WIDaT 2019

Workshop de Informação Dados e Tecnologia

ANAIS DO EVENTO

UnB, sua linda

UnB Programa de Pós-graduação Ciência da Informação - PPGCINF FACULDADE DE CIÊNCIA DA INFORMAÇÃO

FACULDADE DE CIÊNCIA DA INFORMAÇÃO

ISBN: 978-65-86503-01-2

WORKSHOP DE INFORMAÇÃO, DADOS E TECNOLOGIA
Universidade de Brasília – UnB
27, 28 e 29 de novembro de 2019, Brasília – Distrito Federal

ANAIS WIDAT 2019
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO
UNIVERSIDADE DE BRASÍLIA

Organizadores

Dalton Lopes Martins (PPGCINF/UnB)
José Eduardo Santarem Segundo (PPGCI/UNESP - USP)
Márcio Bezerra da Silva ((PPGCINF/UnB)
Marcio Victorino ((PPGCINF/UnB)

Brasília
2020

W926 Workshop de informação, dados e tecnologia (3. : 2019 : Brasília).
Workshop de informação, dados e tecnologia (WIDAT 2019) : anais do evento / Dalton Lopes Martins ... [et al.], organizadores. – Brasília : Universidade de Brasília, Faculdade de Ciência da Informação, 2019.
150 p. il.

Modo de acesso: World Wide Web: http://widat2019.fci.unb.br/.

ISBN 978-65-86503-01-2.
Inclui bibliografia.

1. Ciência da informação – Workshop. 2. Tecnologia. I. Martins, Dalton Lopes, (org.). II. Título.

CDU 02

Organização WIDaT 2019

- **Coordenação**:

Dalton Lopes Martins (PPGCInf/UNB)

- **Organização Geral**:

José Eduardo Santarem Segundo (PPGCI/UNESP - USP)
Márcio Bezerra da Silva (PPGCInf/UNB)
Marcio Victorino (PPGCI/UNB)

- **Coordenação da Comissão Científica**:

José Eduardo Santarem Segundo - Coordenador - (PPGCI/UNESP - USP)

- **Comissão Científica**:

Adilson Luiz Pinto (PGCIN-UFSC)
Ana Carolina Simionato (PPGCI-UFSCar)
Dalton Lopes Martins (PPGCInf/UNB)
Denysson Axel Ribeiro Mota (PPGB/UFCA)
Douglas Dyllon Jeronimo de Macedo (PGCIN-UFSC)
Elaine Parra Afonso (Fatec-Presidente Prudente/SP)
Enrique Muriel Torrado (PGCIN-UFSC)
Guilherme Ataíde Dias (PPGCI-UFPB)
Henry Pôncio (PPGCI-UFPB)
Joyce Siqueira (PPGCInf/UNB)
Leonardo Castro Botega (UNIVEM-Marília)
Luis Felipe Rosa de Oliveira (PPGCInf/UNB)
Marcello Bax (PPGGOC-UFMG)
Marckson Roberto Ferreira de Sousa (PPGCI-UFPB)
Márcio Bezerra da Silva (PPGCInf/UNB)
Márcio Matias (PGCIN-UFSC)
Márcio Victorino (PPGCInf/UNB)
Marcos Mucheroni (CBD/USP)
Moisés Lima Dutra (PGCIN-UFSC)
Plácida Santos (PPGCI-UNESP)
Robson Rodrigues Lemos (UFSC-Araranguá)
Sandro Rautenberg (DECOMP-UNICENTRO)
Silvana Aparecida Borsetti Gregorio Vidotti (PPGCI-UNESP)
Wander Jacome Queiroz (Western University)

- **Comissão Técnica Local**:

Calíope Victor Spíndola de Miranda Dias (UNB)
Eduardo Alves Silva (NOVA IMS - UNL)
Joyce Siqueira (UNB)
Luis Felipe Rosa de Oliveira (UNB)

Sumário

A CURADORIA DE DADOS CIENTÍFICOS NA CIÊNCIA DA INFORMAÇÃO: LEVANTAMENTO DO CENÁRIO NACIONAL

CURATION SCIENTIFIC DATA IN INFORMATION SCIENCE:
Survey National Scenario

Liliane Chaves de Resende[1], Marcello Peixoto Bax[(2)]
Universidade Federal de Minas Gerais, UFMG, lilianederesende@gmail.com
Universidade Federal de Minas Gerais, UFMG, bax@ufmg.br

Resumo:
Para a ciência contemporânea, o compartilhamento e reutilização de dados científicos constituem elementos primordiais para a colaboração entre comunidades científicas e progresso da ciência. Para se adequar a esse cenário, os profissionais da informação necessitam desenvolver habilidades para realizar atividades de curadoria digital dos dados científicos. O objetivo da pesquisa é levantar junto aos pesquisadores brasileiros da área da Ciência da Informação, sua percepção sobre o grau de envolvimento da área no *momentum* internacional da curadoria digital de dados científicos. A pesquisa, do tipo exploratória e descritiva, utilizou procedimentos de pesquisa com *Survey* para obter opiniões da comunidade científica da área da ciência da informação sobre o tema. As informações analisadas revelam que para o campo da Ciência da Informação no Brasil, o desenvolvimento da curadoria digital de dados científicos está em fase inicial. É necessária uma mudança evolutiva considerável na formação disciplinar teórica, prática e técnica desses pesquisadores para fortalecer a área da Ciência da Informação brasileira para assumir a curadoria digital como parte de sua missão.

Palavras-chave: Curadoria Digital; Dados Científicos; *e-Science*; Profissional da Informação.

Abstract:
For contemporary science, sharing and reusing scientific data are key elements for collaboration between scientific communities and the advancement of science. To suit this scenario, information professionals need to develop skills to perform digital curation activities for scientific data. The objective of the research is to raise with the Brazilian researchers in the area of Information Science, their perception about the degree of involvement of the area in the international momentum of digital curation of scientific data. The research, exploratory and descriptive, used research procedures with Survey to obtain opinions from the scientific community in the area of information science on the subject. The information analyzed reveals that for the field of Information Science in Brazil, the development of digital curation of scientific data is at an early stage. Considerable evolutionary change in the theoretical, practical and technical disciplinary training of these researchers is required to strengthen the Brazilian Information Science field to take on digital curation as part of their mission.

Keywords: Digital curation; Scientific data; e-Science; Information professional

1. Introdução

As comunidades acadêmicas estão se conscientizando das atividades necessárias para o gerenciamento da informação científica, como recurso que fortalece o desenvolvimento da ciência e proporciona novas descobertas do conhecimento.

O contexto de novas demandas ao gerenciamento dos dados científicos, fez surgir a emergente área de estudo denominada curadoria digital (CD). O compartilhamento dados científicos tornou-se fundamental para o progresso da ciência. Influencia a colaboração entre comunidades científicas. Portanto, gerenciar dados científicos para uma determinada área de pesquisa é inerente às particularidades da área e do conhecimento que se deseja transmitir (SAYÃO; SALES, 2016).

O uso e prática das atividades de Curadoria Digital de Dados Científicos (CDDC) já são realizadas em bibliotecas acadêmicas de pesquisa em universidades, principalmente, de países como EUA, Canadá e Reino Unido. A

execução de atividades necessárias às práticas da CDDC pode transformar as bibliotecas de pesquisas acadêmicas em *lócus* de gerência e curadoria de dados científicos. Sobretudo pelo auxílio que pode ser dado pelos bibliotecários aos pesquisadores na realização de atividades de curadoria.

No Brasil, esse cenário, ainda, está em processo inicial. Percebe-se que há interesse por parte dos pesquisadores em desenvolver a CDDC produzido em suas pesquisas, porém ainda não se tem desenvolvido uma política sólida, com definições claras sobre o assunto e de como isso poderá se tornar realidade nas instituições de pesquisa acadêmicas.

2. Objetivos

O objetivo desse estudo é investigar a relevância e o grau de adesão que as atividades de curadoria digital de dados científicos têm para a área da Ciência da Informação no cenário brasileiro.

Segundo Kouper (2016), a iniciativa de obter informações sobre atividades de CD diretamente daqueles que pesquisam assuntos da área, amplia o conhecimento existente na área e aprimora a compreensão dos conhecimentos, valores e experiência cotidianas desses profissionais (KOUPER, 2016). Partindo desse pressuposto, obter opiniões acerca das atividades de CDDC executadas por pesquisadores brasileiros da área da CI é imprescindível para que possamos descrever como está evoluindo a área de CD no Brasil e qual seu estado atual, especificamente para a área de CI.

3. Procedimentos Metodológicos

A pesquisa segue abordagem quali-quantitativa, do tipo exploratória e descritiva. Utiliza procedimentos de uma pesquisa com *Survey*, para obter informações de atividades de CDDC dos pesquisadores brasileiros da área da Ciência da Informação. O método utilizado é da amostragem aleatória simples.

Para compor a população alvo foi definido os pesquisadores que atuam em programas de pós-graduação de mestrado e doutorado na área da CI, avaliados pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

A amostra foi auto selecionada, ou seja, composta por pesquisadores que aceitaram o convite em participar do levantamento. Os pesquisadores foram convidados a participar do levantamento e tiveram a mesma chance de participar da pesquisa.

Os pesquisadores foram contatados por correio eletrônico (*e-mail*). Os endereços de *e-mail* foram obtidos por meio de informações públicas, disponibilizadas nas páginas oficiais dos programas de pós-graduação ou disponibilizados em artigos científicos publicados em periódicos de acesso aberto.

A elaboração de um questionário, como instrumento de pesquisa, visou definir às seguintes questões: 1) caracterizar o perfil dos pesquisadores brasileiros que atuam em programas de pós-graduação de mestrado e doutorado na área de CI; 2) descrever o nível de conhecimento desses pesquisadores sobre a CDDC; 3) descobrir o grau de envolvimento nas atividades de CDDC quando desenvolvem suas pesquisas; 4) levantar opiniões desses pesquisadores a respeito da CDDC, para estimar tendências na área.

Os dados foram coletados entre os meses de maio a agosto de 2019, por meio de um questionário on-line, auto aplicado, enviado via e-mail contendo o link para acesso ao questionário disponibilizado pelo *Google Forms*. Optou-se pelo uso dessa ferramenta devido a sua simplicidade, facilidade de uso e agilidade na obtenção dos dados.

Realizou-se um pré-teste da aplicação do questionário para verificar possíveis falhas e testar o processo de coleta de dados. O pré-teste foi aplicado a pesquisadores da Escola de Ciência da Informação da UFMG, pertencentes do programa de pós-graduação, em condições similares à população alvo pesquisada (GIL, 2008).

