UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Campus de Marília

OFICINA UNIVERSITÁRIA
CULTURA ACADÊMICA Editora

Software Livre para a Análise de Informação Estruturada:

Análise sob a Perspectiva do Conhecimento Aberto

Lillian Maria Araújo de Rezende Álvares
Luc Quoniam
Kira Tarapanoff

Como citar: ÁLVARES, L. M. A. D. R.; QUONIAM, L. TARAPANOFF, K. *Software* Livre para a Análise de Informação Estruturada: Análise sob a Perspectiva do Conhecimento Aberto. *In* : VALENTIM, M. L. P.; MÁS-BASNUEVO, A. (org.). **Inteligência organizacional**. Marília: Oficina Universitária; São Paulo: Cultura Acadêmica, 2015. p.35-52. DOI: https://doi.org/10.36311/2015.978-85-7983-678-7.p35-52

Capítulo 2
Software Livre para a Análise de Informação Estruturada: Análise sob a Perspectiva do Conhecimento Aberto

Lillian Maria Araújo de Rezende Álvares
Luc Quoniam
Kira Tarapanoff

1 Introdução

Um *software* é considerado de livre uso quando respeita a liberdade de executar, copiar, distribuir, estudar, mudar e melhorar. Ele garante ao usuário o acesso ao código fonte (*source code*) a fim de modificar e compartilhar o programa. Para a *Free Software Foundation* (FSF, 1983) [em edição consultada de 2006] este deve respeitar quatro liberdades essenciais: a liberdade de executar o programa como desejar, para qualquer propósito; a liberdade de estudar como o programa funciona e adaptá-lo às suas necessidades; a liberdade de redistribuir cópias originais; e a liberdade de distribuir cópias das versões modificadas.

A disseminação do *software* livre promove a possibilidade de oportunidades iguais para todos, indo ao encontro de agendas internacionais de sustentabilidade. Como, a Década das Nações Unidas para um desenvolvimento sustentável 2005-2014, que tinha como objetivo integrar valores inerentes ao desenvolvimento sustentável, em todos os aspectos da aprendizagem, com o intuito de fomentar mudanças de comportamento que permitissem criar uma sociedade sustentável e justa para todos[1].

Software livre não significa necessariamente um *software* sem custo. Nem sempre as soluções apresentadas vão se adequar plenamente às

https://doi.org/10.36311/2015.978-85-7983-678-7.p35-52

necessidades institucionais, devendo ser customizadas e complementadas. Igualmente, nem sempre será fácil a sua implantação, exigindo a participação de especialistas. Essas considerações levam a refletir sobre as vantagens financeiras do uso do *software* livre. Em princípio, pode-se afirmar que sua utilização é vantajosa, pois no mínimo terá a garantia da licença perpétua, o que o *software* proprietário raramente oferece. Também se deve considerar a customização perfeita, que minimiza custos adicionais com outros elementos de ajuste. E finalmente, a questão da responsabilidade social, uma vez que a comunidade se beneficia com as novas versões melhoradas, ampliadas e ajustadas a interesses diversos.

A história do *software* livre tem início com Richard Stallman por ocasião do anúncio do Projeto GNU nos Anos 1980. Ele, intensamente dedicado à liberdade na computação, fundou a *Free Software Foundation*, lançando o Movimento do *Software* Livre que com o tempo transformaria a indústria de *software* (STALLMAN et al., 2006).

Nos Anos 1990 o movimento intensificou-se com o lançamento do Linux, de Linus Torvalds. Centenas de desenvolvedores se juntaram ao projeto para integrar o sistema GNU ao Linux. Na mesma década, Eric Raymond e Bruce Perens criam a *Open Source Initiative* (OSI), defendendo a adoção do *software* livre também por razões técnicas e sugerindo o uso da expressão *open source* ao invés de *free software*, evitando a ambiguidade do termo *free*, que pode significar tanto livre quanto gratuito na língua inglesa. Em português é traduzido como *software* livre, *software* de código aberto ou *software* aberto.

A definição de *software* livre da FSF concentra-se prioritariamente na questão da liberdade do usuário, já a definição de *software* aberto da OSI abrange as mesmas questões, mas inclui alguns elementos técnicos adicionais. Para se definir um *software* de código livre devem ser observados os 10 requisitos a seguir (OPEN..., 1998):

1. a licença não pode restringir ninguém, proibindo que se venda, ou doe o *software* a terceiros;
2. o programa precisa obrigatoriamente incluir o código-fonte e permitir sua distribuição tanto quanto do programa já compilado;

3. a licença deve permitir modificações e obras derivadas que possam ser redistribuídas dentro dos mesmos termos da licença original;
4. a licença pode proibir que se distribua o código-fonte original modificado desde que a licença permita a distribuição de *patch files* com a finalidade de modificar o programa em tempo de construção;
5. a licença não pode discriminar contra pessoas ou grupos;
6. a licença não pode restringir os usuários de fazer uso do programa em uma área específica;
7. os direitos associados ao programa por meio da licença são automaticamente repassados a todas as pessoas às quais o programa é redistribuído sem a necessidade de definição ou aceitação de uma nova licença;
8. os direitos associados a um programa não dependem de qual distribuição em particular aquele programa está inserido; se o programa é retirado de uma distribuição, os direitos garantidos por sua licença continuam valendo;
9. a licença não pode colocar restrições em relação a outros programas que sejam distribuídos junto com o *software* em questão;
10. nenhuma exigência da licença pode ser específica a uma determinada tecnologia ou estilo de interface.

