Revista Famecos
mídia, cultura e tecnologia

ISSN: 1415-0549
e-ISSN: 1980-3729

DOI: http://dx.doi.org/10.15448/1980-3729.2017.3.25666

Jornalismo

Jornalismo estruturado: uso de metadados para enriquecimento de bases noticiosas na web[1]

Structured journalism: using metadata for enrichment of news databases on the web

Walter Teixeira Lima Junior

Jornalista e pesquisador. Professor da Universidade Federal do Amapá. Pós-doutorando do Departamento de Mecatrônica da Universidade de São Paulo.
<digital@walterlima.jor.br>

Andre Rosa de Oliveira

Professor das Faculdades Integradas Rio Branco. Doutorando em Comunicação Social na Universidade Metodista de São Paulo.
<andrerosa.jor@gmail.com>

RESUMO

Bases de dados abastecidas com notícias produzidas para a Web representam um repositório de informação estruturada com potencial tecnológico de ser reutilizada de inúmeras formas e por outras plataformas digitais conectadas via redes. No entanto, a dinâmica de recuperação deste material costuma ser limitada a busca por palavras-chave; da mesma forma, a sua organização é composta por categorizações simples ou apenas por marcações em HTML, não permitindo a flexibilização do seu uso de outras formas. Com base na observação de veículos e produtos jornalísticos, especialmente a britânica BBC, o artigo trata de novas ferramentas baseadas na adoção de vocabulários controlados, ontologias formais e outros padrões de metadados estruturam melhor a recuperação da informação jornalística inserida em banco de dados. Ressalta, dessa forma, a importância em estruturar a informação jornalística por meio de metadados e estendê-la além de seu uso como repositório, mas na obtenção de "relações invisíveis" de temas e contextos, entre outros fatores de diferenciação na qualidade informativa entre grupos de mídia.

Palavras-chave: Jornalismo. Metadados. Multidisciplinaridade.

ABSTRACT

Databases supplied with news produced for the Web represent a repository of information structured with technological potential to be reused in many ways and with other digital platforms connected via networks. However, the recovery dynamics of this material is usually limited to keywords search; moreover, that organization is composed of simple categorizations or just by tags in HTML, not allowing the flexibility of its use in other ways. Based on the observation of vehicles and journalistic products, especially BBC, this paper discusses new tools based on the adoption of controlled vocabularies, formal ontologies and other metadata standards properly structure the recovery of journalistic information inputed into databases. In this way, highlights the importance of journalistic information structured through metadata and extend it aside from its use as a repository, but in achieving "invisible relations" issues and contexts, among other differentiating factors in information quality between media groups.

Keywords: Journalism. Metadata. Multidisciplinarity.

1 Trabalho apresentado ao Grupo de Trabalho Estudos de Jornalismo do XXV Encontro Anual da Compós, na Universidade Federal de Goiás, Goiânia, de 7 a 10 de junho de 2016.

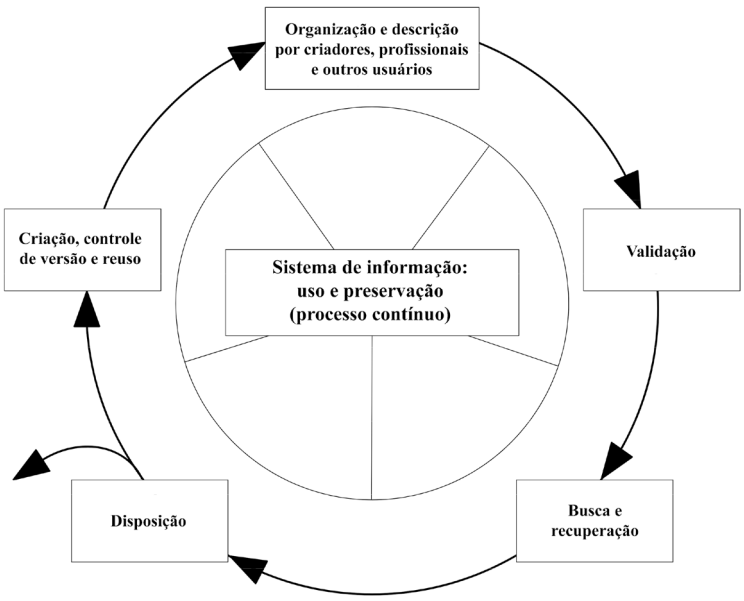
INTRODUÇÃO

Desde janeiro de 1994, quando o semanário Palo Alto Weekly reproduziu parte do material de sua edição impressa na Web, o jornalismo busca as melhores alternativas para compartilhar e armazenar informação neste ambiente, composto por documentos codificados em marcação hipertextual e relacionados entre si, acessadas por meio de softwares específicos (navegadores). Entretanto, o desejo humano de extrair conhecimento através do relacionamento de dados e informações de diversas fontes remonta antes do advento das tecnologias digitais conectadas existe desde as formulações do filósofo e cientista Gottfried Wilhelm von Leibniz (Biblioteca Universal), do dispositivo modulado por Vannevar Bush capaz de armazenar e recuperar informação (Memex), passando pela cooperação entre homem e máquina imaginada por J.C.R. Licklider (Libraries of future), a formalização da rede de informações através de hyperlinks criada por Tim Berners-Lee (Web), até a formatação de estrutura para colaboração e para obtenção de conhecimento implantada por Jimmy Wales (Wikipedia).