Os *e-mails* foram a trezentos e noventa (390) pesquisadores, e noventa e sete (97) responderam o questionário. A amostra foi autosselecionada e

dependente do número de pesquisadores que optaram por participar. Dado o número de entrevistados que responderam, a amostragem foi considerada representativa para generalizar os dados. Esse estudo demandou uma compreensão básica dos conceitos de CDDC por parte dos entrevistados. Dada a natureza emergente da disciplina, pode ter sido esse um motivo pelo qual a participação de respondentes foi baixa.

Para contabilizar os dados, utilizou-se uma versão gratuita de teste do software SPSS Statistics Subscription (Statistical Package for the Social Sciences – Pacote Estatístico para as Ciências Sociais). O software SPSS foi desenvolvido pela IBM e é considerado um software muito útil para apoiar análises estatísticas. As respostas foram tabuladas e analisadas por meio de suas frequências absolutas.

4. Resultados e Discussões

Para levantar o cenário nacional da CDDC na área da CI, os pesquisadores foram analisados da seguinte forma: 1) o perfil do pesquisador para saber suas características e o que ele conhece de CDDC; 2) as preferências em CD, para saber quais atividades ele executa e quais suas necessidades de formação técnica para praticar atividades de CD; 3) as tendências de da CDDC, para saber sua opinião sobre o desenvolvimento da área no Brasil.

4.1 Perfil dos pesquisadores da CI

Os pesquisadores da área de CI no Brasil possuem idade acima de 40 anos (74%), a maioria é feminina (59%), com mais de 10 anos de experiência de atuação em pesquisa (55%). Essa porcentagem está em acordo com a súmula estatística dos pesquisadores realizada pela CNPq em 2016, que informa que a relação da presença feminina entre pesquisadores brasileiros é maior do que a presença masculina.

Realizam pesquisas, principalmente, nas universidades em que trabalham (84%), no setor público (46%). A maior parte desses pesquisadores buscam financiamentos em institutos nacionais de fomento (87%). Interage com outras universidades brasileiras foi de 58,2% dos casos em relação a 32,8% de universidades internacionais. Possuem formação básica não somente em áreas correlatas à CI (62%), mas há uma grande participação de outras áreas de conhecimento (38%), comprovando o que informa Saracevic (1996), ser a CI nacional também uma área multidisciplinar (SARACEVIC, 1996).

Cerca de 79% dos entrevistados tem conhecimento do Manifesto de Acesso Aberto a Dados de Pesquisa Brasileira para a Ciência Cidadã, lançado em 2016 pelo Instituto Brasileiro de Ciência e Tecnologia (IBICT). Este manifesto trouxe recomendações para as instituições brasileiras sobre os dados de pesquisa, consolidando o movimento mundial de acesso aberto à informação e dados científicos no Brasil.

Cerca de 92% dos pesquisadores conhecem repositórios de dados científicos sugeridos na pesquisa (BDC/UFPR – Base de dados científicos da UFPR, CIS – Consórcio de informações sociais, GLOBE – *Global Collaboration Engine*, IBGE – Instituto Brasileiro de Geografia e Estatística e IBICT *Dataverse Network*) e somente 5% sugeriram outros repositórios. Poucos respondentes participaram de cursos de capacitação ou treinamento específicos de curadoria digital (18%) e quase todos (82%) não participam na elaboração de políticas e normas nacionais para a efetiva implantação da curadoria digital no cenário nacional.

O perfil dos pesquisadores brasileiros da CI demonstra que ainda não há um envolvimento da comunidade científica da área que contribuía significativamente com a evolução da CDDC no Brasil. Poucos profissionais parecem estar se movimentando para desenvolver conhecimentos práticos de CD.

4.2 Preferências em atividades de CD

Esse estudo buscou levantar informações sobre atividades básicas de CDDC exercidas pelos pesquisadores da CI. Em torno de 43% dos pesquisadores informaram que utilizam o *DSpace* para

reutilizar dados de pesquisa e 46% para armazenar seus dados. Mas 49% afirmam nunca terem usado uma plataforma tecnológica de dados científicos.

O gerenciamento de dados científicos produzidos nas pesquisas realizadas por estes profissionais é feito armazenando seus dados em computador pessoal (78%) ou na nuvem (74%). Somente 22% informaram que armazenam os dados em plataforma fornecida pela instituição onde trabalham. Mas ao serem questionados sobre qual modelo de referência utilizam para elaboração do seu plano de gestão de dados, 47% informaram que não sabem dizer que modelo utilizar. Somente 18% informaram que utilizam o modelo do *Digital Curation Centre* (DCC), 14% o modelo OAIS e 11% utilizam modelos oferecidos pelas Instituições de fomento.

Sobre o que fazem com esses dados, 63% já compartilhou seus dados, 46% disponibilizou e 61% informaram que reutilizam dados de suas pesquisas. Em torno de 49% afirmam já ter citado dados científicos de outros pesquisadores e 47% já ter publicados seus próprios dados. Os pesquisadores afirmaram que publicam seus dados junto com os artigos (56%) ou divulgando em eventos científicos (50%).

Quando questionados sobre quais padrões de metadados utilizam para descrever seus dados científicos, 52% não souberam informar e 41% informaram que utilizam o padrão *Dublin Core*. O que se observa é que ainda não há uma conscientização sobre a importância das atividades de curadoria digital. Mesmo tendo informado a condução de atividades como armazenar, acessar, disponibilizar e reutilizar dados de pesquisa, o quantitativo de respondentes que afirmam disponibilizar seus dados de pesquisa ainda é baixo, cerca de 46%. Além disso, quando gerenciam seus dados de pesquisa somente 25% demonstram disponibilizá- los de fato.

Dessa forma, conclui-se que a área de CI brasileira ainda não está suficientemente envolvida com atividades de CDDC.

4.3 Tendências sobre a CD no Brasil

Uma questão foi idealizada visando levantar a opinião dos pesquisadores sobre possíveis necessidades de formação técnica curricular para praticarem a CDDC. Dentre muitas disciplinas elencadas, as principais questões necessárias apontadas pelos respondentes foram: compreensão de pesquisa com dados secundários; treinamento técnico na área; compreender como a informação é compartilhada e criada na rede; cursos complementares na área; cursos específicos sobre o tema; treinamento em plataformas de dados científicos; maior conhecimento sobre as práticas e plataformas disponíveis; conhecimento dos *softwares* de CD; maior conscientização sobre CD.

Algumas referências às disciplinas foram extraídas da grade curricular de um curso de especialização em CDDC da Universidade de Illinois em Urbana-Champaign, nos EUA e foram apresentadas aos respondentes para que eles elencassem uma classificação quanto ao grau de importância dessas disciplinas na preparação técnica de um profissional da informação para executar atividades de CDDC. As consideradas mais importantes foram Organização da informação (59%) e preservação digital (57%).

Na opinião dos respondentes, garantir a preservação de dados científicos foi a atividade mais relevante (71%) para que um profissional da informação possa interagir com pesquisadores de domínio específico. Sobre o auxílio que um profissional da informação pode fornecer a um pesquisador considerado mais relevante é atividade de encontrar dados e publicações para reúso (55%).

Entretanto, os respondentes consideram que para que um profissional atue em repositórios digitais de dados científicos, é importante que ele domine questões sobre a Ética da pesquisa científica (91%); Métodos de pesquisa (82%); Comunicação científica (83%); Propriedade intelectual (85%); Acesso a dados digitais (89%); Padrões de metadados (84%) e; Marcos legais regulatórios e políticas de direitos autorais (86%).

Em torno de 77% dos respondentes concordam com o fato de que estudos que disponibilizam dados científicos são mais citados do que aqueles que não disponibilizam. Os pesquisadores acreditam que a questão mais crítica para a pesquisa brasileira é a sustentabilidade e manutenção dos dados científicos em ambientes tecnológicos, no longo prazo (74%).

Os respondentes também opinaram sobre quais necessidades eles acreditam ter sobre a formação técnica curricular para a CDDC, sendo: compreensão de pesquisa com dados secundários; treinamento técnico; compreender como a informação é compartilhada e criada na rede; cursos complementares; cursos específicos sobre o tema; treinamento em plataformas de dados científicos; maior conhecimento sobre as práticas e plataformas disponíveis; conhecimento dos *softwares* de CD; maior conscientização sobre CD.

Por fim, ao serem provocados a emitir opinião sobre como está a CDDC no Brasil, muitos responderam que a área ainda está em fase embrionária, com poucas discussões e em passos lentos.

4.4 Análises e Discussões

Apesar dos indícios de já possuírem um conhecimento básico sobre o tema, parece não haver uma participação para aprimoramentos próprios para se envolver com o desenvolvimento da área de CD no cenário nacional. Esta afirmação pode ser corroborada com as necessidades de aprimoramento elencadas pelos próprios pesquisadores.

O percentual médio de respondentes que afirmaram nunca utilizar plataformas tecnológicas para armazenar e reutilizar dados científicos é de 49%, apesar de afirmarem conhecer repositórios de dados científicos, em torno de 99%.

Mesmo que 79% afirmar conhecer o Manifesto do acesso aberto a dados científicos lançado pelo IBICT em 2016, cerca de 84% dos pesquisadores permanecem com seus dados de pesquisas armazenados em computadores pessoais ou no *google drive*, restringindo compartilhamento aberto desses dados.

Pavão et al. (2018) afirmam que, na pesquisa brasileira, a prática de compartilhamento de dados ainda não é algo comum. As respostas que este estudo obteve confirma essa afirmativa. Somente 2% afirmam disponibilizar dados quando indagados sobre como gerenciam seus dados. Mas parece ter a intenção de deixar os dados disponíveis (46%) e de compartilhar (63%), quando indagamos as ações sobre dados produzidos em suas pesquisas. Essa percentagem é pouco menor em relação ao percentual de respondentes que reutilizam dados de pesquisa (61%).

De acordo com Pavão et al. (2018), a justificativa está na falta de condições para a preservação e segurança dos dados científicos. Estudos também alegam que o avanço do arcabouço normativo-legal referente aos repositórios institucionais de dados de pesquisa no Brasil é lento e que se assentam em âmbitos políticos, legais, econômicos e culturais (De Oliveira; Da Silva, 2016, PAVÃO et al., 2018).

Portanto, para que a área da CI possa contribuir com avanços na CDDC, são necessários esforços em várias instâncias, envolvendo, principalmente, todos os elementos que se relacionam com a ciência: próprio pesquisadores, instituições, financiadores e estado.

