam limites para bases de dados mais extensas, levando ao desenvolvimento de softwares especializados em analisar conteúdos não estruturados e extrair conceitos e metadados de forma automática. Nesse sentido, vale destacar o audacioso projeto *Global Data on Events, Location and Tone* (GDELT)[15], plataforma que monitora a mídia e acumula informações datadas de 1979, codificando-as e estruturando-as a partir de um esquema denominado CAMEO. Mais do que isso: conecta pessoas, organizações, localizações e temas, sendo capaz de identificar padrões e identificar focos de conflito e violência política com antecedência. São 100.000 novos eventos de todo o planeta diariamente a partir de fontes como *Associated Press*, *France Presse* e a chinesa *Xinhua*, compondo um *dataset* com caracrerísticas consideradas necessárias para construir predições em escala global: abrangência global, densidade, codificação geográfica, precisão e disponibilidade de acesso futuro (YONAMINE, 2013).

Outras técnicas relacionando algoritmos ao conteúdo editorial fazem mais barulho, ainda que não representem ideias essencialmente novas. Em 2001, doze anos antes de Jeff Bezos adquirir o *The Washington Post*, Shayne Bowman e Chris Willis propunham[16]: e se um portal pudesse oferecer notícias da mesma forma que a *Amazon* recomenda seus livros? No mesmo ano, a Universidade de Columbia iniciava o projeto *NewsBlaster*, sistema inteligente que explorava, extraía, agrupava, organizava e classificava notícias por "clusters" e tópicos, gerando páginas Web automaticamente (GRASSI et al., 2013).

A classificação e recomendação de notícias proposta pelo *NewsBlaster* abriu outras frentes de pesquisa, tais como o incremento deste agrupamento por meio de *tags* e bookmarks produzidos pelos usuários (RAMAGE et al., 2009), ou mesmo a linha de um projeto que tornou-se mais popular nessa linha: o *Google News*, que se aproveita ainda do histórico de navegação do usuário (LIU; DOLAN; PEDERSEN, 2010). O agregador automatizado gerou controvérsia em veículos informativos na Bélgica, no Brasil e, mais recentemente, na

[15] Disponível em <http://gdeltproject.org/>. Acesso em 12 March 2015
[16] Disponível em <http://www.hypergene.net/ideas/amazon.html>. Acesso em 8.mai.2015.

Espanha[17], onde os periódicos se desconectaram da ferramenta alegando prejuízos, já que "há um aproveitamento do conteúdo" pela ferramenta sem uma compensação.

Se os primeiros exemplos ou discussões se assemelham a uma abstração tecnológica e os segundos correspondem a uma ameaça ao modelo estabelecido, ambos incapazes de ressoar com as demandas econômicas dos veículos, qualquer atributo destes sistemas parece, em princípio, irrelevante. Ao contrário: dão voz a iniciativas que propõem a "exclusão da internet do conteúdo dos jornais como porta de saída da crise mundial da mídia impressa"[18]. Em síntese, uma derrota "cultural" ou "financeira".

> Claramente, estes dois olhares são defasados. E que provavelmente ambos estão errados, ou melhor, estão nitidamente incompletos. Trata-se do medo histórico de autonomia das máquinas, aplicado a uma profissão que sofre, não aproveita a plenitude de incertezas que atravessa e que poderiam contar a seu favor. Entender os algoritmos de outra maneira, menos reducionista, é uma missão intelectual capital para repensar o que fazemos nesta indústria onde, supunha-se, nada poderia ser automatizado. (MANCINI, 2011, p. 45, tradução nossa[19])

4 FERRAMENTAS SEMÂNTICAS COMO SSD

"Somos uma empresa de notícias, não uma empresa de jornal". A frase, pinçada de um memorando interno do jornal *The New York Times* enviado por Arthur Sulzberger e Janet Robinson, é lembrada como exemplo de compromisso com a informação seja qual for a plataforma. A área de desenvolvedores do jornal[20] inclui *datasets* específicos e informações relacionadas ao acervo do jornal, publicado a partir de 1851, digitalizado. Desde 2009[21], um vocabulário formado por pessoas, organizações, exemplos e outras descrições é disponibilizado como *linked open data* para utilização em outras aplicações[22].