Seguindo a linha de pensamento e tentando construir mais uma ponte para obtenção do conhecimento através da interligação de repositório de dados, o cientista da computação Calvin Mooers dedica-se ao tema e elabora três perguntas que, mesmo elaboradas nos anos 1950, permanecem atuais: como descrever intelectualmente a informação? Como especificar intelectualmente a busca por ela? Quais sistemas, técnicas ou máquinas devem ser utilizados para isso? (Saracevic, 1996). Estas indagações se aplicam à Web, que se transformou um ambiente amigável de navegação, permitiu o desenvolvimento de ferramentas para produção e compartilhamento de conteúdos com facilidade e tornou-se um poderoso (e complexo) repositório de informação, na contramão de uma "galáxia de notícias" capaz de auxiliar usuários a explorar e organizar com facilidade informações relacionadas entre si (Rennison, 1994).

A Ciência da Informação nasce deste problema, e com da necessidade de organizar informação, surgem os metadados. Eles representam o que pode ser descrito a respeito de um objeto de informação em qualquer nível. Nesse contexto, objetos de informação podem ser definidos como qualquer coisa que pode ser endereçado e manipulado por um ser humano ou um sistema de informação. Um objeto corresponde a um item isolado, vários itens juntos ou uma base de dados inteira (Gilliland, 2008).

Figura 1 ciclo de um objeto de informação



Fonte: Gilliland (2008)

Dessa forma, o encadeamento de células informativas na Web – que constitui uma "memória múltipla, instantânea e cumulativa" (Palacios, 2008) – pode ser entendida como uma relação entre objetos de informação. A Figura 1 ilustra o ciclo de vida contínuo destes objetos, desde sua criação até sua disponibilização em sistemas de informação que se relacionam com bases de dados. Assim, o uso destas bases 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 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, p. 154)

Entre as preocupações para a efetiva estruturação de textos jornalísticos em bases de dados e sua visualização em páginas Web, verifica-se 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 (Corrêa; Bertocchi, 2012). Isso não é suficiente: mesmo com estas ações, informações jornalísticas são facilmente descontextualizadas, perdendo relevância. Ao apresentar seu modelo de sistema narrativo, Bertocchi (2014)sugere que o mesmo está subordinado a uma costura computacional solta de dados, metadados e formatos realizada por atores humanos e não-humanos, exigindo novas experimentações e oportunidades.

Este pensamento ganha força a partir da visão de Tim Berners-Lee, que aponta para uma Web Semântica, ambiente onde agentes de software identifiquem padrões de informação nas páginas e sejam capazes de executar tarefas complexas, permitindo ainda que computadores e pessoas trabalhem em cooperação com usuários (Berners-Lee; Hendler; Lassila, 2001). Tal visão se relaciona com o conceito de *linked data*, que refere-se a dados publicados na web, de tal forma a serem descobertos e legíveis por máquinas capazes de utilizá-los e relacioná-los em aplicações diversas (Bizer; Heath; Berners-Lee, 2009).

Há uma expectativa de que a transição entre uma Web que conecte documentos para a Web de Dados, permeada pelos princípios do *linked data*, abram as portas dos silos informativos e habilitem os efeitos da rede. Em um artigo onde discute as limitações do OpenGraph, esquema de metadados adotados pelo Facebook, o especialista em dados Tyler Bell sintetiza: mais do que um padrão, *linked data*é um *ethos*, focado em produção de contexto, desambiguação e descobertas não triviais. Isso pode ser entregue por meio de dados, plataforma e aplicação trabalhando juntos (Bell, 2010).

Desta visão emergiu a ideia do "jornalismo estruturado", termo que surgiu pela primeira vez numa proposta do editor de inovação e dados da Thomson Reuters, Reginald Chua (2010). Em essência, propõe a fragmentação de narrativas jornalísticas em partes reunidas e relacionadas entre si. Alexis Lloyd, diretora criativa do Laboratório de Pesquisa e Desenvolvimento do *The New York Times*, revelou que o Project Editor, por exemplo, "analisa a forma como alguns metadados granulares podem ser criados por meio de sistemas colaborativos que dependem fortemente de aprendizado de máquina, bem como inputs editoriais" (Lloyd, 2015).Ao apresentar seu protótipo Structured Stories, que coleta fragmentos de notícias relacionados a eventos específicos e, a partir de uma codificação prévia desses elementos, oferece ao usuário narrativas maiores, David Caswell apresenta sua definição do termo.

> Jornalismo estruturado é uma nova forma de jornalismo baseada em reportagens como componentes estruturados em uma base de

> dados, e posterior recuperação destes componentes estruturados para gerar produtos informativos. A abordagem ainda é incipiente, mas lida diretamente com diversos problemas sistêmicos enfrentados por produtores e consumidores de notícias em um ecossistema de mídia digital, e pode potencialmente facilitar o rearranjo do Jornalismo em redes, bem como a criação de produtos informativos controlados pelo consumidor num contexto que se estende além do artigo (CASWELL; RUSSELL; ADAIR, 2015, tradução nossa)[2].

Este artigo apresenta uma relação entre o conteúdo jornalístico armazenado em bases de dados e iniciativas de *linked data*, tendo como base sua estruturação por meio de metadados, delineando um conceito possível de jornalismo estruturado. A partir de aplicações desenvolvidas tanto por pesquisadores quanto grupos de mídia (notadamente o trabalho realizado pela BBC), um quadro descritivo de possibilidades técnicas é apresentado. Experimentações ou incrementos rotineiros pautados por estas possibilidades apontam, por consequência, novas oportunidades para a prática jornalística num cenário pautado por algoritmos, tecnologias semânticas e conexões entre dados para que informações possam ser reutilizadas em sistemas diferentes, como através de *Application Programming Interface* (API).

> Uma API, em seu nível mais básico, permite que seu produto ou serviço dialogue com outros. Desta forma, uma API permite que você abra seus dados e funcionalidades para outros desenvolvedores, para outras empresas ou mesmo entre departamentos e locais dentro de sua companhia. É cada vez maior a forma como as organizações trocam dados, serviços e recursos complexos, tanto internamente, externamente com parceiros, e abertamente ao público (Lane, 2013).