Em se tratando da área da CI, torna-se crítica esse cenário porque os resultados mostram que não há, ainda, uma efetiva participação em questões básicas relacionadas às atividades de CDDC com seus próprios dados. Apesar de conhecer a definição básica do tema, haja vista o alto percentual (93%) que concorda que a CD é um campo de oportunidade de carreira para o profissional da informação, e que dizem já terem citado dados científicos (49%) e de já terem publicado dados científicos (47%), conclui-se que não há um envolvimento significativo da área para assumir a área da CDDC. Os resultados sugerem que é necessária uma maior conscientização e aprofundamento conceitual nas atividades de curadoria digital. É necessário também uma maior emersão destes pesquisadores nas práticas de CDDC de forma que

contribuam, efetivamente, com o desenvolvimento da área.

Na opinião dos pesquisadores, a CDDC no Brasil está "... na sua forma embrionária, dando os primeiros passos", apesar de um dos pesquisadores mencionar uma iniciativa governamental que poderá contribuir com o desenvolvimento da área de CDDC no cenário brasileiro:

> "... em estágio embrionário, pois ainda não há uma política nacional para a gestão e compartilhamento de dados científicos. Porém, merece ser comentado que, em novembro de 2018, o Ministro Kassab (MCTIC) criou o grupo de trabalho para elaborar uma minuta de Decreto para a ciência aberta no Brasil, espaço onde se têm discutido questões complexas, como, por exemplo, papel das agências de fomento, acesso a recursos internacionais para desenvolvimento de pesquisa, disponibilização de dados que envolvem soberania nacional em repositórios internacionais. O Fato é que a curadoria de dados passará a ser uma necessidade a partir do momento de que existir uma diretriz nacional – fomento da CAPES /CNPq para a pesquisa --> dado cientifico aberto, salvo os casos de dados que envolvem propriedade industrial, soberania nacional, defesa, conhecimento tradicional dentre outros. A partir desse momento, as agências de fomento no Brasil passarão a exigir uma Plano de Gestão de Dados para o pesquisador" (Texto de um respondente na íntegra).

Portanto, há longo caminho ainda a ser percorrido para a área de CI para assumir a CDDC como parte da sua missão, no cenário brasileiro.

5. Considerações Finais

Esta pesquisa investigou a relevância e o grau de adesão que as atividades de curadoria digital de dados científicos têm para a área da Ciência da Informação no cenário nacional.

Realizou-se uma pesquisa com Survey para levantar e descrever como a CI, através de seus representantes pesquisadores, percebe o papel importante que tem, frente a este fenômeno, e também como a área está acompanhando e interagindo com o movimento internacional de CDDC.

A análise realizada visa contribuir, também, com a apresentação do cenário de atuação profissional e acadêmica dos pesquisadores da área da CI no tocante às atividades principais que compõem a CDDC. A partir dessa apresentação pode-se iniciar uma discussão mais informada acerca de estratégias para incrementar a relevância das atividades CDDC realizadas por esses profissionais, e, consequentemente, pela própria CI.

Infere-se, em relação aos dados coletados e à opinião dos pesquisadores da CI, que já há uma movimentação para adesão ao gerenciamento de dados científicos e das atividades de CDDC. Já existe uma movimentação de entendimento de que a CDDC é fundamental para a preservação, a disponibilização e a reutilização dos dados científicos.

Os dados permitem afirmar, contudo, que ainda não há uma conscientização suficientemente assimilada pelos pesquisadores da CI em relação à necessidade de envolvimento mais consciente e comprometido nessas atividades. Constatou-se claramente que ainda não há uma efetividade de práticas de CD dos próprios dados produzidos por esses pesquisadores e, portanto, não há suficiente envolvimento com a CDDC por parte desses pesquisadores para poder afirmar que a CI nacional tomará a CDDC como um pilar ou uma parte fundante de sua missão. Sobretudo, porque é necessário um domínio dessas atividades e dos conceitos envolvidos para realização dessas atividades.

Assim, concluiu-se, de forma geral, ser ainda necessária uma mudança evolutiva considerável na formação disciplinar teórica, prática e técnica desses pesquisadores. Tal conclusão é ainda mais relevante para o fortalecimento da CI enquanto área de conhecimento como um todo, pois são esses mesmos pesquisadores que formam os futuros profissionais da informação que irão atuar nas práticas profissionais e em serviços técnicos de informação.

Uma limitação desta pesquisa é que a participação dos pesquisadores da CI nacional poderia ter sido ainda maior numericamente e também mais abrangente. Mesmo assim, acredita-se

que os objetivos inicialmente propostos, foram atingidos.

Como trabalhos futuros, sugerem-se que sejam realizadas novas pesquisas que procurem gerar informações ainda mais detalhadas para subsidiar e aprofundar esta discussão, corroborando ou refutando os resultados alcançados até aqui.

O levantamento detalhado das informações descritivas do cenário da CI frente à CDDC permitirá que se estabeleçam estratégias de evolução do tratamento deste tema pela área da CI no país. Permitirá ainda discutir, de forma embasada empiricamente, sobre lacunas da formação disciplinar do profissional da informação, tanto em nível de graduação quanto pós-graduação.

Outra sugestão poderia ser a realização de pesquisas que aprofundem em temas mais específicos no domínio da CDDC, tais como a integração, o reuso, a publicação e a preservação de dados científicos; o uso de ferramentas de repositórios de dados científicos, não apenas textos (artigos, monografias, teses e dissertações).

O levantamento de informações mais detalhadas sobre tais atividades e ferramentas permitirá que a CI reflita sobre estratégias e meios para conscientizar a comunidade acadêmica para a importância da CDDC para a Ciência da Informação, bem como sobre a importância de se disponibilizar os dados produzidos em suas pesquisas.

Acredita-se que tal discussão permitirá que a CI contribua, de forma definitiva, para melhorar as taxas de reuso de dados e condições de reprodução dos resultados de pesquisas científicas em todas as áreas de conhecimento, assim como de seus próprios pesquisadores.

Referências

ALBAGLI, S. Ciência aberta em questão. **Ciência aberta, questões abertas**. Brasília: Ibict, 2014. p. 9-25.

BABBIE, E. **Métodos de Pesquisas de Survey.** Belo Horizonte: Editora UFMG, 1999. 519p

CORREIA, M. S. B. B. **Probabilidade e estatística**. 2ª ed. Belo Horizonte: PUC Minas Virtual, 2003.

DE OLIVEIRA, A. C. S.; DA SILVA, E. M. Ciência aberta: dimensões para um novo fazer científico. **Informação & Informação**, v. 21, n. 2, p. 5-39, 2016.

GIL, A.C. **Métodos e técnicas de pesquisa social.** 6ª ed. São Paulo: Atlas, 2008.

KOUPER, I. Professional participation in digital curation. **Library & Information Science Research**, v. 38, n. 3, p. 212-223, 2016.

MARCONI, M.A.; LAKATOS, E. M. **Fundamentos de metodologia científica.** 5ª ed. São Paulo: Atlas, 2003.

OLIVEIRA, E. F. T.; GRÁCIO, M. C. C. Análise a respeito do tamanho de amostras aleatórias simples: uma aplicação na área de Ciência da Informação. **Revista de Ciência da Informação**, v. 6, n. 3, p. 1-11, 2005.

PAVÃO, C. M. G. et al. **Acesso aberto a dados de pesquisa no Brasil: repositórios brasileiros de dados de pesquisa: relatório 2018.** 2018.

POOLE, A. H. The conceptual landscape of digital curation. **Journal of Documentation**, v. 72, n. 5, p. 961-986, 2016.

SARACEVIC, T. Ciência da Informação: origem, evolução e relações. **Perspectivas em Ciência da Informação,** Belo Horizonte, v. 1, n. 1, p. 41-62, jan./jun. 1996.

SAYÃO, L. F. et al. Dados abertos de pesquisa: ampliando o conceito de acesso livre. **RECIIS – Rev. Eletron. de Comun. Inf. Inov. Saúde**. 2014 jun.; 8(2) – p.76-92

SAYÃO, L. F.; SALES, L. F. Curadoria digital e dados de pesquisa. **AtoZ: novas práticas em informação e conhecimento**, v. 5, n. 2, p. 67-71, 2016.

TRIPATHI, M.; SHUKLA, A.; SONKAR, S. K. Research data Management Practices in university libraries: a study. DESIDOC **Journal of Library & Information Technology**, v. 37, n. 6, p. 417-424, 2017.

A QUALIDADE DA INFORMAÇÃO EM ONTOLOGIAS TEMPORAIS NO CONTEXTO DE GERENCIAMENTO DE EMERGÊNCIAS

Information Quality in Temporal Ontologies in the Context of Emergency Management

Gustavo Marttos Cáceres Pereira[1], Leonardo Castro Botega[1]
(1) Universidade Estadual Paulista "Júlio de Mesquita Filho", Av. Hygino Muzzi Filho, 737, Marília/SP, gustavo.marttos@unesp.br, leonardo.botega@unesp.br

Resumo:
No contexto de gerenciamento de emergências, onde as informações são provenientes de fontes heterogêneas, é necessário que as tomadas de decisões sejam assertivas e dentro de um intervalo de tempo hábil. O tempo possui grande relevância por ser fundamental no domínio deste contexto, pois é criando uma linha do tempo, ou seja ao decorrer dele, que se torna viável a percepção e compreensão de todas as características de uma situação. A qualidade da informação torna-se imprescindível no contexto de gerenciamento de emergências, considerando a complexidade e dinamicidade dos dados. Este trabalho objetiva a aplicação da dimensão temporal em uma ontologia de domínio visando identificar as alterações comportamentais nas métricas e demais dimensões qualitativas. A natureza deste trabalho é qualitativa, de finalidade teórico-aplicada e de tipo exploratória, sendo sua metodologia guiada por um estudo de caso único, avaliando o comportamento da qualidade informação e buscando inferir novos conhecimentos temporais para que sirvam de insumos para tomadas de decisões mais assertivas.
Palavras-chave: Qualidade da Informação; Ontologia; Gerenciamento de Emergências; Recuperação da Informação.

Abstract:
In the context of emergency management, where information comes from heterogeneous, complex and dynamic sources, decision making is required to be assertive and within a timely interval of time. Time has great relevance, being fundamental in the domain of this context, because it is creating a timeline, that is to say during it, the perception and comprehension of all the characteristics of a situation becomes viable. The quality of information becomes indispensable in the context of emergency management, mainly by dynamic and complex factors. This work aims to apply the temporal dimension in a domain ontology to identify behavioral changes in the metrics and other qualitative dimensions. The nature of this work is qualitative, of theoretical-applied purpose and exploratory type, being its methodology guided by a single case study, evaluating the behavior of quality information and seeking to infer new temporal knowledge so that inputs for more assertive decision-making.
Keywords: Information Quality; Ontology; Emergency Management; Information Retrieval.