Apesar de as diferenças filosóficas entre o movimento do *software* livre e o movimento do código aberto, basicamente caracterizada pelo lado social da FSF e pelo lado técnico e de mercado da OSI, as definições oficiais de ambas referem-se basicamente ao mesmo ideal: a filosofia de desenvolvimento da OSI está centrada no que Raymond (1999) chamou de Modelo Bazar, com inúmeras e diferentes abordagens no desenvolvimento do *software*, ao contrário do tradicional modelo de engenharia de *software*, desenvolvido de forma centralizada e isoladamente.

O modelo Bazar, resumidamente, compreende que os usuários devem ser tratados como codesenvolvedores, incentivados a apresentar adições ao *software*, correções de código, relatórios de *bugs* e documentação, entre outros. Acredita que com mais desenvolvedores, a evolução do *software* acontece aceleradamente, incluindo a identificação de erros

e as soluções de correção. Cabe destacar que cada ambiente de desenvolvimento significa mais ambientes de testes, maximizando a identificação de latentes problemas nos programas. O modelo defende que a primeira versão do *software* deve ser lançada tão cedo quanto possível, de modo a aumentar as chances de encontrar codesenvolvedores. Deve disponibilizar pelo menos duas versões do programa, uma com mais recursos (e, portanto, mais erros, também chamada de versão de desenvolvimento) e uma mais estável, portanto com menos recursos e menos erros. A primeira, para aqueles usuários que desejam usar imediatamente os recursos mais recentes e estão dispostos a aceitar o risco de usar um código que ainda não está completamente testado.

2 Conhecimento livre ou conhecimento aberto

De fato, ao tratar de *software* livre, estamos tratando efetivamente de conhecimento livre ou conhecimento aberto. Na definição da *Open Knowledge Foudation*[2] (2006), é o conhecimento que pode ser adquirido, interpretado e aplicado livremente. Ele pode ser reformulado de acordo com as necessidades de alguém, e compartilhado com os outros para benefício da comunidade.

Diz respeito a quaisquer conteúdos, informações ou dados que as pessoas são livres para usar, reutilizar e redistribuir, sem qualquer restrição legal, tecnológica ou social. Deve capacitar a todos, permitindo que as pessoas trabalhem em conjunto para enfrentar os desafios locais e globais, compreendendo o mundo, expondo ineficiências, combatendo a desigualdade e ainda, responsabilizando governos e empresas a prestar contas de suas ações. A convicção por trás do movimento do conhecimento livre é de que ele deve ser acessível e compartilhável, sem restrições.

Esta orientação baseia-se no entendimento de que ativos intangíveis, como o conhecimento, são propulsores de mudanças sociais. Estas propiciadas pelo tipo de conhecimento que conduz à descoberta, onde o processo de aprendizagem é continuo; onde o conhecimento é produzido na interação com o mundo; onde cada um de nós é produtor de conhecimento; e onde a finalidade do conhecimento é transformar o mundo e cada um de nós (LEV, 2001).

Os defensores do conhecimento aberto acreditam que a liberdade de acesso ao conhecimento está sob ameaça em virtude das tentativas de restringir ou controlar o compartilhamento de informações na internet. A conceituação de conhecimento livre foi formulada com base na definição de *software* livre da *Free Software Foundation*, considerando o sucesso dos sistemas colaborativos que permitiram o desenvolvimento do *software* livre e a edição da maior enciclopédia de todos os tempos, em todas as línguas, a *Wikipedia*.

A mesma *Open Knowledge* defende que dados abertos são os blocos da construção do conhecimento aberto. De fato, outras iniciativas importantes relativas ao conhecimento livre são os dados abertos da pesquisa científica, entendidos como aqueles que estão disponíveis gratuitamente na internet, permitindo a qualquer usuário baixar, copiar, analisar, reprocessar, fazer a captura por *software* ou utilizá-los para qualquer outra finalidade, sem barreiras financeiras, legais ou técnicas além daquelas que dizem respeito à própria internet. Para este fim os dados que dão origem às publicações científicas devem ser explicitamente colocados em domínio público (MURRAY-RUST et al., 2014).

A reutilização dos dados de pesquisa é uma forma de compartilhamento que se insere na gênese da ciência aberta como o compartilhamento ideal de conhecimentos, recursos educacionais e informacionais que são viabilizados por poderosas infraestruturas eletrônicas, transpondo fronteiras institucionais de disciplinas científicas e de nações. Supõe-se que isso pode ser um passo importante para estimular, desde cedo, professores e alunos em suas carreiras de pesquisadores.

Deriva da aprendizagem aberta, a aprendizagem social construcionista, que tem como filosofia o uso de recursos não sujeitos a licenças excessivamente restritivas (XAVIER, 2005). O termo é associado a visões da liberdade de aprender, que seja considerada adequada aos países onde os sistemas de ensino não são capazes de atender a todas as necessidades da sociedade, e incentivam a sociedade civil a tomar a iniciativa de aumentar a oferta e a qualidade dos sistemas de ensino público.