Metadados: definições

Com a emergência da Web, como plataforma de produção e criação de conteúdo, no qual o jornalismo se estabelece com grande desenvoltura, e seu objetivo de tornar seus conteúdos interoperáveis, os metadados surgem como fator importante para implantação de sistemas que ajudem

2 Versão original: "Structured Journalism is a new form of journalism based on reporting news as structured components into a database, and subsequent retrieval of those structured components to generate news products. The approach is still nascent but it directly addresses several systemic problems facing news producers and news consumers in the digital media ecosystem, and it may potentially facilitate the rebundling of journalism as networks and the creation of consumer-controlled news products with context that extends beyond the article".

na melhora da produção e apresentação da informação jornalística. Por meio deles é possível pensar em "interoperabilidade, a habilidade de dois ou mais sistemas de informação de trocar metadados com a mínima perda de informação"(Neiswender e Montgomery, 2009).

Metadados são informações que permitem rotular, catalogar e descrever dados para serem estruturados de modo a serem compreendidos tanto por humanos quanto por máquinas. É importante ressaltar que não se trata apenas de um acréscimo do código HTML, comuns em processos de otimização de páginas Web, mas sim da descrição de objetos e suas relações com outros conceitos, alcançando um grau de uniformidade na descrição por meio de funções e esquemas(Sicilia; Lytras, 2009).

Metadados são fundamentais para a criação, descrição, organização, atualização, reutilização, validação, recuperação, preservação e recontextualização de objetos de informação. Podem ser descritivos (voltados à descoberta e a identificação de objetos), contextuais ou estruturais (que definem relações entre objetos). Dificilmente metadados são utilizados isoladamente: esquemas de metadados podem especificar o significado de um item, regras de armazenamento e sintaxe.

A origem do termo está nas ciências da computação: "meta" é comumente usado como sinônimo de "sobre" – metadados seriam, portanto, dados sobre dados. Mas não é só isso: eles descrevem como uma organização entende suas entidades, pessoas, lugares, entre outros atributos e suas relações formais. A biblioteconomia destaca-se entre as áreas interessadas na aplicação de metadados (Caplan, 2003). Bibliotecas possuem um rico histórico de organização e gerenciamento de informações a partir de sua estruturação, o que reforça sua importância: se extensos catálogos indexados podem ser controlados com eficácia por estas instituições, por que não utilizar alguns de seus princípios com a Web?

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. A adoção de metadados é apenas uma etapa nesse sentido. Não se trata de um caminho simples: para Polleres e outros (2010), existem poucos dados estruturados em meio a grande quantidade de bases de dados disponíveis, sem contar outro volume de bases inconsistentes ou fora das especificações.

> Contrastando com dados não estruturados, dados estruturados são dados que podem ser facilmente organizados. Independentemente de

> sua simplicidade, a maioria dos especialistas da indústria de dados de hoje estima que dados estruturados correspondem a apenas 20% dos dados disponíveis. São limpos, analíticos e normalmente armazenados em bancos de dados (NEMSCHOFF, 2014, tradução nossa)[3].

Existem modelos e esquemas diversos são propostos para representação, armazenamento e manipulação de metadados. O W3C, consórcio que estabelece boas práticas para a Web, recomenda especificações baseadas em *eXtensible Markup Language* (XML), como a *Resource Description Framework* (RDF), um modelo genérico de dados baseada em gráfico, onde a estrutura de dados se conectam na forma de triplas: sujeito, predicado e objeto. Os três são identificados por URIs; predicados especificam como sujeitos e objetos se relacionam (Bizer; Heath; Berners-Lee, 2009).

Sua evolução é o RDFa: a diferença provocada pelo "a" ao fim da sigla diz respeito a atributos que podem ser definidos no próprio conteúdo, já que o RDF necessita um arquivo separado.Com todo esse potencial surgem dificuldades: a implementação do RDFa provou ser excessivamente complexa para a maioria dos desenvolvedores (Ronallo, 2014).

Assim, outras especificações mais simples tornaram-se mais populares entre os desenvolvedores. É o caso dos microformatos, um tipo simples de marcação usado com frequência para a marcação de eventos, especificações de pessoas ou organizações. Ou ainda os microdados, uma tentativa interessante de adotar as premissas do RDF pelo HTML5. Os microdados utilizam-se de vocabulários para descrever itens – como o Schema.org, criado em conjunto por três empresas do ramo das buscas (Bing, Google e Yahoo!).

No contexto computacional, as representações do conhecimento expressas por linguagens de marcação representam camadas de base. Acima delas, surgem as ontologias: infraestruturas de representação formal do conhecimento em algum domínio de interesse, percebido 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).

Finalmente, agentes inteligentes, programas baseados em operadores que incluem instruções e expressões regulares, permitem o processamento de informação, "interpretação" e troca de dados com outros softwares. Os trabalhos voltados a jornalismo, comunicação e artes (Lammel; Mielniczuk, 2012; Laurentiz, 2010; Ribas, 2007) reforçam o longo caminho a ser trilhado,

3 Versão original: "Contrasting to unstructured data, structured data is data that can be easily organized. Regardless of its simplicity, most experts in today's data industry estimate that structured data accounts for only 20% of the data available. It is clean, analytical and usually stored in databases."

além de exigir uma abordagem multidisciplinar entre comunicação e outras áreas do conhecimento.

> Para o contexto computacional é aquela área que definirá um vocabulário comum entre homens e máquinas para que compartilhem informação... Definir ontologias é tarefa complicada, pois prevê um conjunto de métodos e técnicas automáticas ou semi-automáticas para aquisição de conhecimento utilizando textos, dados estruturados e semiestruturados, esquemas relacionais e outras bases do conhecimento (Laurentiz, 2010).