1. Introdução

A dimensão temporal é um artefato de grande importância e relevância no domínio de gerenciamento de emergências. Entretanto, ambientes informacionais que dependam de tal métrica, tal qual a atualidade dos dados, estão sujeitos a limitações de qualidade, as quais podem causar riscos à vida, ao meio ambiente e ao patrimônio (TAO *et al*., 2010).

Portanto, torna-se um fator crucial para o entendimento de situações críticas e a tomada de decisão de operadores de emergências.

Para que a análise de situações de emergência seja mais assertiva, é indispensável a presença de elementos que indiciem a qualidade da informação, principalmente devido à sua heterogeneidade, complexidade e dinamicidade (BOTEGA *et al*., 2017).

Sob a perspectiva da Ciência da Informação, Nehmy e Paim (1998) e Oleto (2006) discorrem sobre a qualidade da informação sendo conceituada como um conjunto de atributos relacionados e mensuráveis em relação ao valor informacional.

Em trabalhos anteriores dos autores, Silva *et al*. (2018) discorrem a respeito das características dos dados emergenciais, os quais são heterogêneos, imprevisíveis, complexos e dinâmicos, o que acaba por limitar a representabilidade e a recuperabilidade de tais dados por meio de modelos computacionais sintáticos. A

informação proveniente de tais dados tem sua qualidade comprometida, afinal ela pode estar incompleta, imprecisa e difusa.

Previamente, neste mesmo trabalho, uma ontologia de domínio foi desenvolvida para o contexto de gerenciamento de emergências, na qual foi aplicada uma metodologia de gestão de qualidade para qualificar e quantificar as informações utilizadas na mesma, contribuindo para a melhoria dos processos de inferência sobre situações de emergências. Entretanto, não houve o emprego e análise da dimensão temporal em sua estrutura.

Desta maneira, o objetivo deste trabalho é a inserção e análise da dimensão temporal nesta ontologia de domínio, a fim de identificar possíveis demandas de alterações nas demais dimensões e métricas qualitativas, principalmente o que remete à atualidade e relevância informacional.

A partir de resultados representados e recuperados na ontologia de Silva *et al*. (2018), espera-se que a inclusão da dimensão temporal viabilize, por meio de inferências baseadas em novos relacionamentos propostos entre os elementos da ontologia, novas descobertas informacionais, podendo servir como insumos para a melhoria do processo de tomada de decisão.

O arcabouço metodológico se sustenta sob natureza qualitativa, de finalidade teórico-aplicada e de tipo exploratória. A pesquisa é orientada a um estudo de caso único referente ao uso de ontologia de domínio no contexto de gerenciamento de emergências e a avaliação do comportamento da qualidade da informação, onde a dimensão temporal foi incluída em sua estrutura.

Serão apresentados as contextualizações de qualidade da informação na Seção 2, ontologias temporais na Seção 3, a discussão sobre como as ontologias temporais afetam a qualidade da informação na Seção 4, juntamente com um estudo de caso e, por fim, as considerações finais na Seção 5.

2. Qualidade da Informação

Nehmy e Paim (1998) e Oleto (2006), como dito anteriormente, argumentam a respeito da qualidade da informação como produto, isto é, um conjunto de atributos relacionados, mensuráveis e multidimensionais.

Para Oleto (2006), o conjunto de atributos consiste em diversas relações entre eles, tais como a abrangência, acessibilidade, atualidade, confiabilidade, objetividade, precisão e validade.

Pereira Junior, Pereira e Botega (2019) também vêm a qualidade da informação como produto, entretanto enfatizam que ela é variável e que pode ser subjetiva, pois deve ser definida de acordo com as necessidades informacionais requeridas pelo domínio, evitando possíveis problemas ou falhas de interpretação.

Aspectos internos do domínio crítico e externos podem afetar diretamente a qualidade da informação, uma vez que há diversas fontes de dados, tornando-as heterogêneas, complexas e dinâmicas.

Calazans (2008) argumenta que a falta da qualidade da informação pode causar impactos, devendo ser diagnosticados, providenciando soluções o quanto antes.

No contexto de gerenciamento de emergências, Botega *et al*. (2019) argumentam que a qualidade da informação pode beneficiar tanto os processos automatizados, como por exemplo as inferências de uma ontologia, quanto a compreensão humana perante situações de emergências. Isto ocorre devido a presença das dimensões qualitativas, pois podem auxiliar os operadores quanto à confiabilidade informacional.

3. Ontologias Temporais

Tao *et al*. (2010) discorrem sobre a importância da inclusão da dimensão temporal em uma ontologia que já possui alguma característica qualitativa. Para os autores, esta dimensão é fundamental para o raciocínio temporal, ou seja, respostas que podem mudar ao decorrer do tempo e criam, portanto, uma linha do tempo que pode ser analisada durante sua recuperação e representação e consequentemente pode servir de insumo para que os operadores tenham uma melhor percepção e compreensão acerca das informações.

O modelo proposto pelos autores citados acima conta com duas principais classes OWL (Web Ontology Language): Evento e Tempo. A primeira refere-se a qualquer tipo de ocorrência, estado, percepção, procedimento, sintoma ou situação que ocorra em uma linha do tempo. A segunda é dividida em outras quatro classes: Instante, Intervalo, Fase e Período.

A classe Instante refere-se a um ponto específico de tempo dentro de uma linha do tempo, onde existem fatores granulares, como data (ano, mês e dia) e horário (hora, minuto e, se necessário, segundo). Essas granularidades permitem que a linha do tempo seja representada e recuperada de maneira correta pela ontologia, além de auxiliar nos processos de inferências para que novos conhecimentos temporais sejam descobertos.

A classe Intervalo representa a duração de tempo, ou seja, há um relacionamento de início e fim. Cada parte do relacionamento torna-se uma instância de Instante.

A classe Fase representa cada ocorrência de um intervalo repetido, também tendo início e fim. Por fim, a classe Período específica a medida de frequência que uma Fase repete.

Toda informação que remete à horário, independente de qual classe temporal for, deve ser representada pela classe Duração, onde deve-se conter a unidade de tempo utilizada junto ao seu respectivo valor. A unidade de tempo é dada pelo fator granular mencionado acima, isto é, pode ser “ano”, enquanto seu valor é “2019”.

Apesar dessas classes estarem presentes no modelo semântico, elas não cumprirão seus objetivos se não houver um relacionamento consistente entre elas. Portanto, o relacionamento temporal se dá entre duas instâncias de Evento ou de Evento com alguma instância de Tempo.

Não obstante, Okeyo, Chen e Wang (2014) reiteram a importância do relacionamento temporal, pois de acordo com seus estudos, representar conhecimento temporal usando OWL é um desafio, pois esta tecnologia suporta apenas relações unárias e binárias, enquanto uma relação temporal depende de, no mínimo, uma relação ternária.

Com o relacionamento temporal estabelecido, pode-se inferir novas informações, obtendo conhecimento temporal. Para tanto é necessário que todas as instâncias estejam com seus atributos granulares, pois assim a linha do tempo pode se formar.

Uma granularidade é a normalização de datas, ou seja, deixar as datas de modo que sejam interpretáveis por mecanismos computacionais. Uma expressão de tempo dada por “*dois dias atrás*” deve ser normalizada para “*2019-09-12*”. Outras expressões, como “*antes*”, “*depois*” e “durante” também são válidas (HASANUZZAMAN *et al*., 2014).

De acordo com Tao *et al*. (2010), a dimensão temporal em relação à análise de dados emergenciais possui diversas aplicabilidades, tais como (1) a descoberta de padrões temporais em uma situação de incêndio florestal em um determinado bioma; (2) a explicação de situações passadas, buscando trazer as prováveis causas que acarretam em situações de emergência; e (3) projeção de estados futuros, como a possibilidade do fogo de um incêndio florestal se alastrar para outras áreas.

4. Qualidade da Informação em Ontologias Temporais no Contexto de Gerenciamento de Emergências

No trabalho proposto por Silva *et al*. (2018), a metodologia utilizada no desenvolvimento da ontologia de domínio engloba os aspectos informacionais relativos à qualidade. Entretanto não há a verificação temporal junto à ontologia, ou seja, apesar de existir a dimensão de atualidade, a qual se refere ao ritmo de produção informacional ao decorrer do tempo, ela não é devidamente expressada de forma ontológica visando a criação da dimensão temporal.

Os autores definiram as métricas e dimensões utilizando a metodologia IQESA (*Information Quality Assessment Methodology in the Context of Emergency Situational Awareness*) proposta por Botega *et al*. (2017), ilustrando todas as fases para avaliar e representar a qualidade como parte de um processo de avaliação de informações no contexto de gerenciamento de emergências.

A IQESA faz uso das dimensões qualitativas: atualidade, completude, consistência, relevância e certeza. Cada dimensão conta com uma fórmula específica para que seja possível quantificar seu índice qualitativo.

No estudo de caso, os autores argumentam que a dimensão de consistência só pode ser calculada a partir do segundo evento emitido sobre a mesma situação, desde que a confiabilidade da informação seja positiva.

Percebe-se, então, que a dimensão temporal poderia ser útil e altamente relacionável nesse cenário, pois em um primeiro momento há a emissão de um evento que se refere a uma situação com índices qualitativos baixos, enquanto no segundo momento há um novo evento que complementa as informações desta situação, melhorando seu índice qualitativo.

Desta maneira, propõe-se a implementação da dimensão temporal, de acordo com as especificações de Tao *et al*. (2010) e Okeyo, Chen e Wang (2014), baseando-se no estudo de caso de Silva *et al*. (2018).

Dois eventos foram criados. O primeiro foi emitido por um cidadão, é do tipo alerta, possui confiabilidade do emissor e o horário da denúncia foi às 14h23. O segundo foi emitido por um bombeiro, também é do tipo alerta e possui confiabilidade do emissor, sendo o horário da denúncia às 14h25.

Ambos os eventos podem ser instâncias de Evento, conforme proposto por Tao *et al*. (2010), portanto o relacionamento temporal pode ser criado ao vincular uma instância de Instante, a qual se refere a um momento de tempo.

A unidade de tempo das duas instâncias deve ser especificada como horário e os atributos de tempo precisam ter seus valores normalizados para serem interpretados por máquinas, portanto os horários "*14h23*" e "*14h25*" devem ser normalizados para "*14:23*" e "*14:25*", respectivamente, conforme a ISO 8601 - Formato de Data e Hora.

A representação das instâncias de eventos e do relacionamento temporal é apresentada pela Figura 1A e Figura 1B do Apêndice A.