---

[17] Informações sobre este histórico em <http://www.xataka.com/aplicaciones/google-news-cierra-en-espana-pero-que-ha-pasado-ante-situaciones-similares-por-el-mundo>. Acesso em 8. mai.2015

[18] "Bala de Prata". Disponível em <http://observatoriodaimprensa.com.br/imprensa-em-questao/_ed837_bala_de_prata/>. Acesso em 8.mai.2015.

[19] Versão original: "Está claro que esas dos miradas atrasan. Y que probablemente las dos estén equivocadas o, mejor dicho, sean marcadamente incompletas. Se trata del miedo histórico a la autonomía de las máquinas aplicado a una profesión que padece y no aprovecha la plena alza de incertidumbre que atraviesa y podria contar a su favor. Entender a los algoritmos de otra manera, menos reduccionista, es una misión intelectual mayúscula para repensar lo que hacemos en esta industria donde, se suponía, nada podía ser automatizado"

[20] Disponível em <http://developer.nytimes.com>. Acesso em 12.mar.2015.

[21] Anúncio foi feito em <http://open.blogs.nytimes.com/2009/06/26/nyt-to-release-thesaurus-and-enter-linked-data-cloud/>. Acesso em 12.mar.2015.

[22] Disponível em <http://data.nytimes.com>. Acesso em 12.mar.2015.

O *The Guardian*, por sua vez, disponibiliza um mecanismo que permite acesso aos artigos publicados no site desde 1999, bem como dados estruturados sobre temas gerais em sua *Open Platform*[23], dividida em quatro módulos: *Content API*, *Data Store*, *Politics API* e *MicroApps*. Todas permitem a reutilização dos conteúdos disponibilizados pelo veículo.

Mas é a *British Broadcasting Corporation* (BBC) que demonstra forte relação entre desenvolvedores e seu conteúdo. Seu exemplo pioneiro, o site *BBC Wildlife*[24], tornou-se um dos primeiros repositórios utilizados como complemento, por meio de tecnologias semânticas, a outros produtos jornalísticos da BBC. Isto é, sistemas que decidem como os conteúdos devem ser publicados a partir do processamento de metadados, enriquecendo o produto final (LAMMEL; MIELNICZUK, 2012). Na esteira desta experiência, buscaram enriquecer informações utilizando metadados durante a Copa de 2010[25], esforço ampliado durante os Jogos Olímpicos de 2012, em Londres[26].

A partir da adoção de ferramentas semânticas no esporte, a *BBC Future Media* apresentou a nova versão de suas ontologias[27],base para sua plataforma de *linked data*, em abril de 2014. Como resultado, o serviço *BBC Things*[28], lançado em setembro do mesmo ano, oferece acesso público a estes conceitos, permitindo a criação de aplicações a partir de seus dados – na prática, o site da BBC funciona como uma API.

Equipes multidisciplinares aprendem novos conceitos e tomam decisões a partir de protótipos desenvolvidos e compartilhados de forma aberta em um laboratório, o *BBC News Labs*, aprendendo sobre novas tecnologias e construindo um legado de informações estruturadas em suas bases de dados. Em um destes experimentos, elaborados em um evento Hack/Hackers[29], surgiu o algoritmo *Datastringer*. Ele permite ao jornalista monitorar com

[23] Disponível em <http://open-platform.theguardian.com/>. Acesso em 12.mar.2015.
[24] Disponível em <http://www.bbc.co.uk/nature/wildlife>, Acesso em 16.jan.2015
[25] Disponível em <http://www.bbc.co.uk/blogs/legacy/bbcinternet/2012/04/sports_dynamic_semantic.html>. Acesso em 16.jan.2015
[26] Disponível em <http://www.bbc.co.uk/blogs/legacy/bbcinternet/2012/04/sports_dynamic_semantic.html>. Acesso em 16.jan.2015
[27] Disponível em <http://www.bbc.co.uk/blogs/internet/entries/78d4a720-8796-30bd-830d-648de6fc9508>. Acesso em 23.fev.2015
[28] Disponível em <http://www.bbc.co.uk/things>. Acesso em 23.fev.2015
[29] From the BBC News Labs: Datastringer. <https://source.opennews.org/en-US/articles/bbc-news-labs-datastringer/>. Acesso em 8.mai.2015.

facilidade bases de dados externas a partir de critérios definidos por uma pauta, com alertas programados (KOBILAROV et al., 2009). O algoritmo é *open-source*, ou seja, está disponível para uso, ajustes ou melhorias[30].