Metadados e jornalismo

Além dos conteúdos publicados originalmente na Web a partir dos anos 1990, a digitalização de acervos jornalísticos também representam objetos de informação indexáveis. Em 2002, o projeto *ProQuest Historical Newspapers* anunciou a digitalização completa do acervo do *The New York Times*, abrindo um serviço de consulta online a partir de sua primeira edição. Outros jornais históricos norte-americanos, incluindo edições descontinuadas, fazem parte do projeto. No Brasil, apesar de grandes veículos contarem com acervo disponível para consultas, a transformação do processo manual para o informatizado é lento. O exemplo mais eficiente é o do *Acervo Estadão*, que disponibiliza as edições impressas do periódico desde 1875, incluindo períodos censurados durante a ditadura. A recuperação da informação, no entanto, é limitada ao uso de palavras-chave simples.

O *The Guardian*, por sua vez, está na vanguarda das iniciativas relacionadas a jornalismo e computação, além de peça-chave na iniciativa de dados abertos no Reino Unido – como no episódio envolvendo a análise de documentos ligados à despesa de parlamentares britânicos (Daniel; Flew, 2010). O periódico disponibiliza um mecanismo que permite acesso aos artigos publicados no site desde 1999, bem como dados estruturados sobre temas gerais em seu *Data Store*.

O *The New York Times* é outro exemplo. A área de desenvolvedores do jornal inclui *datasets* específicos (atuação de congressistas, gastos em campanhas presidenciais) e algumas informações relacionadas ao acervo (títulos, resumos e links relacionados aos textos do jornal desde 1851, metadados das URLs mais populares). Desde 2009, um vocabulário formado por pessoas, organizações, exemplos e outras descrições é disponibilizado como *linked open data* para utilização em aplicações. Durante os Jogos Olímpicos de 2012, o *hotsite* do evento aproveitou dados oferecidos pelo

Comitê Olímpico Internacional. Informações sobre atletas e resultados de provas, codificados em XML, eram relacionados à cobertura factual.

No Brasil, o caso mais relevante diz respeito à adoção de tecnologias semânticas pelos sites de notícia da *Globo.com*, especialmente a adoção de anotações semânticas manuais (Pena, 2012). Para estruturar um sistema interno de organização, existem funções específicas - como a do Editor de Dados, responsáveis por manter bases de dados atualizadas e organizadas ao longo do tempo(Pena, 2012). Técnicas de anotações semânticas capazes de associar metadados ao conteúdo jornalístico de forma amigável são comuns. O *PundIt*, por exemplo, é uma ferramenta desenvolvida para que qualquer usuário pudesse criar estrutura de dados semânticos em conteúdos Web (Grassi e outros, 2013). Outro exemplo, a ferramenta *Hermes*, foi pensada especificamente para ser um *framework* (modelo) capaz de personalizar notícias a partir de uma combinação de técnicas (Frasincar; Borsje; Levering, 2009). Por fim, os criadores do *Loomp*, software que torna intuitivo o processo de anotações em conteúdos(Luczak-Rösch; Heese, 2009).

São poucos os veículos de mídia que se posicionam declaradamente ao redor do *linked data*, tendo como base a estruturação de objetos de informação com metadados. A versão online da BBC é o que melhor aproveita o das tecnologias semânticas – uma descrição detalhada é apresentada na Seção 4 deste trabalho.

Existem softwares especializados em analisar conteúdos não estruturados e extrair conceitos e metadados de forma automática. É o caso do *OpenCalais*, serviço lançado pela *Thomson Reuters*. Outro projeto nesta linha, bastante audacioso, é o *GDELT*, plataforma que monitora a mídia e acumula informações desde 1979, codificando-as e estruturando-as. Mais do que isso: conecta pessoas, organizações, localizações e temas. Outras plataformas promovem discussões e oferecem ferramentas baseadas em dados e APIs para discutir o futuro da mídia online: é o caso do *Media Cloud*, parceria entre as universidades *Harvard* e *MIT*.

Sistemas de anotação ou métodos de extração poderiam ser utilizados para identificar metadados em acervos desestruturados. Esta possibilidade é favorecida a partir de uma discussão envolvendo a complexidade dos padrões estabelecidos pelo W3C e alternativas propostas por desenvolvedores, como a adoção de microdados interpretados pelos navegadores, associados a esquemas como o Schema.org, proposto por Google, Yahoo e Microsoft (Ronallo, 2014). Pode-se verificar, no entanto, que há um abismo entre as possibilidades técnicas e a aplicação destas.

> Atualmente, metadados para notícias são bastante heterogêneos e difíceis de serem enriquecidos ou detalhados o suficiente para cobrir todo o conhecimento que estes documentos contém. 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 de informação seria extraído (Kallipolitis e outros, 2012, tradução nossa)[4].

O reflexo destes obstáculos pode ser representado pelo projeto *Neptuno*, desenvolvido pelo *Information Retrieval Group*, ligado à escola politécnica da Universidade Autónoma de Madrid. Ele propôs a construção e gestão do acervo digital do jornal *Diari SEGRE*, preocupando-se com a ontologia adequada, a semântica das palavras-chaves, arquitetura e formas de navegação e visualização. Além da redação e duas instituições (Universidad Autónoma de Madrid e Universitat de Lleida), o projeto envolveu ainda uma empresa provedora de tecnologia. Como resultados, além de algumas respostas, surgiram mais perguntas.

> O tamanho e complexidade das informações armazenadas, bem como as limitações de tempo ao catalogar, descrever e ordenar informações de entrada, fazem dos acervos digitais um corpus relativamente desorganizado e difícil de gerenciar. Nesse sentido, compartilham as características e problemas da web, e as soluções propostas para a web semântica são pertinentes aqui (Castells e outros, 2004, tradução nossa)[5].