De acordo com o estudo de Silva *et al*. (2018), ocorreram dois eventos ("*evento_1*" e "*evento_2*") os quais se referem a mesma situação ("*situacao_1*"), entretanto, caso não houvesse a dimensão, não seria possível dizer ontologicamente qual evento ocorreu antes do outro, por exemplo.

Dada a inclusão da dimensão temporal, há a possibilidade de inferir novos conhecimentos temporais a partir dos novos relacionamentos criados, como por exemplo o Evento 1 ter ocorrido antes do Evento 2 utilizando a propriedade "*time:hasNormalizedTime*" das instâncias "*instante_1*" e "*instante_2*".

As métricas qualitativas e quantitativas são afetadas, uma vez que uma das dimensões de qualidade passou a ter relacionamentos ternários passíveis de novas mensurações.

5. Considerações finais

No domínio de gerenciamento de emergências, sabe-se que utilizar uma ontologia como modelo semântico para melhorar a recuperabilidade e representabilidade da informação é vantajoso em ambientes informacionais que adotam sistemas de apoio à tomada de decisão.

Acoplar a dimensão de tempo, mudando a característica da ontologia para temporal pode ser útil para inferir novos conhecimentos, além de tornar todos os atributos qualitativos como orientados ao tempo.

Isto possibilita a avaliação a qualidade da informação ao decorrer do tempo, ou seja, em uma linha temporal, permitindo que novas análises sejam realizadas a fim de proporcionar aos operadores melhores insumos para suas tomadas de decisões.

Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Referências

BOTEGA, Leonardo Castro et al. Methodology for data and information quality assessment in the context of emergency

situational awareness. **Universal Access in the Information Society**, v. 16, n. 4, p. 889-902, 2017.

BOTEGA, Leonardo Castro *et al*. Quantify: An Information Fusion Model Based on Syntactic and Semantic Analysis and Quality Assessments to Enhance Situation Awareness. In: **Information Quality in Information Fusion and Decision Making**. Springer, Cham, p. 563-586, 2019.

CALAZANS, Angelica Toffano Seidel. Qualidade da informação: conceitos e aplicações. **TransInformação**, v. 20, n. 1, p. 29-45, 2008.

HASANUZZAMAN, Mohammed *et al*. Propagation strategies for building temporal ontologies. In: **Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers**. p. 6- 11, 2014.

NEHMY, Rosa Maria Quadros; PAIM, Isis. A desconstrução do conceito de" qualidade da informação". **Ciência da Informação**, v. 27, n. 1, 1998.

OKEYO, George; CHEN, Liming; WANG, Hui. Combining ontological and temporal formalisms for composite activity modelling and recognition in smart homes. **Future Generation Computer Systems**, v. 39, p. 29-43, 2014.

OLETO, Ronaldo Ronan. Percepção da qualidade da informação. **Ciência da informação**, v. 35, n. 1, 2006.

PEREIRA JUNIOR, Valdir Amancio; PEREIRA, Gustavo Marttos Cáceres; BOTEGA, Leonardo Castro. Towards a Process for Criminal Semantic Information Fusion to Obtain Situational Projections. In: **The Human Position in an Artificial World: Creativity, Ethics and AI in Knowledge Organization**. Ergon-Verlag p. 51-72, 2019.

SILVA, Jordana Nogueira *et al*. Desenvolvimento de ontologia ciente de qualidade de informações para o domínio de gerenciamento de emergências. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, Florianópolis, v. 23, n. 53, p. 184-200, 2018.

TAO, Cui *et al*. CNTRO: a semantic web ontology for temporal relation inferencing in clinical narratives. In: **AMIA annual symposium proceedings**. American Medical Informatics Association, p. 787, 2010.

6. Apêndice A – Representação de instância da ontologia com a dimensão temporal

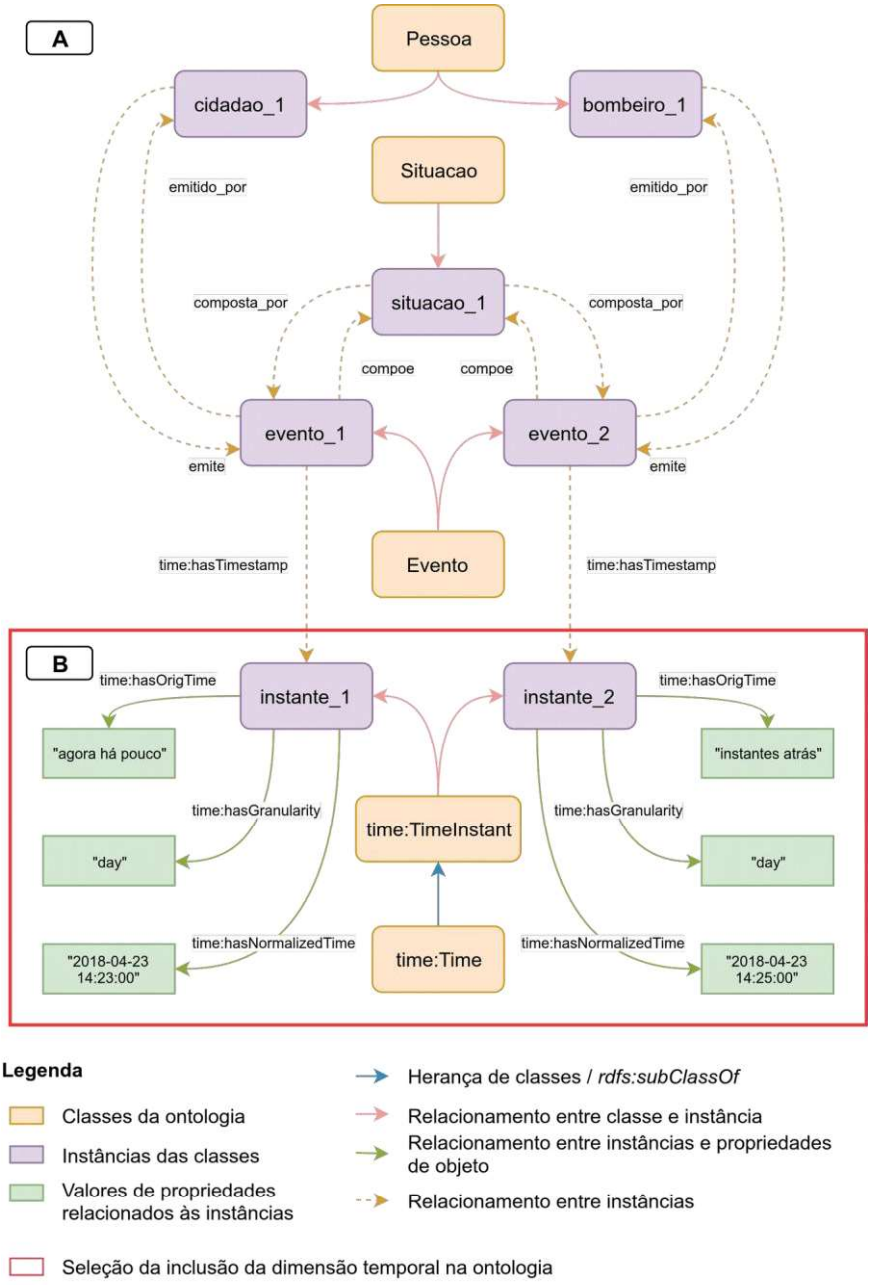


Figura 1: (A) Representação da ontologia sem a dimensão temporal. Os objetos com preenchimento na cor laranja representam as classes da ontologia de domínio proposta por Silva *et al*. (2018), enquanto os objetos com preenchimento na cor roxa representam as instâncias dessas classes, as quais remetem ao estudo de caso proposto na Seção 4. As setas contínuas rosas referem-se a qual classe pertence uma instância específica, enquanto as linhas tracejadas laranjas referem-se ao relacionamento entre instâncias de outras classes. **(B)** Representação da dimensão temporal relacionada à ontologia. Os objetos com preenchimento na cor verde são os valores das propriedades relacionados às instâncias. A linha contínua azul refere-se à subclasse e a contínua verde remete ao vínculo entre propriedade e valor específico. Para fins de destaque, a inclusão da dimensão temporal é representada pelo todo que se encontra dentro da borda vermelha.

ANOTAÇÃO DE DADOS PARA GERAÇÃO DE INDICADORES DE DESEMPENHO EM ORGANIZAÇÕES

Data Annotation for Generating Performance Indicators in Organizations

Marcello Peixoto Bax[1], Evaldo de Oliveira da Silva[2]

(1) Programa de Pós-Graduação em Ciência da Informação – UFMG, Av. Pres. Antônio Carlos, 6627 - Pampulha, Belo Horizonte - MG, 31270-901, bax@eci.ufmg.br

(2) Centro de Ensino Superior de Juiz de Fora (CESJF), Rua Halfeld, 1.179, Centro Campus Academia - MG, 36016-000, evaldosilva@cesjf.br.

Resumo:

Key Performance Indicators (KPIs) são usados por organizações para avaliar o exercício de suas atividades, apoiando a decisão. Com base nesses indicadores, elas revêem seus processos buscando a melhoria contínua das atividades. Modelos de dados dimensionais estruturam os dados agrupados em "fatos" e "dimensões". Os fatos são representados por campos numéricos que alavancam a geração de KPIs. Observa-se, contudo, a necessidade de boas práticas para nomear e anotar dados com metadados. Assim, diferentes usuários compreendem melhor o conjunto de dados, evitando interpretações divergentes. Descreve-se um processo de anotação semântica usando dicionário de dados, que associa dados a conceitos, permitindo a geração de KPIs. Apresenta-se como se dá a geração de desses indicadores pelo enriquecimento semântico dos dados com ontologias.

Palavras-chave: Modelos Dimensionais, Indicadores de Desempenho, KPI, Dicionário de Dados, Ontologia, Anotação Semântica

Abstract:

Key Performance Indicators (KPIs) are used by organizations to evaluate the performance of their activities and decision support. Based on these indicators, they review their processes seeking continuous improvement of activities. Dimensional data models structure data grouped into "facts" and "dimensions." Facts are represented by numeric fields that leverage the KPIs generation. There is, nevertheless, a need for good practices of naming and annotating data with metadata. Thus, different users better understand the dataset, avoiding divergent interpretations. We describe a semantic annotation process using data dictionary, which associates data with concepts, allowing the generation of KPIs. We present how these indicators are generated by semantic enrichment of data with ontologies.