Uma breve e não exaustiva linha do tempo com manifestações sobre o conhecimento livre pode ser apresentada, entre tantas iniciativas ocorridas, da seguinte maneira:

- 1909: a publicação Hind Swaraj de Mahatma Gandhi é reconhecida como o projeto intelectual do movimento de libertação da Índia. No livro, o autor afirma que não há direitos reservados para o conhecimento (THE SWARAJ..., 1999);
- 1948: o artigo 27 da Declaração dos Direitos Humanos assegura que todos têm o direito de participar livremente da vida cultural da comunidade, de fruir as artes e de participar do progresso científico e de seus benefícios e que toda pessoa tem direito à proteção dos interesses morais e materiais decorrentes de qualquer produção científica, literária ou artística da qual seja autor (ORGANIZAÇÃO..., 1948);
- 1954: Na coletânea *Selected poems*, Mark Van Doren defende o direito do homem ao conhecimento e ao uso livre do mesmo (POETRY..., 2015);
- 1995: a *Knowledge Ecology International*, organização não governamental, defende o movimento A2K[3] (Acesso ao Conhecimento) que se preocupa com leis de direitos autorais e outros regulamentos. Tem a perspectiva social, tratando o tema como acesso aos bens de conhecimento, incluindo acesso à informação, à educação, à saúde pública (em torno de patentes e medicamentos) (KNOWLEDGE..., 1995);
- 2000: Lawrence Lessig inicia a publicação de uma série de livros sobre o conhecimento aberto: *Code and other laws of cyberspace* (2000); *The future of ideas* (2001); *free culture* (2004); *Code: version 2.0* (2006) (HARVARD..., [s.d. a);
- 2001: Criação da *Creative Commons,* com o lançamento do seu primeiro conjunto de licenças gratuitas de direitos autorais. As licenças são inspiradas na Licença GNU da *Free Software Foundation* (GNU GPL) ao lado de uma plataforma de aplicações *Web* (SCIENCE..., 2015);
- 2001: Criação por Patrick Brown da *Public Library of Science*[4] (PLOS), projeto sem fins lucrativos que tem o objetivo de criar uma biblioteca de revistas científicas e publicações afins dentro do modelo

de licenciamento de conteúdo aberto, fazendo uso, especificamente, da *Creative Commons* (PUBLIC..., 2015);

- 2002: Editoras acadêmicas começam a pensar em acesso aberto e muitos lançam conteúdo científico sob a licença *Creative Commons. O Directory of Open Access Journals* enumera muitos, mas com variados graus de liberdade de acesso (DIRECTORY..., 2015);
- 2002: A Budapest *Open Access Initiative* (BOAI) formaliza o movimento do acesso aberto para a pesquisa em todos os campos. Este pequeno grupo de indivíduos é reconhecido como fundadores do movimento de acesso aberto para a ciência, por meio da distribuição mundial eletrônica da literatura científica revisada por pares, e acesso totalmente gratuito e irrestrito a ele para todos os cientistas, estudiosos, professores, estudantes e outras mentes curiosas. Por ocasião do 10º aniversário da iniciativa, em 2012 os compromissos foram reafirmados (BAILEY JR., 2006);
- 2002: Peter Suber dá início a uma série de publicações sobre acesso aberto ao conhecimento como *Open access to the scientific journal literature* (2002); *Open access: other ways* (2003); *Open access, impact, and demand* (2005) (HARVARD..., 2013);
- 2003: *Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities*, a mais significativa iniciativa internacional de acesso aberto ao conhecimento. Em 2007 mais de 240 organizações científicas haviam assinado a declaração e em 2013, 451 (BAILEY JR., 2006);
- 2003: *Bethesda Statement on Open Access Publishing* se une à *Budapest Open Access Initiative* e à Declaração de Berlim sobre o Acesso Livre ao Conhecimento nas Ciências e Humanidades e reafirma o acesso aberto como o termo para descrever iniciativas para tornar a informação mais ampla e facilmente disponível. Ela complementa as declarações e acrescenta a condição de que os praticantes do acesso aberto poderão reutilizar o conhecimento, incluindo o direito de fazer obras derivadas (BAILEY JR., 2006);
- 2004: Yochai Benkler publica dois títulos para falar do novo modelo de produção: *Coase's penguin or, Linux and the nature of the firm* (2004) e *The wealth of networks* (2006) (HARVARD..., [s.d. b);

- 2006: divulgada a definição de conhecimento livre, inspirada na definição de *software* Livre e por uma postagem no *blog* de Jimmy Wales (cofundador e promotor da *Wikipedia*) que afirma que conhecimento livre requer *software* e arquivos livres (JIMMY, 2015);
- 2007: Charlotte Hess e Elinor Ostrom editam uma coletânea intitulada *Understanding knowledge as a commons* que discute o movimento de acesso aberto, a estrutura de código aberto ao conhecimento científico, e os efeitos sobre comunicação científica, entre outros temas relativos (P2P..., 2015);
- 2007: Kim Tucker lança o *Say Libre* ensaio que trata dos Recursos Educacionais Abertos. Cabe destacar a visão que o autor quer compartilhar: conhecimento para todos, liberdade de aprender, no sentido de sabedoria coletiva, permitir que as pessoas se fortaleçam com o conhecimento e compartilhá-lo para o benefício da comunidade[9] (Say..., 2015);
- 2011: Aaron Hillel Swartz disponibiliza livremente os arquivos da editora científica JSTOR na internet, por não acreditar no modelo das grandes revistas científicas que compensam financeiramente as editoras, e não os autores, e de cobrar o acesso aos artigos, limitando assim o acesso para finalidades acadêmicas. Por esse crime foi preso e não chegou a ser julgado, pois as acusações contra ele foram retiradas após sua morte por aparente suicídio (REMEMBER..., 2015).