Já existem formatos de metadados voltados para sistematizar processos de arquivamento e digitalização de informações jornalísticas. Destaque para o NITF (*News Industry Text Format*), 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

4 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."

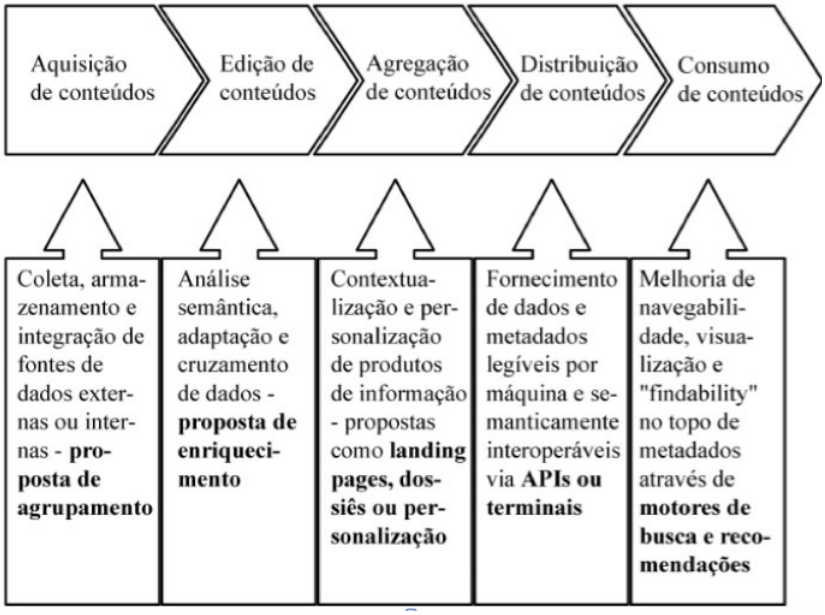
5 Versão original: "The size and complexity of the stored information, and the time limitations for cataloguing, describing and ordering the incoming information, make newspaper archives a relatively disorganised and difficult to manage corpus. In this sense, they share many of the characteristics and problems of the WWW, and therefore the solutions proposed in the Semantic Web vision are pertinent here."

metadados e ontologias a objetos como textos, fotografias, áudios e vídeos, maximizando a interoperabilidade de informação e produzindo conexões significativas(Troncy, 2008).Mesmo sendo uma iniciativa conhecida e adotada por grandes jornais e agências de notícias, o pesquisador Tassilo Pellegrini identifica um obstáculo em sua utilização.

> A adoção prática dos códigos do IPTC entre a indústria de notícias e seu uso em sistemas de gerenciamento de conteúdo editorial e aplicativos é limitada a uma pequena fração do vocabulário existente, o que por um lado é um forte indicador de especificações em excesso e, por outro, a falta de uma elaborada "cultura de metadados" na gestão da informação dentro dos fluxos de trabalho editoriais (Pellegrini, 2012, tradução nossa)[6].

Pellegrini menciona Michael Porter, professor de Harvard e referência no universo de economia e negócios, para adaptar o conceito de "cadeia de valor" à produção de notícias. Seguindo uma lógica de produção, cada etapa pode ser reforçada por metadados. A Figura 2 ilustra potenciais contribuições de valor, por meio dos *linked data*, nessas etapas do processo de produção e distribuição de conteúdo.

■ **Figura 2: cadeia de valor por meio de** *linked data*



Fonte: Pellegrini, 2012, p. 127

6 Versão original: "The practical uptake of the IPTC codes among the news industry and its usage in editorial content management systems and applications is limited to a small fraction of the existing vocabulary which is a strong indicator for over-specification on the one side and a lack of an elaborated "metadata culture" in the management of information within editorial workflows on the other."

O universo de dados abertos estruturados disponíveis (como*DBPedia* ou *Freebase*) representa uma oportunidade para o processo de aquisição de conteúdos, onde profissionais coletam, armazenam e relacionam itens que vão se tornar notícia. Mas é no processo de edição, por meio de técnicas de anotação semântica, que a informação pode ser enriquecida. Aqui, a discussão dos processos editoriais torna-se imprescindível. A terceira etapa diz respeito a contextualização e personalização de conteúdos, o que inclui modelos de metadados relacionados ao comportamento do usuário. Na etapa de distribuição, ocorre o diálogo com máquinas, especialmente por meio de APIs. Finalmente, no consumo de conteúdos, usuários interagem da forma mais agradável possível.

O aumento na disponibilidade de dados estruturados como parte da estratégia de governos, organizações ou iniciativas colaborativas faz surgir uma questão: de que forma a indústria da mídia pode se beneficiar deste processo? Em 2010, o boletim do IPTC (Mirror, 2010)repercutiu a seguinte questão entre seus leitores: "a mídia consegue utilizar *linked data* por um futuro mais forte"? "Responder a pergunta 'linked data pode funcionar' é apenas o começo: 'existe um business case para ele' é o complemento dessa questão", observa o texto. Um olhar mais detalhado em redações, segundo Pellegrini, revela um descompasso entre debates científicos e a utilização de metadados na indústria da mídia.

> 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)[7].

A partir do interesse em adicionar valor à notícia, das ferramentas semânticas existentes e da constatação de projetos desenvolvidos, é possível identificar procedimentos técnicos capazes de estruturar objetos de informação por meio de metadados. A adoção destes instrumentos, em maior ou menor grau a partir dos obstáculos, permite estruturar a informação jornalística na Web, contribuindo para uma análise mais adequada de veículos que experimentam estas práticas (Palacios, 2011) e encaminhando-a para o patamar de sistema. O

7 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."

Quadro 1 sintetiza estas possibilidades, relacionando-as a partir da observação descrita anteriormente. Entre os grupos de mídia observados, a BBC pode ser reconhecido como referência, capaz inclusive de determinar os parâmetros.