Keywords: Dimensional Data, Performance Indicators, KPI, Data Dictionary, Ontology, Semantic Annotation

1. Introdução

Um indicador chave de desempenho (KPI, *Key Performance Indicador*) é um valor que pode ser medido e que demonstra a eficácia da organização em alcançar resultados (PARMENTER, 2015). KPIs permitem avaliar o atingimento de metas, avaliar resultados e rever processos capacitando a melhoria contínua das atividades. Valores KPI criam base analítica para tomada de decisões que priorizam ações avaliadas (empiricamente) como as mais relevantes.

KPIs são, p.ex., receitas, lucros, preços e custos, medidas de qualidade ou satisfação. Gestores e executivos interpretam KPIs para decidirem com base científica, empírica. Exemplo comum de mensuração é o percentual de aderência da realização de atividades com o planejamento. KPIs podem ser vistos também no meio acadêmico. De acordo com o *Central European Research Infrastructure Consortium* (CERIC) KPIs podem avaliar o grau do alcance de objetivos de instituições de ensino ou programas de pesquisa. KPIs são insumos para gerenciar e monitorar o atingimento de objetivos e auxiliar o planejamento estratégico (KOLAR, HARRISON e GLIKSOHN, 2019).

Para Kimball e Ross (2013) a criação de KPIs deve ser disciplinada em suas práticas de nomeação de dados. Assim, caso seja impossível entender o conjunto de dados (*datasets*) a ser utilizado para gerar os cálculos, nomes diferentes serão atribuídos a diferentes interpretações. Com isso, os KPIs acabam resultando de combinações de dados incompatíveis, comprometendo os valores e prejudicando a tomada de decisão.

É necessário garantir a qualidade dos dados (MEDEIROS, 2018) e a Curadoria Digital propõe técnicas de descrição com metadados que favorecem qualidade, preservação e facilitam a descoberta de novos conhecimentos pelo reúso de dados. No entanto, somente a definição dos metadados não basta para extrair e compartilhar *datasets*. Dados usados para geração de KPIs podem vir de estruturas e modelos de dados distintos e requerem informações adicionais para que seus significados sejam explicitados. Bastante aplicados na descrição de *datasets*, os dicionários de dados apoiam atividades de gerenciamento, procedimentos de conversão, validação e critérios para armazenar dados. Ontologias e tecnologias semânticas enriquecem e formalizam o significado dos KPIs, evitando interpretações discrepantes.

Desta forma, descreve-se aqui um processo de anotação baseado em Dicionários Semânticos de Dados (SDDs) que contribui com a curadoria, dentre outros elementos, por estar alinhado com princípios FAIR (*Findable*, *Accessible*, *Interoperable*, *Reusable*) de Wilkinson et. al, 2016). O "caso de uso" apresentado neste artigo anota os dados de um modelo dimensional para cálculos de KPIs.

A Seção 2 traz o conceito de modelagem dimensional de dados para KPIs e trabalhos correlatos. A Seção 3 descreve o processo de anotação proposto por Rashid et. al (2017). A Seção 4 relata a anotação para criação de um KPI para monitoramento de projetos de pesquisa. A Seção 5 faz considerações finais e sugere trabalhos futuros.

2. Modelagem de KPIs

Um modelo de dados dimensional é construído, agrupando dimensões ao redor de dados numéricos. Os fatos são estruturados relacionando dados e suas dimensões. A análise dos fatos usa as dimensões (facetas) combinando filtros que atendem as necessidades do usuário, na tomada de decisão (KIMBAL e ROSS, 2013). O modelo dimensional da Figura 1 (Apêndice A) usa esquema estrela e permite o cálculo dos montantes de publicações pelas dimensões: fator de impacto, centro de pesquisa, mês e ano.

2.1. A Ontologia KPIOnto

A anotação pelo SDD, exige compreender o domínio por sua modelagem conceitual prévia. Deve-se selecionar os dados e encontrar os termos/vocabulários existentes que referenciam os conceitos do domínio, explicitando e formalizando sua semântica com o uso de ontologias. A anotação ontológica, permite a geração de fragmentos (declarações em formato de triplas) do conhecimento do domínio em RDF (*Resource Description Framework*).

Diamantini, Potena e Storti (2016) propõem a KPIOnto que usamos para anotação e alinhamento conceitual de diferentes profissionais sobre os KPIs. A KPIOnto constitui-se de classes como: Indicador, Dimensão e Fórmula; sendo "Indicador" a classe principal. Ela especifica um indicador pelas propriedades: hasDimension, hasFormula e hasAggrFunction (para uso de funções de agregação).

2.2. Dicionário Semântico de Dados

Rashid et. al (2017) utiliza padrões de metadados para configurar a anotação semântica por um SDD. Recomenda ainda a utilização da ontologia SIO (S*emanticscience Integrated Ontology*) que fornece propriedades para descrever os relacionamentos entre objetos e atributos como modelo de representação do conhecimento. A anotação semântica proposta por Rashid et. al (2017) utiliza os seguintes documentos:

- *InfoSheet*: referências para descrição dos SDDs;
- *Dictionary Mapping*: anotação semântica das colunas das coleções de dados;
- *CodeBook*: códigos correspondentes a conceitos de ontologia;
- *Code Mapping*: mapeamento de termos dos *datasets* que correspondem a conceitos existentes na ontologia;
- *TimeLine*: anotação de intervalos temporais;
- *Properties Table*: para fins de customizar a descrição por outras ontologias de topo.

A ferramenta sdd2rdf (SEMANTIC DATA DICTIONARY, 2019) interpreta o SDD e "ingere" os dados, formando um grafo RDF. Para acessar os dados anotados, o sdd2rdf

cria consultas no formato SPARQL[1] . São geradas também regras SWRL[2] que auxiliam em novas inferências. O grafo RDF gerado pelo script *sdd2rdf* utiliza o vocabulário formal ontológico, e possibilita a interoperabilidade dos dados.

2.3. Trabalhos Correlatos

Kritikos (2017) descreve que os dados vinculados (*Linked Data*) representam um grande mecanismo para a integração de informações entre fontes distintas, permitindo a realização de inferências para derivar conhecimento. Utiliza esta ideia no contexto do processo de negócios como serviço (BPaaS) a fim de coletar e vincular informações originadas de diferentes sistemas. Propõe o uso de ontologias principais que visa melhorar a comparação de KPIs gerados dos dados integrados entre os sistemas.

Wetzstein, Ma e Leymann (2008) propõe que KPIs sejam modelados por analistas de negócios que exploram anotações semânticas de processos de negócios. Os modelos de KPI são automaticamente calculados para serem geridos por meio de um painel de monitoramento em tempo real.

Kourtesis e Alvarez-Rodrigues (2014) sugerem uma estrutura semântica para gerenciamento de QoS (*Quality of Service*). Utilizam abordagens para o gerenciamento de QoS baseado em semântica, bem como os principais métodos, técnicas para explorar diversos dados.

Silva et. al. (2018) propõe um conjunto de funções para compor a estrutura semântica para definição de dicionário de dados. Apresenta ainda como estrutura semântica está relacionada à configuração sintática dos dicionários de dados, a fim de identificar padrões que possam ser usados no desenvolvimento de procedimentos para extração de informações e modelos semânticos.

3. Processo de Anotação Semântica

A anotação baseia-se em Rashid et. al (2017), que segue princípios FAIR e permite gerar o grafo RDF (*script sdd2rdf*) persistido no *triplestore*. A ontologia formaliza o vocabulário e abre caminho para interoperabilidade de dados. Após escolher que dados do *dataset* anotar, segue-se para a criação dos artefatos abaixo, em cada etapa do processo:

1. Ontologia de domínio. Criação/ajuste de ontologia de domínio para formalização dos conceitos tratados no problema de pesquisa. Buscar reutilizar ontologias consolidadas no domínio do problema.
2. *Dictionary Mapping*. Cada linha do DM mapeia uma coluna do *dataset*, formalizando-a conceitualmente e também suas relações e proveniência.
3. *CodeBook*. Permite a criação dos seguintes campos: Coluna (entidade a ser anotada), Código, Descrição e a Classe da Ontologia.
4. *Infosheet*. Metadados de um SDD que organiza e descreve a coleção de arquivos de metadados (planilhas do Excel) usados pelo SDD em questão.
5. Grafo RDF. Interpretação da dupla: "SDD + Dados" pelo script sdd2rdf, gerando o RDF e armazenando-o em *triplestore* para consulta posterior.

Os dados dos objetos mapeados pelo SDD são as colunas do próprio *dataset*. Porém, Rashid et al. (2017) afirmam que os objetos descritos no *dataset* podem encontrar-se ali explícita ou implicitamente. Ou seja, no mesmo *dataset* podem aparecer também atributos de outros objetos implicitamente representados alí. Estes objetos serão explicitados no SDD e formalizados no grafo final gerado (pelo script *sdd2rdf*), favorecendo a sua integração nos níveis conceituais (ou intencionais) mais abstratos do projeto.

4. Anotação de Dados e Geração de KPIs

Descreve-se exemplo de anotação de dados para geração de KPIs a partir da necessidade do acompanhamento de índices de publicação em centros de pesquisa. O modelo da Figura 1 foi utilizado como fonte de dados. Abaixo segue a descrição da execução do processo de anotação:

[1] SPARQL Protocol and RDF Query Language

[2] Semantic Web Rule Language

Coleta de dados. O *dataset* a ser anotado foi obtido por uma *view* criada a partir do modelo da Figura 1.

Dictionary Mapping (DM). O DM (Tabelas 2 e 3) mapeia para ontologias (Sio e KPIOnto) as seguintes características dos KPIs: ResearchField, ImpactFactor e PubQuantity. A Tabela 1 traz o Codebook, que descreve os dados categoriais do *dataset*: DTempo, DFatorImpacto e DCentroPesquisa.

Criação da tabela de *Infosheet*. A *Infosheet* (Tabela 4) possui as propriedades:

- dct:creator: Responsável pelo preenchimento.
- dct:contributor: Contribuidores na criação do *Infosheet* e execução do processo.
- dct:created: Data de criação.
- dct:description: Propósito do *Infosheet*.
- owl:imports: Endereço da Ontologia.
- schema:keywords: Palavras-chave.
- dct:publisher: Responsável por publicar.
- dct:title: Título do *Infosheet*.

Grafo RDF. Geração dos RDFs para representar os fragmentos de conhecimento a respeito do KPI (do exemplo apresentado). O RDF foi persistido no Virtuoso (ERLING e MIKHAILOV, 2009).

Visualização dos dados. Um *dashboard* genérico conecta-se ao Virtuoso, via ODBC[3], e executa consultas SPARQL para ilustrar como os dados, extraídos do grafo, aparecem no *dashboard* (Figura 2).