Feita a contextualização do cenário em que se insere o trabalho, seu foco recai na identificação de *software* de acesso livre e código aberto com objetivo de promover o conhecimento aberto e o aprendizado contínuo no tratamento e análise de informações estruturadas em bases de dados.

3 Análise de informações estruturadas

A terminologia Informação Estruturada pode ter várias definições e diferentes aplicações. Para esse trabalho, o conceito adotado segue o expresso por Cavalcante e Valentim (2010), baseado em um padrão formal pré-estabelecido. Quando se relacionam com o padrão, são chamadas informações estruturadas e quando não se relacionam a nenhum padrão formal, são chamadas informações não estruturadas.

Deve-se considerar que o padrão está relacionado a uma base de conhecimento específica e, portanto, não é possível falar genericamente de estrutura da informação. Isto é, a noção intangível de estrutura precisa ser explicitada a partir das unidades fundamentais da informação: o documento e o conteúdo.

Piwowarski (2003) define o grau de estruturação de um documento a partir das seguintes questões: (i) qual a plataforma de registro do documento (impresso, digital, voz, vídeo, fotografia, etc.); (ii) como o documento se estrutura (conjunto de atributos que o constroem) e quais elementos de comunicação estão contidos (imagens, texto, áudio, etc.); (iii) e como a estrutura do documento é descrito globalmente (se existem ou não formatos universais de representação e descrição do conteúdo).

De outra perspectiva, o documento pode ser considerado estruturado quando está disponível explicitamente no ambiente a que pertence. Quando seu conteúdo estiver ao alcance da análise, de forma confiável, cumprindo os critérios de elegibilidade da informação: útil (quando tem algum uso ou serve para algo); mensurável (passível de ser lastreada); datada (cuja data de obtenção e de produção é possível determinar); específica (pertencente exclusivamente a uma situação); acurada (elaborada ou obtida com rigor) e atualizada (própria para o momento atual).

A análise de informações estruturadas em base de dados requer uma série de ações, cada qual exigindo recursos diferentes. Pode-se indicar pelo menos o seguinte sequenciamento das etapas de trabalho contidas: decisão do problema a ser estudado; escolha da fonte de dados; conservação do sistema operacional; extração dos dados de trabalho; limpeza e preparação dos dados; tratamento e análise dos dados.

a) Plataforma de tratamento

É importante nas primeiras etapas mapear com precisão o campo a ser estudado e levantado, bem como coletar com responsabilidade, completeza e de forma padronizada os dados. Nesta fase o foco é no controle da qualidade dos dados a serem coletados.

A primeira e a segunda etapa, em princípio, não requerem a intervenção de *software*. A partir da terceira etapa é obrigatório o uso de *software* livre ou proprietário. É desejável que a plataforma de trabalho tenha uma boa interoperabilidade, a disponibilização de um ponto focal de recursos já disponíveis que permita seu compartilhamento e a atualização permanente do sistema operacional, geralmente feita de modo automático.

No que se refere à conservação do sistema operacional, hoje, qualquer exposição de um computador na *web* provoca o risco de sua contaminação por vírus. Para manter um bom desempenho da máquina é desejável sempre mantê-lo limpo. Para este fim, existem algumas ferramentas indispensáveis a serem utilizadas frequentemente. Algumas delas são: AdwCleaner[5], Geek Uninstaller[6], Ccleaner[7] e Spybot[8].

b) Limpeza e preparação dos dados

Ao realizar o tratamento da informação para análise, tomada de decisão e agregação de valor, a etapa mais demorada e exaustiva é a limpeza dos dados, que deve ser efetuada antes da análise propriamente dita. Essa etapa, apesar de consumir bastante tempo ao longo do processo, é fundamental por ter grande repercussão na qualidade dos resultados, tanto em termos de significado como de custo de produção. É importante ter esta informação em mente para dosar o difícil equilíbro entre o rigor desejado na informação obtida da limpeza e nos resultados esperados.

Esta fase é considerada a menos nobre da análise, tanto na perspectiva da ciência da informação como da ciência da computação. Mas o fato é que ela é indispensável e necessita de uma boa integração entre cientistas da informação e da computação para que haja bom resultado.