No Brasil, o caso mais relevante diz respeito a adoção de tecnologias semânticas pelos sites de notícia da *Globo.com*[31]. As três áreas que compõem o portal (notícias, esportes e entretenimento) possuem, muitas vezes, assuntos semelhantes com pontos de vista diferentes: Romário, por exemplo, pode ser entendido como ex-jogador pelo *Globoesporte.com*, senador da República pelo *G1* e celebridade pelo *Gshow*. É possível interligar este conteúdo por meio de ferramentas da web semântica?

Figura 1: protótipo de publicador com anotações semânticas da Globo.com[32]

[30] Disponível em <https://github.com/BBC-News-Labs/datastringer>. Acesso em 8.mai.2015.

[31] Alguns exemplos desta implantação podem ser visualizados no portfólio do desenvolvedor Renan Oliveira: http://renanoliveira.net>. Acesso em 12.mar.2015.

[32] Reprodução de < http://pt.slideshare.net/renangpa/ontologias-e-sua-utilizao-em-aplicaes-semnticas-uff-casi-2014>. Acesso em 8.mai.2015.

Em 2009, o portal iniciou o projeto desenvolvendo ferramentas de anotação manual, integrada ao publicador, e ontologias. Ambas se adaptam por conta dos campos e valores disponíveis no próprio publicador, sugestões que garantem o uso de metadados que dialogam com vocabulários controlados. As anotações ficam armazenadas em um banco de triplas, isto é, relações entre conceitos e objetos. Está no radar do projeto a extração automática de termos com processamento de linguagem natural, bem como exportar conceitos por meio de *linked open data*.

5 CONSIDERAÇÕES FINAIS

A apropriação do Waze por uma emissora de TV, mencionado no princípio deste artigo, é um exemplo da sobreposição de tecnologias digitais no jornalismo, procurando uma reconfiguração. Da mesma forma, durante os anos 1990, redações buscavam adaptações com a introdução maciça de computadores pessoais, compreendidos neste princípio como um "tipo avançado de máquina de escrever".

A lição, nos dois casos, é a mesma: para extrair relevância de suas bases de dados, alimentadas diariamente com informações que transformam-se em documentos com conexões limitadas, e ao mesmo tempo não ser surpreendido, o Jornalismo precisa dominar tecnologias digitais. Ferramentas da web semântica, pautadas por metadados, vocabulários, esquemas e ontologias, correspondem a uma forma de reforçar a informação.

Todos os casos apresentados apresentam relações entre a notícia, entendida como um objeto de informação estruturado por metadados, e o desenvolvimento sistemas que permitem sua formalização semântica, recuperação e reutilização para aplicações variadas. Representam, acima de tudo, um ganho de informação obtido a partir de um necessário esforço multidisciplinar: desenvolvedores, jornalistas ou novos atores interessados podem atuar neste ambiente, desde que as ferramentas se apresentem de forma aberta – compartilhadas em serviços como *GitHub*[33], por exemplo.

Ao mesmo tempo, ainda que algoritmos sejam capazes de agrupar termos e organizar clusters de dados, computadores não substituirão o caráter humano. Softwares podem organizar e relacionar dados e ideias com facilidade, atuando como Sistemas de Suporte à

[33] Disponível em <http://github.com/>. Acesso em 8.mai.2015.

Decisão. Mas é o jornalista que deverá explorar estes caminhos com detalhes: ao reduzir as dificuldades nas atividades rotineiras de investigação, estas ferramentas digitais não substituem a percepção do profissional, responsável pela elaboração e interpretação da narrativa.

6 REFERÊNCIAS

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The Semantic Web. **Scientific American**, n. May 2001, p. 34–43, 2001.

BERTOCCHI, D. **Dos dados aos formatos: o sistema narrativo no jornalismo digital**XXIII Encontro Anual da Compós. **Anais**...Belém, PA: 2014

BIZER, C.; HEATH, T.; BERNERS-LEE, T. Linked Data - The Story So Far. **International Journal on Semantic Web and Information Systems (IJSWIS)**, 2009.

CASTELLS, P.; PERDRIX, F.; PULIDO, E. Neptuno: Semantic web technologies for a digital newspaper archive. 2004.

CIAMPAGLIA, G. L. et al. Computational fact checking from knowledge networks. p. 1–20, 2015.

COHEN, S.; HAMILTON, J. T.; TURNER, F. Computational journalism. **Communications of the Association for Computing Machinery**, v. 54, n. 10, p. 66–71, 1 out. 2011.