5. Considerações Finais

O processo especificado neste trabalho visa organizar etapas para anotação com dicionários semânticos de dados para gerar fragmentos de conhecimento em RDF, i.e., conjunto de fatos originados da combinação de dados de diferentes fontes. Um exemplo usando um modelo dimensional para geração de KPI na área de publicação de pesquisa ilustrou o processo constituindo uma validação preliminar do método ("prova de conceito").

No contexto organizacional, a modelagem conceitual adequada dos dados envolve questões complexas de interpretação conceitual e negociação de significados sobre entidades, relacionamentos e regras de negócios, todas envolvidas no processo de comunicação entre as "partes interessadas". Argumentou-se neste texto sobre como o processo, fundamentado em SDDs, contribui para organização e integração conceitual dos dados oriundos de diferentes nichos da organização, gerando informações que fundamentam a estruturação de conhecimentos sobre diversos indicadores empresariais (KPIs). Isso facilita os alinhamentos dos KPIs a partir de uma abordagem de modelagem de dados ampla, do tipo *top down*, e não apenas *bottom up*.

Futuras pesquisas investigarão as possibilidades da modelagem com SDD, tal como foi apresentada, constituir alternativa superior à modelagem dimensional do tipo *"data mart*" ou "*data warehouse*". Espera-se poder alavancar a flexibilidade de modelos conceituais "livres de esquemas" (*schema free*) para facilitar a geração de KPIs. Isso tornaria a evolução do conhecimento sobre os indicadores de desempenho das organizações mais flexível, incremental e conceitualmente enriquecido, agregando ainda a explicitação da semântica formal, advinda do uso de ontologias representadas em Lógica de Descrições (*Description Logic*).

Referências

DIAMANTINI, C., POTENA, D. and STORTI, E. SemPI: A Semantic Framework for the Collaborative Construction and Maintenance of a Shared Dictionary of Performance Indicators. Future Generation Computer Systems (FGCS), vol. 54, pages 352-365, Elsevier, 2016.

ERLING, Orri; MIKHAILOV, Ivan. RDF Support in the Virtuoso DBMS. In: Networked Knowledge-Networked Media. Springer, Berlin, Heidelberg, 2009. p. 7-24.

KIMBALL, Ralph; ROSS, Margy. The data warehouse toolkit: The definitive guide to dimensional modeling. John Wiley & Sons, 2013.

KOLAR, Jana; HARRISON, Andrew e GLIKSOHN, Florian. Key performance indicators of Research Infrastructures. Disponível em: https://www.ceric-eric.eu/2018/08/30/key-performance-

[3]Open Database Conectivity

indicators-of-research-infrastructures/. 30 de Ago de 2018.
KOURTESIS, Dimitrios; ALVAREZ-RODRÍGUEZ, Jose María; PARASKAKIS, Iraklis. Semantic-based QoS management in cloud systems: Current status and future challenges. Future Generation Computer Systems, v. 32, p. 307-323, 2014.
KRITIKOS, Kyriakos; PLEXOUSAKIS, Dimitris; WOITSCH, Robert. Towards Semantic KPI Measurement. In: CLOSER. 2017. p. 63-74.
MEDEIROS, Claudia B. Gestão de Dados Científicos – da coleta à preservação. Disponível em https://blog.scielo.org/blog/2018/06/22/gestao-de-dados-cientificos-da-coleta-a-preservacao/#.XXZ82ChKjIV. Acesso em 04 de Set de 2019.
PARMENTER, David. Key performance indicators: developing, implementing, and using winning KPIs. John Wiley & Sons, 2015.
RASHID, Sabbir M. et al. The Semantic Data Dictionary Approach to Data Annotation & Integration. In: SemSci@ ISWC. 2017. p. 47-54.
SEMANTIC DATA DICTIONARY. SDD Specification. Disponível em: https://github.com/tetherless-world/SemanticDataDictionary. Acesso em 22 de set de 2019.
SILVA, Vivian S.; HANDSCHUH, Siegfried; FREITAS, André. Categorization of semantic roles for dictionary definitions. arXiv preprint arXiv:1806.07711, 2018.
WETZSTEIN, Branimir; MA, Zhilei; LEYMANN, Frank. Towards measuring key performance indicators of semantic business processes. In: International Conference on Business Information Systems. Springer, Berlin, Heidelberg, 2008. p. 227-238.
WILKINSON, M. D., DUMONTIER, M., AALBERSBERG, I. J., APPLETON, G., Axton, M., BAAK, A., and BOUWMAN, J. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific data, 3.

Apêndice A

Figura 1. Exemplo de modelo dimensional de dados.

Fonte: Elaborada pelos autores.

Tabela 1. Codebook as dimensões DTempo, DFatorImpacto e DCentroPesquisa.

Column	Code	Label	Class
DCentroPesquisa	1	CIÊNCIA DA INFORMAÇÃO	kpionto:researchField
DCentroPesquisa	2	CIÊNCIA DA COMPUTAÇÃO	kpionto:researchField
DCentroPesquisa	3	LINGUÍSTICA	kpionto:researchField
...			
DFatorImpacto	40	IMPACTO ENTRE 2 E 4	kpionto:ImpactFactor
DFatorImpacto	41	IMPACTO ENTRE 5 E 7	kpionto:ImpactFactor
DFatorImpacto	42	IMPACTO ENTRE 7 E 10	kpionto:ImpactFactor
...			

Fonte: Elaborada pelos autores.

Tabela 2. Especificação do DM para dados explícitos.

Column	**Attibute**	**Label**	**AttributeOf**
Id_Kpi	sio:Identifier	Identificador do KPI	??KpiPublication
ResearchField	kpiOnto:hasDimension	Centro de Pesquisa	??KpiPublicationo
ImpactFactor	kpiOnto:hasDimension	Nível do Fator de Impacto	??KpiPublication
PubQuantity	kpiOnto:hasAggFunction	Quantidade de Publicação	??KpiPublication

Fonte: Elaborada pelos autores.

Tabela 3. Especificação do DM para dados implícitos.

Column	Entity	Role	InRelationTo
??KpiPublication	kpiOnto:Indicator		??researchInstitute
??researchInstitute	sio:institute		

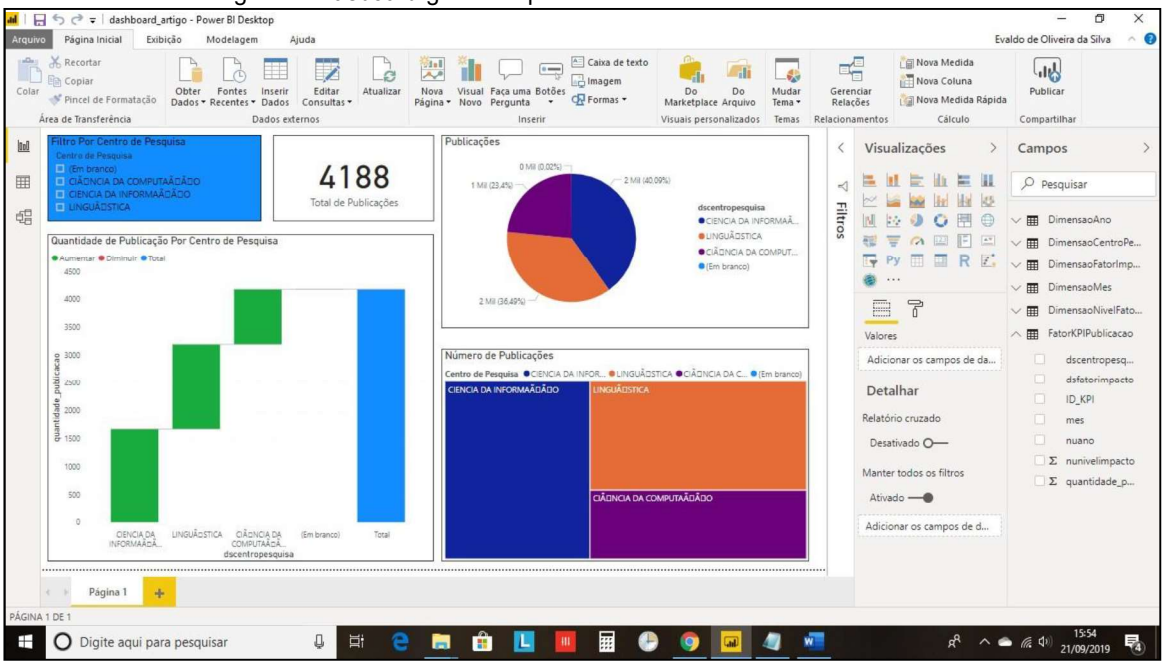
Fonte: Elaborada pelos autores.

Tabela 4. Especificação do DM para dados implícitos.

Atributo	Valor
dct:creator	Marcello P. Bax e Evaldo de Oliveira da Silva
dct:contributor	Marcello P. Bax
dct:created	19/09/2019
dct:description	Anotação semântica do dicionário de dados para geração do KPI de Publicação
owl:imports	http://semanticscience.org/ontology/sio-subset-labels.owl
schema:keywords	KPI, Publicação
dct:publisher	Evaldo de Oliveira da Silva
dct:title	Geração de KPIs com base na anotação semântica de modelos de dados dimensionais

Fonte: Elaborada pelos autores.

Figura 2. *Dasboard* gerado a partir dos RDFs inseridos no Virtuoso .



Fonte: Elaborada pelos autores.

Trechos dos códigos em RDF e SPARQL utilizados na geração do *dashboard*.

```
...
<#ID_KPI:1> <kpionto/hasdimension/mes> "JANEIRO" .
<#ID_KPI:1> <kpionto/hasdimension/ano>2000 .
<#ID_KPI:1> <kpionto/hasdimension/ResearchField>"CIENCIA DA INFORMAÇÃO" .
<#ID_KPI:1> <kpionto/hasdimension/fatorimpacto>"IMPACTO ENTRE 2 E 4"   .
<#ID_KPI:1> <kpionto/hasdimension/nivelfatorimpacto>4 .
<#ID_KPI:1> <kpionto/hasggrfunction/contagempublic> 6 .
...
```

```
SELECT DISTINCT
   kpi_pubic.s as id_kpi,
   kpi_pubic.p as labelDimensao,
   kpi_pubic.o as dsDimensao
FROM
(
SPARQL
SELECT ?s ?p ?o
FROM <http://kpi_public_pesq>
WHERE
   {  ?s ?p ?o
      FILTER (?o IN (2000, "JANEIRO", "FEVEREIRO", "MARÇO"))     }
) as kpi_pubic
```

Fonte: Elaborada pelos autores.