A grande maioria dos formatos de arquivos manipulados no âmbito da análise de informação bibliográfica são arquivos estruturados do tipo texto. Neste sentido podem ser trabalhados com um editor, com vistas a sua padronização. Existem várias soluções possíveis desde que sejam contemplados os seguintes requisitos: conversão do fim de linha; conversão maiúscula/minúscula (palavra inteira ou primeira letra); procura/substitui avançado; procura/substitui na forma *regex*[10]; capacidade de edição de arquivo grande; conversão de caracteres; possibilidade de ordenação; visuali-

zação dos caracteres especiais (tab, fim de linha etc); função anula, repita; manipulação simultânea de vários arquivos; trabalhar em modo coluna; sistema de macro comandos; coloração do texto conforme *template*[11] fornecido; uso de conversão XML (XSL) e reconhecimento de XPATH.

É possível indicar várias soluções para atender a esses requisitos. A maioria dos editores que atendem a estas características é voltada à programação. Portanto, a etapa de tratamento da informação de maneira eficiente necessita de aprendizado da parte do usuário com formação em ciência da informação. Algumas opções são *Notepad*++[12] e *Wreplace*[13].

O tratamento de texto pode exigir muito do analista. Entre outros, porque pode ir até à análise linguística e semântica do documento. Ainda assim, as soluções rápidas, que podem parecer simples, também são capazes de fornecer bons resultados e acima de tudo uma boa relação custo e qualidade. Uma parte deste tratamento solicita a remoção de palavras vazias ou *stopwords.* Estas listas de *stopwords* existem em várias línguas e podem ser obtidas na *web*.

Além dos supracitados, recomenda-se aprofundar a análise com a utilização dos Regex[14], Xpath[15] e XSL/XSLT[16].

c) Tratamento e análise dos dados

Uma vez limpos e preparados, tem início o gratificante momento da análise de dados, atividade de organizar, estruturar e dar significado a determinada situação representada pelos dados, de forma racional, a partir de métodos de análise qualitativos ou quantitativos e obter informações de valor agregado, suficientes para inferir conclusões, dimensionar fatos e ampliar o conhecimento do universo contido na análise.

Suas várias dimensões têm início com a definição do método de análise das variáveis. Nesse trabalho, serão exploradas brevemente as análises univariada, bivariada e multivariada. A escolha da análise pretendida é decisiva e está relacionada com os objetivos do estudo.

Na análise univariada cada variável é estudada isoladamente e de forma descritiva (os dados são organizados em tabelas e gráficos e onde se calculam características como a moda, a mediana, a média, o desvio pa-

drão, entre outras), a partir de uma única variável. Muito embora *software* como Excel ou Libre Office tenham possibilidades sofisticadas de fazer tabelas cruzadas, é preferível o uso do *software* PivotTable.js[17] ou semelhante.

A estatística bivariada inclui métodos de análise de duas variáveis, podendo ser ou não estabelecida uma relação de causa e efeito entre elas. Exemplo típico desse método é o estudo da relação linear entre duas variáveis.

A análise multivariada inclui as relações de múltiplas variáveis dependentes e ou independentes, quer se estabeleçam ou não relações de causa e efeito entre elas. Somente nesse caso é possível explorar o desempenho conjunto das variáveis e analisar como se comportam, umas em relação às outras. Nesse caso, as tecnologias mais indicadas para tratamento e análise são o Pajek[18,] Netdraw[19], R statistical package[20], Gephi[21], CSV, GDF, Gexf[22], *Command line execution*[23], SPARQL *plugin*[24] e Twitter com Gephi[25]. Existem soluções simples e elegantes de análise que não precisam de análise linguística detalhada envolvendo gramática, como os recomendados Treecloud[26], Cowo[27] e Vosviewer[28], JsLDA[29], Iramuteq e Ngrams.

d) Bases de dados de acesso aberto

Hoje existem bases de dados que dão acesso a grandes e valiosos recursos de informação. Para mantenedores de iniciativas de acesso aberto, pode ser valioso usar este recurso para melhorar seu desempenho ou incluir complementos de informação, e ao mesmo tempo, aumentar o valor e a consistência do que está compartilhando. Aqui serão descritas brevemente três bases de dados, de fundamental importância para promoção do conhecimento livre.

e) Plataforma Lattes

A Plataforma Lattes de currículos *vitae* de pesquisadores brasileiros se tornou um sistema imprescindível para a ciência nacional. Foi uma estratégia engenhosa disponibilizar a produção científica brasileira a baixo custo. Ela se baseia no modelo 2.0, onde cada pesquisador registra e mantém atualizada sua produção científica, de forma descentralizada,

sem grande controle externo. Hoje ela contempla mais de 1.5 milhões de currículos, somente de doutores. A identificação personalizada de cada currículo, com um número único (ID Lattes) a torna um recurso valioso para interação com outros sistemas e para análise dela própria, que contém conhecimento variado e não analisado sistematicamente, da pesquisa produzida no país.

Além das ferramentas de busca e visualização disponíveis na Plataforma Lattes, conta-se com algumas outras que permitem o processamento e compilação de relatórios, com destaque para o *Scriptlattes*[30], ferramenta para extração e visualização de conhecimento. Foi desenvolvida para a extração e compilação automática de: produções bibliográficas, produções técnicas, produções artísticas, orientações, projetos de pesquisa, prêmios e títulos, grafo de colaborações, mapa de geolocalização, coautoria e internacionalização, do conjunto de pesquisadores cadastrados na plataforma Lattes.