■ Quadro 1: Relação entre procedimentos técnicos e grupos de mídia

	The New York Times (EUA)	BBC (UK)	The Guardian (UK)	Globo.com (BR)
Aproveitar dados externos com informações sobre conceitos (sujeitos, objetos ou lugares) para enriquecer suas próprias bases	Hotsite dos Jogos Olímpicos de 2012	Relação de músicas e programas por meio da DBPedia; projeto BBC Wildlife	Data Store: dados estruturados sobre temas gerais	
Codificar fragmentos de informação manualmente, a partir do CMS, utilizando anotações semânticas		Anotações manuais do canal BBC Sports na Copa de 2010 e nos Jogos de 2012		Projeto interno de anotações semânticas em seu CMS
Analisar (parsing) e codificar fragmentos de informação (páginas, bases de dados) com metadados por meio de software (codificação automática)		Projeto The News Juicer do BBC News Labs		
Oferecer conceitos ou conteúdos por meio de uma API, permitindo a criação e interoperabilidade de dados para múltiplos dispositivos e plataformas	Datasets sobre Congresso dos EUA e informações do acervo	Projeto BBC Things	API para acesso aos artigos do site e ao Data Store	
Relacionar conceitos (sujeitos, objetos ou lugares) por meio de triplas usando tecnologias como RDF		Ontologia específica para cobertura das Eleições 2014		
Desenvolver agentes inteligentes capazes de reconhecer e aproveitar o ecossistema de *linked open data* (LOD)		Desenvolvimento do algoritmo Datastringer		

Fonte: produzido pelo autor

Uso de metadados pela bbc

A BBC, *British Broadcast Corporation*, utiliza metadados associados a ferramentas semânticas desde 2009, sendo o primeiro grupo de mídia a fazê-lo. Já identificando uma grande quantidade de conteúdo online (incluindo notícias e entretenimento), mas que não dialogavam entre si,

iniciou projetos que relacionavam internamente programas e músicas utilizando a DBPedia como vocabulário controlado (Kobilarov e outros, 2009).

No âmbito das notícias, a BBC também já enriquece informações utilizando metadados por meio de um sistema de publicação e gerenciamento de conteúdos – a começar com a organização do material relacionado à editoria "esporte" durante a Copa de 2010. As 700 páginas agregadoras de entrada, incluindo informações sobre grupos, seleções e jogadores, eram criadas a partir das informações codificadas manualmente em cada notícia publicada no sistema, baseado em RDF e *linked data*. A experiência resultou na continuidade do processo nas notícias sobre futebol do site *BBC Sports*. Esforço ampliado durante os Jogos Olímpicos de 2012, em Londres.

Outro exemplo pioneiro, o site *BBC Wildlife,* reúne informações sobre animais selvagens, plantas, entre outros dados do mundo natural. Para cada espécie, há uma página única, gerada dinamicamente, a partir de uma base de dados estruturada– que permite ainda a sugestão de conteúdos relacionados. Tornou-se ainda 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).

A cultura de metadados, adaptação e reutilização de conteúdos iniciada por estes projetos, tendo como premissa a identificação de cada item de interesse da BBC em uma URI específica, contribuiu para impulsionar a divisão *BBC Future Media*, guarda-chuva das inovações associadas aos serviços digitais, criada em 2011.Um ano depois, em 2012, a divisão *BBC Connected Studio* lançou um projeto de inovação para explorar oportunidades para seus produtos noticiosos a partir de tecnologias criativas: o *BBC News Labs*. Tratam-se das áreas mais envolvidas em desenvolvimento de aplicações que culminam com tecnologias de *linked data*. Um dos projetos desenvolvidos pela equipe do Labs, batizado de *The News Juicer*, consistiu em um protótipo para extração de conceitos, seu relacionamento com a DBPedia e anotação automática nos arquivos da BBC.

Em abril de 2014, a *BBC Future Media* apresentou a nova versão de suas ontologias, base para sua plataforma de *linked data*. O site procurou organizar de maneira apropriada o resultado dos projetos e esquemas hospedados na organização desde suas primeiras experiências. Dessa forma, mantém-se inserida no ecossistema de *Linked Open Data* (LOD). Como resultado deste processo, o serviço *BBC Things*, lançado em setembro de 2014, 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.

A expertise em arquitetura de dados estimula o desenvolvimento de novas ações, como a cobertura das eleições locais britânicas em maio de 2014. Para viabilizar as anotações semânticas no conteúdo, foi desenvolvida uma ontologia específica para a cobertura política: candidatos, partidos, entre outras instâncias precisam ter sua própria URI de acordo com os padrões do W3C, bem como relações estabelecidas entre objetos. Com estas amarrações e ferramentas, a equipe é capaz descobrir quantas vezes um determinado partido foi mencionado durante a cobertura das eleições. Ou ainda quais expressões e personagens aparecem com mais frequência ao lado de cada um deles.

Por meio do laboratório, equipes multidisciplinares aprendem novos conceitos e tomam decisões a partir dos protótipos desenvolvidos, aprendendo sobre novas tecnologias e construindo um legado de informações estruturadas em suas bases de dados. O algoritmo *Datastringer* é um dos exemplos mais recentes: ele que permite ao jornalista monitorar com facilidade bases de dados externas a partir de critérios definidos por uma pauta(Shearer; Simon; Geiger, 2014). Além deste histórico revelar a capacidade de inovação da BBC, um manifesto ao jornalismo estruturado reforça a escolha deste veículo como referência neste campo:

> Acreditamos que o jornalismo estruturado tornará a BBC news mais inteligente, eficiente e envolvente. Acreditamos que o jornalismo estruturado permitirá nosso engajamento com o mundo em formas que reconhecem sua verdadeira complexidade. Finalmente, acreditamos que o jornalismo estruturado nos tornará melhores jornalistas - aqueles que têm o poder de mostrar seu trabalho, abrir seus dados, permitir que o público contribua significativamente e criar uma sociedade mais informada (A MANIFESTO FOR STRUCTURED JOURNALISM, 2015, tradução nossa)[8].