DANIEL, A.; FLEW, T. The Guardian Reportage of the UK MP Expenses Scandal: a Case Study of Computational Journalism. **Communications Policy and Research Forum**, n. November, 2010.

FRASINCAR, F.; BORSJE, J.; LEVERING, L. A semantic web-based approach for building personalized news services. **International Journal of E-Business …**, n. 2, 2009.

GARCÍA, R.; PERDRIX, F.; GIL, R. **Ontological infrastructure for a semantic newspaper**. [s.l: s.n.]. Disponível em: <http://www.image.ntua.gr/swamm2006/resources/paper07.pdf>. Acesso em: 6 jun. 2013.

GRASSI, M. et al. Pundit: augmenting web contents with semantics. **Literary and Linguistic Computing**, v. 28, n. 4, p. 640–659, 18 set. 2013.

GUPTA, A. et al. **TweetCred: Real-Time Credibility Assessment of Content on Twitter**Proc. 6th International Conference on Social Informatics (SocInfo). **Anais**...2014

KALLIPOLITIS, L.; KARPIS, V.; KARALI, I. Semantic search in the World News domain using automatically extracted metadata files. **Knowledge-Based Systems**, v. 27, p. 38–50, mar. 2012.

KOBILAROV, G. et al. Media Meets Semantic Web – How the BBC Uses DBpedia and Linked Data to Make Connections. **ESWC 2009**, p. 723–737, 2009.

LIMA JUNIOR, W. T. O uso dos Sistemas de Suporte à Decisão (SSD) visando à melhora da qualidade do conteúdo jornalístico. **Revista FAMECOS**, v. 38, p. 79–85, 2009.

LIU, J.; DOLAN, P.; PEDERSEN, E. R. **Personalized news recommendation based on click behavior**Proceedings of the 15th international conference on Intelligent user interfaces - IUI '10. **Anais**...New York, New York, USA: ACM Press, 2010Disponível em: <http://dl.acm.org/citation.cfm?doid=1719970.1719976>

LUCZAK-RÖSCH, M.; HEESE, R. **Linked Data Authoring for Non-Experts.**WWW2009. **Anais**...Madri: 2009Disponível em: <http://ceur-ws.org/Vol-538/ldow2009_paper4.pdf>. Acesso em: 15 set. 2014

MANCINI, P. **Hackear el periodismo - Manual de laboratorio**. Buenos Aires: La Crujía, 2011.

PALACIOS, M. Ruptura, Continuidade e Potencialização no Jornalismo Online: o Lugar da Memória. In: MACHADO, E.; PALACIOS, M. (Eds.). . **Modelos do Jornalismo Digital**. Salvador: Editora Calandra, 2003.

PELLEGRINI, T. **Semantic Metadata in the News Production Process - Achievements and Challenges**MindTrek. **Anais**...Tampere, Finland: 2012

PETASIS, G. The SYNC3 Collaborative Annotation Tool. **Lrec**, p. 363–370, 2012.

PIETOSO, C. R. **Newspapers as Platforms: How Open APIs Can Impact Journalism**London, UKCity University, , 2009.

POLLERES, A. et al. Can we ever catch up with the Web? **IOS Press**, p. 1–5, 2010.

RAMAGE, D. et al. Clustering the tagged web. **Proceedings of the Second ACM International Conference on Web Search and Data Mining WSDM 09**, v. 42, p. 54, 2009.

SICILIA, M.-A.; LYTRAS, M. **Metadata and Semantics**. New York, NY: Springer Science+Business Media, LLC, 2009.

STEINER, C. **Automate This: how algorithms came to rule the world**. London: Portfolio / Penguin, 2012.

TURNER, F.; HAMILTON, J. T. **Accountability Through Algorithm: Developing the Field of Computational Journalism**. Disponível em: <http://dewitt.sanford.duke.edu/wp-content/uploads/2011/12/About-3-Research-B-cj-1-finalreport.pdf>. Acesso em: 18 set. 2012.

USHER, N.; LEWIS, S. C. **Open source and journalism: Toward new frameworks for imagining news innovation**Paper accepted for presentation to the Journalism Studies Division of ICA, Phoenix, AZ. **Anais**...2012

YONAMINE, J. E. Predicting Future Levels of Violence in Afghanistan Districts. **The 3rd Annual Meeting Of The European Political Science Association**, p. 1–32, 2013.