O *ScriptLattes* coleta automaticamente os currículos Lattes em formato HTML de um grupo de pessoas de interesse, compila as listas de sua produção, tratando apropriadamente as produções duplicadas e similares. São geradas páginas HTML com listas de produções e orientações, separadas por tipo e colocadas em ordem cronológica invertida. Adicionalmente são criados automaticamente vários grafos (redes) de coautoria entre os membros do grupo de interesse e um mapa de geolocalização dos membros e alunos (de pós-doutorado, doutorado e mestrado) com orientação concluída. Os relatórios gerados permitem avaliar, analisar e documentar a produção de grupos de pesquisa (SCRIPTLATTES, 2015).

f) Wikipedia

A *Wikipedia* é uma enciclopédia multilíngue, colaborativa e livre que contém entradas (artigos) em mais de 300 línguas, sendo o inglês o mais representativo, com quase 3 milhões de entradas. Os 10 idiomas em ordem de ocorrência são o inglês (2.826.000); alemão (888.000); francês (786.000); polonês (593.000); italiano (576.000); japonês (556.000); holandês (528.000); português (470.000); espanhol (460.000) e russo (376.000). Interessante notar que a *Wikipedia* não é a tradução de um

artigo em várias línguas, mas sim, cada artigo criado de forma independente por diferentes usuários (OTERO; LÓPEZ, 2010).

Ela representa um enorme avanço para o conhecimento aberto, ainda que o sistema colaborativo possa levar a algumas informações imprecisas. A renomada revista *Nature* escolheu artigos da *Encyclopedia Britannica* e da *Wikipedia* e enviou para a revisão por pares. Ao final do processo, o jornal encontrou apenas oito erros graves nos artigos. Destes, quatro veio de cada uma. No entanto, descobriram uma série de erros pontuais, omissões ou declarações imprecisas, somando ao todo 162 da *Wikipedia* e 123 da *Encyclopedia Britannica*. Isso dá em média de 2,92 erros por artigo para a *Britannica* e 3,86 para a *Wikipedia*, o que leva à conclusão por parte da *Wikipedia* de que o resultado foi razoavelmente favorável (Terdiman, 2005).

O modelo adotado pela *Wikipedia* se assemelha ao da comunidade de *software* de código aberto, onde o código é acessível a todos e milhares de pessoas colaboram para sua realização. A diferença reside no fato de que no *software* aberto, uma versão final emerge como oficial, mas o conteúdo da *Wikipedia* nunca está finalizado, nunca há uma versão oficial de um artigo (Berinstein, 2006).

4 Classificação Internacional de Patentes[31] (IPC)[32]

A Classificação Internacional de Patentes (IPC), mantida pela WIPO[33] (Organização Internacional de Propriedade Intelectual) desde 1968, é um sistema complexo de descrição do conteúdo das patentes e modelos de utilidade, de acordo com as diferentes áreas tecnológicas a que pertencem, sendo utilizada por mais de 100 países. Seus objetivos são criar uma ferramenta de organização, busca, recuperação e disseminação de documentos de patentes e ser a base para investigação do estado da arte em inovação.

A IPC é mantida em duas línguas (inglês e francês), disponível para consulta *online* e para *download*. Todos os arquivos podem ser livremente baixados, armazenados, copiados, modificados (*layout* ou formato), impressos, usados para criar produtos derivados e publicamente apresentados.

Dentre os recursos de suporte à IPC está a ferramenta de categorização, especialmente projetado para auxiliar o processo de classificação. Outra ferramenta, a *IPC Inventory Green,* fornece previsões da classificação com base na análise estatística dos documentos de patentes que contêm os termos de pesquisa especificados, facilitando a procura por informações sobre patentes relativas a tecnologias ambientalmente saudáveis.

5 Conclusão

As soluções aqui apresentadas atendem aos requisitos de código aberto e *software* livre e são em sua maioria desenvolvidos na esfera acadêmica. A utilização de *software* relacionado a este trabalho, nem sempre representa o investimento mais econômico, mas as despesas associadas à sua utilização, em seu conjunto, são seguramente menores. Além disso, é preciso considerar as questões relacionadas à responsabilidade social associada à decisão.

Para contribuir com o ideal do conhecimento aberto, e valorizar o esforço daqueles que estão construindo soluções e compartilhando o seu conhecimento com a sociedade, esse trabalho encerra destacando que os autores aqui apresentados, tanto da perspectiva política, quanto da perspectiva tecnológica, têm um papel singular na sociedade contemporânea: são eles que contribuem com suas iniciativas para o fortalecimento do conhecimento aberto, preparando o mundo para ser mais justo e equânime.

Referências

BAILEY JR., C. W. *What is open access?* 2006. In: Portal da Digital Scholarship. Disponível em: <http://www.digital-scholarship.org/cwb/WhatIsOA.htm>. Acesso em: 7 fev. 2015.

BERINSTEIN, P. Wikipedia and Britannica. *Information Today*, v.14, n.3, Mar. 2006. Disponível em: <http://www.infotoday.com/searcher/mar06/berinstein.shtml>. Acesso em: 7 fev. 2015.

CAVALCANTE, L. F. B.; VALENTIM, M. L. P. Informação e conhecimento no contexto de ambientes organizacionais. In: VALENTIM, M. L. P. (Org.). *Gestão, mediação e uso da informação.* São Paulo: Cultura Acadêmica, 2010.