Considerações finais

Diante da possibilidade de qualquer pessoa se aprofundar em fontes de dados e encontrar informação relevante, o jornalismo produzido com o auxílio de bases de dados representa o acesso das ferramentas, técnicas e métodos a qualquer interessado que deseja aprender, algo anteriormente utilizado

8 Versão original: "We believe that structured journalism will make BBC News smarter, more efficient, and more engaging. We believe that structured journalism will allow us all to engage with the world in ways that acknowledges its true complexity. And, finally, we believe structured journalism will make better journalists - ones who are empowered to show their work, open their data, allow the public to meaningfully contribute, and create a more informed society."

exclusivamente por especialistas: repórteres investigativos, cientistas sociais, estatísticos ou analistas. Práticas podem ser compreendidas por meio de cursos livres, sites especializados, encontros denominados *hack day*. Isso representa uma transformação no Modelo Padrão de Jornalismo, desenhado por Walter Lippman nos anos 1920, bem como uma reconfiguração da profissão(Lima Junior, 2012).

Com a emergência da Web como uma plataforma,está claro que "as bases de dados são consideradas plataformas tecnológicas fundamentais para o desenvolvimento do jornalismo contemporâneo em redes digitais"(Lammel; Mielniczuk, 2012). Da mesma forma, não há como ignorar o protagonismo dos metadados na construção de um jornalismo estruturado. Consequentemente, a utilização de padrões semânticos na Web, a adoção dos princípios do *Linked Data* e a disponibilização de APIs representam um trajeto árduo, mas possível, para estimular práticas multidisciplinares e buscar práticas inovadoras em redações.

As práticas e experimentações produzidas por veículos de mídia, especialmente a BBC, indicam a procura pelo aperfeiçoamento do processo de armazenamento e recuperação da informação em bases de dados, estabelecendo conexões entre computação e jornalismo por meio de ferramentas semânticas. Além disso, reforça a necessidade de diálogo entre estas áreas do conhecimento: isoladamente, os profissionais de mídia terão dificuldade em construir estas conexões. Além do estímulo à formação de equipes multidisciplinares, a opção da BBC por dados e plataformas abertas permitem seu apoderamento por qualquer usuário, ampliando a possibilidade de aplicações e, consequentemente, a relevância deste conteúdo.

Por conta do caráter exploratório dos veículos de mídia proposto neste artigo, não se trata de uma avaliação do melhor ou pior trabalho na utilização de metadados como fator de interoperabilidade em sistemas informativos por organizações de mídia, seja para melhora da produção jornalística ou para automatização de produtos noticiosos distribuídos em diferentes plataformas, principalmente no ambiente dos dispositivos móveis conectados.

A avaliação nesse nível comparativo não é possível, pois muitos desses sistemas estão rodando internamente (*privated access*), não permitindo acesso aos pesquisadores ao seu funcionamento e modelagem, ou as configurações tecnológicas que permitem apresentar o resultado das relações entre *datasets* são imperceptíveis ao usuário através da interface na qual acessa o conteúdo noticioso, mas que proporcionam um ganho informativo considerável.

Assim, o artigo sinaliza quais são os esforços dos grupos de mídia mencionados na busca implantar soluções para tratamento de dados e informações através de metadados e sistemas interoperáveis, buscando fornecer para o produtor de informação noticiosa, o jornalista, melhores opções para a construção

da narrativa, enriquecendo o material jornalístico produzido e otimizando o trabalho de armazenamento, recuperação, relacionamento, distribuição de dados em função da melhora dos produtos jornalísticos espalhados por diversas plataformas digitais, mas tendo como base uma única modelagem tecnológica.

A utilização de sistemas com base em metadados para construção a informação jornalística, seja na ponta da produção de narrativas produzidas por jornalistas ou na estrutura máquina para máquina (automatizados), podem ser fatores de diferenciação na qualidade informativa entre grupos de mídia, pois esses sistemas podem enriquecer o conteúdo jornalístico com informações não-triviais ao produtor e ao consumidor de notícias.

Referências

A MANIFESTO FOR structured journalism. **BBC News Labs**, Londres, 7 jul. 2015. Disponível em: http://bbcnewslabs.co.uk/2015/07/07/a-manifesto-for-structured-journalism. Acesso em: 6 nov. 2015.

BARBOSA, S.; TORRES, V. O paradigma "Jornalismo Digital em Base de Dados": modos de narrar, formatos e visualização para conteúdos. **Revista Galáxia**, São Paulo, n. 25, p. 152–164, 2013.

BELL, T. Where the semantic web stumbled, linked data will succeed. **Radar O'Reilly**, Sebastopol, Califórnia, [s.l.], 15 nov. 2010. Disponível em: http://radar.oreilly.com/2010/11/semantic-web-linked-data.html. Acesso em: 25 mar. 2015.

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The Semantic Web. **Scientific American**, New York, n. May 2001, p. 34–43, 2001. https://doi.org/10.1038/scientificamerican0501-34

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.

CAPLAN, P. **Metadata Fundamentals for All Librarians**. Chicago: American Library Association, 2003.

CASWELL, D. A.; RUSSELL, F.; ADAIR, B. Editorial aspects of reporting into structured narratives. In: COMPUTATION+JOURNALISM SYMPOSIUM, 2015, New York. **Anais...** New York: [s. n.], 2015.