DIRECTORY of Open Access Journals. Disponível em: <http://doaj.org/>. Acesso em: 7 fev. 2015.

GASPAR, A. C.; ÁLVARES. L.; PEREIRA, M. N. F. Gestão dos dados de pesquisa: oportunidades e desafios. In: SEMINARIO HISPANO-BRASILEÑO DE INVESTIGACIÓN, DOCUMENTACIÓN Y SOCIEDAD, 3., Madrid. *Anais...* Madrid: Universidade Complutense de Madrid, 2014.

GERGEN, K. J.; GERGEN, M. *Construcionismo social:* um convite ao diálogo. Rio de Janeiro: Instituto Noos, 2010. 120p.

HARVARD Law School. *Lawrence Lessig.* [s.d. a]. Disponível em: <http://hls.harvard.edu/faculty/directory/10519/Lessig>. Acesso em: 7 fev. 2015.

HARVARD Law School. *Yochai Benkler.* [s.d. b]. Disponível em: <http://hls.harvard.edu/faculty/directory/10071/Benkler>. Acesso em: 7 fev. 2015.

HARVARD Open Access Project. *Peter Suber.* Disponível em: <http://cyber.law.harvard.edu/node/8322>. Acesso em: 7 fev. 2015.

HAYEK, F. The use of knowledge in society. *American Economic Review,* v.35, n.4, p.519-530, 1945. Disponível em: <http://home.uchicago.edu/~vlima/courses/econ200/spring01/hayek.pdf>. Acesso em: 30 jan. 2015.

JIMMY, Wales: Free knowledge for free minds. Disponível em: <http://jimmywales.com/>. Acesso em: 7 fev. 2015.

KELTY, C. M. *Two bits:* the cultural significance of free software. London: Duke University Press, 2008. Disponível em: <http://twobits.net/pub/Kelty-TwoBits.pdf>. Acesso em: 30 jan. 2015.

KNOWLEDGE Ecology International. *Access to knowledge.* 1995. Disponível em: <http://www.cptech.org/a2k/>. Acesso em: 7 fev. 2015.

LEV, B. *Intangibles:* management, measurement, and reporting. Washington: Brookings, 2001. Disponível em: <http://www.redalyc.org/articulo.oa?id=195416332006>. Acesso em: 1 fev. 2015.

MURRAY-RUST, P. et al. *Panton principles:* principles for open data in science. Disponível em: <http://pantonprinciples.org>. Acesso em: 24 set. 2014.

OPEN Knowledge Foundation. Disponível em: <https://okfn.org/>. Acesso em: 30 jan. 2015.

OTERO, P. G.; LÓPEZ, I. G. Wikipedia as multilingual source of comparable corpora. In: WORKSHOP ON BUILDING AND USING COMPARABLE CORPORA (LREC 2010), 3., 2010. *Proceedings…* Malta, 2010. p.21–25

OPEN Source Iniciative. Disponível em: <http://opensource.org/docs/definition.php>. Acesso em: 7 fev. 2015.

ORGANIZAÇÃO das Nações Unidas. *Declaração Universal dos Diretos Humanos.* Nova York: ONU, 1948.

P2P Foundation. *Charlotte Hess.* Disponível em: <http://p2pfoundation.net/Charlotte_Hess>. Acesso em: 7 fev. 2015.

PIWOWARSKI, B. *Techniques d'apprentissage pour le traitement d'informations structurées:* application à la recherche d'information. Paris, 2003. Tese (Doutorado). L'Université Paris 6, 2002.

POETRY Foundation. *Mark Van Doren.* Disponível em: <http://www.poetryfoundation.org/bio/mark-van-doren>. Acesso em: 7 fev. 2015.

PUBLIC Library of Science. Disponível em: <http://www.plos.org/>. Acesso em: 7 fev. 2015.

QUONIAM, L. *Levantamento dos softwares livres disponíveis para análise de informação estruturada.* Brasília: 2014. (Relatório de Consultoria ao IBICT).

RAYMOND, E. S. *The cathedral and the bazaar.* Sebastopol (CA): O'Reilly, 1999.

REMEMBER Aaron Swartz. Disponível em: <http://www.rememberaaronsw.com/memories/>. Acesso em: 7 fev. 2015.

SAY Libre. Disponível em: <http://wikieducator.org/Say_Libre>. Acesso em: 7 fev. 2015.

SCIENCE Commons. Disponível em: <http://sciencecommons.org/about/whoweare/wilbanks/>. Acesso em: 7 fev. 2015.

SCRIPTLATTES. Disponível em: <http://scriptlattes.sourceforge.net/>. Acesso em: 7 fev. 2015. STALLMAN, R. M.; LAWRENCE, L.; GAY, J. *Free software, free society:* selected essays of Richard M. Stallman. Boston: Free Software Foundation, 2006.

TERDIMAN, D. Study*:* Wikipedia as accurate as Britannica. *CNET News,* Dic., 2015. Disponível em: <http://news.cnet.com/Study-*Wikipedia*-as-accurate-as-Britannica/2100-1038_3-5997332.html>. Acesso em: 7 fev. 2015.