CASTELLS, P.; PERDRIX, F.; PULIDO, E. Neptuno: semantic web technologies for a digital newspaper archive. In: **The Semantic Web: Research and Applications**. Athens: Springer Berlin Heidelberg, 2004. https://doi.org/10.1007/978-3-540-25956-5_31

CHUA, Reg. Structured Journalism. **(Re)Structuring Journalism**, [s.l.], 2 aug. 2010. Disponível em: https://structureofnews.wordpress.com/structured-journalism. Acesso em: 6 nov. 2015.

CORRÊA, E. N. S.; BERTOCCHI, D. A cena cibercultural do jornalismo contemporâneo: web semântica, algoritmos, aplicativos e curadoria. **Matrizes**, São Paulo, v. 5, n. 2, p. 123–144, 2012.

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

GILLIAND, A. Setting the Stage. In:BACA, M. **Introduction to Metadata**. Los Angeles, CA: Getty Publications, 2008.

GRASSI, M. e outros. Pundit: augmenting web contents with semantics. **Literary and Linguistic Computing**, v. 28, n. 4, p. 640–659, 18 set. 2013. https://doi.org/10.1093/llc/fqt060

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. https://doi.org/10.1016/j.knosys.2011.12.007

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

LAMMEL, I.; MIELNICZUK, L. Aplicação da Web Semântica no jornalismo. **Estudos em Jornalismo e Mídia**, Florianópolis, v. 9, n. 1, p. 180–195, 5 jul. 2012. https://doi.org/10.5007/1984-6924.2012v9n1p180

LANE, K. **What Is An API**. Disponível em: https://s3.amazonaws.com/kinlane-productions/whitepapers/API+Evangelist+-+API+101.pdf. Acesso em: 20 fev. 2017.

LAURENTIZ, S. Tags e metatags? De Ted Nelson a Tim Berners-Lee. **Revista Porto Arte**, Porto Alegre, v. 17, n. 28, p. 17–33, 2010.

LIMA JUNIOR, W. T. Big Data, Jornalismo Computacional e Data Journalism: estrutura, pensamento e prática profissional na Web de dados. **Estudos em Comunicação**, Curitiba, n. 12, p. 207–222, 2012.

LLOYD, Alexis. The Future of News is not an Article. **The New York Times Research and Development**, Nova York, 20 out. 2015. Disponível em http://nytlabs.com/blog/2015/10/20/particles. Acesso em: 6 nov. 2015.

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

MIRROR, I. Can news media use linked data for a stronger future? **IPTC**, Londres, n. 1, p. 2–7, fev. 2010.

NEISWENDER, C; MONTGOMERY, E. Metadata Interoperability — What Is It, and Why Is It Important? **The MMI Guides: Navigating the World of Marine Metadata**, [s.l.], 2009. Disponível em: http://marinemetadata.org/guides/mdataintro/mdatainteroperability. Acesso em: 6 nov 2015.

NEMSCHOFF, M. A Quick Guide to Structured and Unstructured Data. **SmartDataCollective**, [s. l.], 28 jun. 2014. Disponível em: http://smartdatacollective.com/michelenemschoff/206391/quick-guide-structured-and-unstructured-data. Acesso em: 28 mar. 2015.

PALACIOS, M. A memória como critério de aferição de qualidade no ciberjornalismo: alguns apontamentos. **Revista FAMECOS**, Porto Alegre, v. 37, 2008.

_____. **Ferramentas para Análise de Qualidade no Ciberjornalismo (Volume 1: Modelos)**. Covilhã/Portugal: LabCom Books, 2011.

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

PENA, R. A. P. **Suporte semântico à publicação de conteúdo jornalístico na Web.** 2012. 105 f. Dissertação (Mestrado em Comunicação Social) – Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2012.

POLLERES, A. e outros. Can we ever catch up with the Web? IOS Press, **Lansdale: Pennsylvania/EUA**, p. 1–5, 2010.

RENNISON, E. Galaxy of News: An approach to visualizing and understanding expansive news landscapes. Proceedings of the 7th annual ACM symposium on User interface software and technology. **Anais...** New York, NY: ACM, 1994.

RIBAS, B. Web Semântica e produção de notícias: Anotações para o estudo da aplicação da tecnologia ao campo do Jornalismo. V Encontro Nacional de Pesquisadores em Jornalismo - SBPJor. **Anais**...Aracaju: 2007.

RONALLO, J. HTML5 Microdata and Schema.org. **The Code4Lib Journal**, Los Angeles/EUA, n. 16, p. 1–17, 2014.

SARACEVIC, T. Ciência da informação: origem, evolução e relações. **Perspectivas em Ciencia da Informação**, Belo Horizonte, v. 1, n. 1, p. 41–62, 1996.

SHEARER, M.; SIMON, B.; GEIGER, C. Datastringer: easy dataset monitoring for journalists. In: COMPUTATION+JOURNALISM SYMPOSIUM. **Anais**...New York, NY: Columbia Journalism Schools, 2014.

SICILIA, M.-A.; LYTRAS, M. **Metadata and Semantics**. New York, NY: Springer Science+Business Media, LLC, 2009. https://doi.org/10.1007/978-0-387-77745-0

TRONCY, R. Bringing the IPTC news architecture into the semantic web. 7TH INTERNATIONAL SEMANTIC WEB CONFERENCE. **Anais**... Karlsruhe, Germany: 2008.

Recebido em: 3/11/2016
Aceito em: 20/3/2017

Endereço dos autores:
Walter Teixeira Lima Junior <digital@walterlima.jor.br>
Universidade Federal do Amapá
Rodovia Juscelino Kubitscheck, km 2, s/nº – Jardim Marco Zero
68903-419 – Macapá (AP) – Brasil

Andre Rosa de Oliveira <andrerosa.jor@gmail.com>
Faculdades Integradas Rio Branco
Avenida José Maria de Faria, 111 – Lapa de Baixo
05038-190 – São Paulo (SP) – Brasil