THE SWARAJ Foundation. *Intepreting Gandhi's Hind Swaraj.* 1999. Disponível em: <http://www.swaraj.org/interpreting.htm>. Acesso em: 7 fev. 2015.

XAVIER, A. C. As tecnologias e a aprendizagem (re)construcionista no século XXI. In: XAVIER, A. C.; MARCUSCHI, L. A. (Orgs.). *Hipertexto e gêneros digitais.* Recife: Parábola Editorial, 2005. Disponível em: <https://www.ufpe.br/nehte/revista/artigo-xavier.pdf>. Acesso em: 7 fev. 2015.

WORLD Intellectual Property Organization. *International Patent Classification* (IPC). Disponível em: <http://www.wipo.int/classifications/ipc/en/>. Acesso em: 7 fev. 2015.

Notas

1. Disponível em: <http://unesdoc.unesco.org/images/0013/001399/139937por.pdf>.
2. Organização não governamental sem fins lucrativos, fundada em 2004 na cidade de Cambridge (Inglaterra) na *St. John's Innovation Centre* (SJIC), incubadora de negócios em ciência e tecnologia, que promove o conhecimento aberto, os dados abertos e os conteúdos abertos, hoje com o nome de Open Knowledge
3. Access to Knowledge.
4. Biblioteca Pública de Ciência.
5. Disponível em: <http://www.bleepingcomputer.com/download/adwcleaner/>. Acesso em: 30 jan. 2015.
6. Disponível em: <http://www.geekuninstaller.com/>. Acesso em: 30 de jan. 2015.
7. Disponível em: <https://www.piriform.com/CCLEANER>. Acesso em: 30 de jan. 2015.
8. Disponível em: <http://www.safer-networking.org/>. Acesso em: 30 de jan. 2015.
9. Knowledge for all, freedom to learn, towards collective wisdom, enabling people to empower themselves with knowledge, and to share it for community benefit.
10. Em Ciência da Computação, uma expressão regular (ou o estrangeirismo *regex*, abreviação do inglês Regular Expression) provê uma forma concisa e flexível de identificar cadeias de caracteres de interesse, como caracteres particulares, palavras ou padrões de caracteres. Expressões regulares são escritas numa linguagem formal que pode ser interpretada por um processador de expressão regular, um programa que ou serve um gerador de analisador sintático ou examina o texto e identifica partes que casam com a especificação dada. O uso atual de expressões regulares inclui procura e substituição de texto em editores de texto e linguagens de programação, validação de formatos de texto (validação de protocolos ou formatos digitais), realce de sintaxe e filtragem de informação.
11. Fonte: <http://en.wikipedia.org/wiki/Regular_expression>.
12. Template é um modelo de documento sem conteúdo, apenas com a desejável apresentação visual e diagramação e informações sobre como e onde os conteúdos devem ser inseridos.
13. Disponível em: <http://notepad-plus-plus.org/>. Acesso em: 30 jan. 2015.
14. Disponível em: <http://www.sharktime.com/us_wReplace.html>. Acesso em: 30 jan. 2015.
15. Disponível em: <http://pt.wikipedia.org/wiki/Express%C3%A3o_regular>. Acesso em: 30 jan. 2015
16. Disponível em: <http://www.w3schools.com/XPath/>. Acesso em: 30 jan. 2015.
17. Disponível em: <http://www.w3schools.com/xsl/default.asp>. Acesso em: 30 jan. 2015.
18. Disponível em: <http://nicolas.kruchten.com/pivottable/examples/index.html>. Acesso em: 30 jan. 2015.
19. Disponível em: <http://pajek.imfm.si/doku.php>. Acesso em: 30 jan. 2015.
20. Disponível em: <http://www.analytictech.com/>. Acesso em: 30 Jan. 2015.
21. Disponível em: <http://www.r-project.org/>. Acesso em: 30 jan. 2015.
22. Disponível em: <https://gephi.github.io/>. Acesso em: 30 jan. 2015.
23. Disponível em: <http://gexf.net/format/>. Acesso em: 30 jan. 2015.
24. Disponível em: <https://gephi.github.io/toolkit/>. Acesso em: 30 jan. 2015.
25. Disponível em: <https://marketplace.gephi.org/plugin/semanticwebimport/>. Acesso em: 30 jan. 2015.
26. Disponível em: <http://matthieu-totet.fr/Koumin/tools/naoyun/>. Acesso em: 30 jan. 2015.
27. Disponível em: <http://treecloud.univ-mlv.fr/>. Acesso em: 30 jan. 2015.
28. Disponível em: <https://github.com/seinecle/Cowo/blob/master/README.md>. Acesso em: 30 jan. 2015.
29. Disponível em: <http://www.vosviewer.com/>. Acesso em: 30 jan. 2015.
30. Disponível em: <https://github.com/mimno/jsLDA>. Acesso em: 30 jan. 2015.
31. O ScriptLattes não está vinculado ao CNPq. A ferramenta é o resultado de um esforço (independente) realizado com o único intuito de auxiliar as tarefas mecânicas de compilação de informações cadastradas nos Currículos Lattes (publicamente disponíveis).
32. International Patent Classification.
33. Disponível em: <http://www.wipo.int/classifications/ipc/en/>. Acesso em: 30 jan. 2015.
34. World Intellectual Property Organization.