APLICAÇÃO DE MÉTRICAS PARA DESTAQUE DE ENTIDADES NA ANÁLISE DE GRAFOS

Application of Entity Highlighting Metrics in Graph Analysis

Roberto Zaina[1], Vinicius Faria Culmant Ramos[1], Gustavo Medeiros de Araujo[1]
(1) Universidade Federal de Santa Catarina, R. Eng. Agronômico Andrei Cristian Ferreira, s/n - Trindade, Florianópolis - SC, 88040-900, rzaina@gmail.com, v.ramos@ufsc.br, gustavo.araujo@ufsc.br.

Resumo:
A proposta do presente estudo é a de desenvolver um método para destaque de entidades em Relatórios de Inteligência Financeira a partir de métricas de relevância. Inicialmente, explicou-se que uma das informações usadas em investigações de lavagem de dinheiro é o Relatório de Inteligência Financeira. Este documento pode ser analisado somente pela sua leitura ou, dependendo do volume e da complexidade dos seus dados, o Relatório de Inteligência Financeira precisa ser analisado por meio de ferramenta de análise. Foi desenvolvido um método de processamento de dados por uma ferramenta de *business intelligence* e apresentados os resultados por uma ferramenta de análise de vínculos. Neste método, foram aplicadas duas métricas de relevância: "empresas suspeitas" e "contadores suspeitos". A partir do processamento dos dados, os principais resultados foram a detecção automática de empresas e contadores suspeitos e a posterior visualização em formato de grafos com destaques de entidades relevantes. Este método ajuda o analista, pois facilita o processamento de grande volume de dados e ajuda a diminuir a complexidade das informações de Relatórios de Inteligência Financeira.
Palavras-chave: Lavagem de Dinheiro; *Business Intelligence*; Análise de Vínculos.

Abstract:
The purpose of this study is to develop a method for highlighting entities in Financial Intelligence Reports from relevance metrics. It was initially explained that one of the information used in money laundering investigations is the Financial Intelligence Report. This document can be reviewed by reading it only or, depending on the volume and the complexity of data, the Financial Intelligence Report needs to be analyzed using the analysis tool. A data processing method was developed by a business intelligence tool and the results presented by a link analysis tool. In this method, two relevance metrics were applied: "suspicious companies" and "suspicious accountants". From the data processing, the main results were the automatic detection of suspicious companies and counters and the subsequent visualization in graph format with highlights of relevant entities. This method helps the analyst as it facilitates the processing of large data and helps to reduce the complexity of Financial Intelligence Reporting information.
Keywords: Money laundry; Business intelligence; Link Analysis.

1. Introdução

A lavagem de dinheiro é o processo pelo qual uma pessoa procura dar aparência de legalidade a bens que têm sua origem mediata ou imediata em atividades criminais (HERNÁNDEZ QUINTERO, 2017).

Neste crime, o dinheiro proveniente de atividades criminosas é introduzido nos circuitos financeiros legais, por meio de complexas operações que promovem a desvinculação da origem ilícita dos valores (OLIVEIRA, 2012).

Para a investigação de lavagem de dinheiro as informações básicas a serem analisadas são as de natureza financeira, como as transações bancárias, as declarações fiscais e as operações financeiras suspeitas.

As operações financeiras suspeitas constam em documentos chamados de Relatórios de Inteligência Financeira (RIF), produzidos pela Unidade de Inteligência Financeira (UIF) do Brasil.

Em alguns casos, a análise de RIF é feita somente pela leitura e interpretação das operações descritas no relatório, sem a utilização de programas específicos de análise de dados.

Porém, dependendo do volume e da complexidade de informações contidas em um RIF, essa mera leitura textual é precária, pois dificilmente o analista conseguirá memorizar todas as informações e, ainda, fazer todas as correlações entre as pessoas, empresas e operações financeiras.

Um primeiro método para análise de RIF com o uso de ferramentas analíticas foi descrito no artigo "Identificação de entidades destaque para a melhoria da Análise de Vínculos" do II WIDaT (2018).

2. Objetivos

O objetivo geral deste estudo é aprimorar o método de identificação de "entidades destaque" com a aplicação de tecnologias de *business intelligence* (*BI*) e de análise de vínculos em Relatórios de Inteligência Financeira.

3. Procedimentos Metodológicos

A metodologia usada foi a pesquisa experimental, pela aplicação de programas de *business intelligence* e de análise de vínculos em dados de Relatórios de Inteligência Financeira.

O *business intelligence* pode ser definido como um conjunto de metodologias, processos e tecnologias que transformam dados brutos em informações úteis e significativas (EDELHAUSER e IONICA, 2014). Neste trabalho foi usado o programa *Qlikview*[1] em sua versão *desktop*.

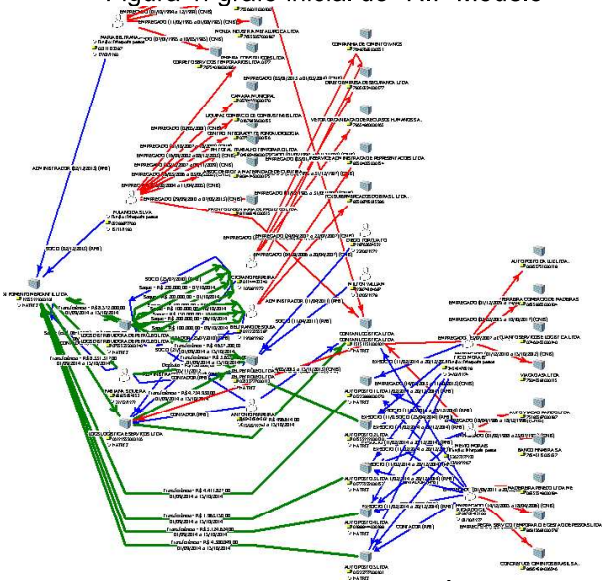
Por sua vez, as ferramentas de análise de vínculos criam grafos, que são diagramas que servem como um retrato gráfico de dados investigativos (COADY, 1985 apud SPARROW, 1991). Neste trabalho foi usado o *Analyst's Notebook*[2].

Para o desenvolvimento do método deste trabalho, partiu-se do modelo desenvolvido e apresentado no II WIDaT (2018), em que o principal resultado foi o desenvolvimento de um método por meio dos programas *Qlikview* e *Analyst's Notebook*.

Naquele método, pelo *Qlikview* foram processados dados tabulados de um "RIF Modelo" e, ainda, dados das empresas e de seus sócios, e o resultado foi a identificação de pessoas suspeitas de serem interpostas pessoas, também chamadas de "laranjas".

O resultado da aplicação deste método no "RIF Modelo" foi a identificação de 4 (quatro) prováveis interpostas pessoas. Isto foi carregado no *Analyst's Notebook*, que gerou um grafo inicial, mostrado na Figura 1.

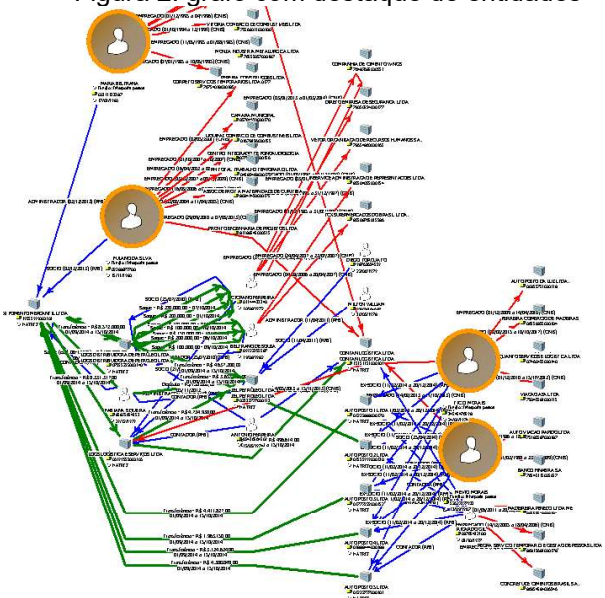
Figura 1: grafo inicial do "RIF Modelo"



Fonte: ZAINA, RAMOS e ARAÚJO (2018)

Posteriormente, com a aplicação de uma formatação no *Analyst's Notebook*, chegou-se a um grafo com destaque das interpostas pessoas, mostrado na Figura 2.

Figura 2: grafo com destaque de entidades



Fonte: ZAINA, RAMOS e ARAÚJO (2018)

[1] Ver mais em: https://www.qlik.com/pt-br. Acesso em 26/07/2019.

[2] Ver mais em: www.ibm.com/br-pt/marketplace/analysts-notebook. Acesso em 26/07/2019.

O grafo da Figura 6, ao destacar as entidades dos prováveis “laranjas”, melhora significativamente a visualização do grafo e amplia instantaneamente a compreensão por parte do investigador que analisa o RIF.

Neste artigo é apresentada uma evolução deste modelo de destaque de entidades, alterando alguns dados de entrada e buscando adotar novas métricas de relevância.

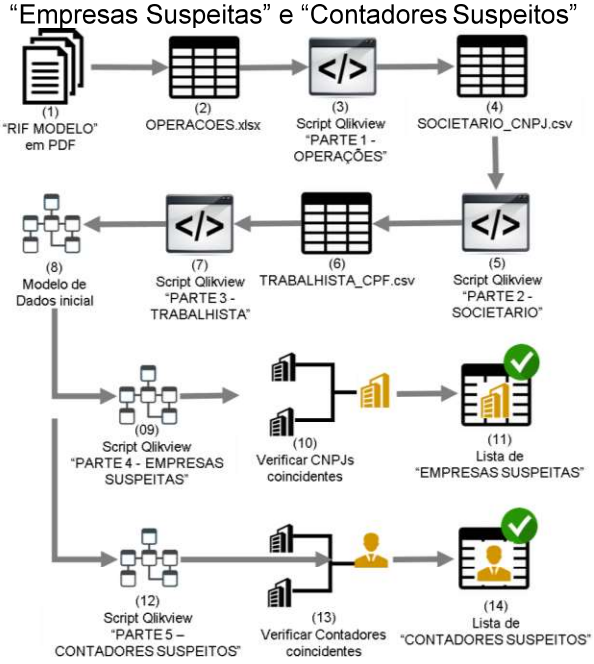
As novas métricas de relevância adotadas foram denominadas de “Empresas Suspeitas” e “Contadores Suspeitos”.

A primeira métrica de relevância “Empresas Suspeitas” é de empresas que possam ter empregados ou ex-empregados usados como sócios “laranjas” em outras empresas de um mesmo RIF.

A segunda métrica de relevância “Contadores Suspeitos” é de pessoas que atuam como contadoras, ao mesmo tempo, em mais de uma empresa relacionada em um mesmo RIF.

Os procedimentos para identificar essas métricas nos dados do “RIF Modelo” foram os mostrados na Figura 3.

Figura 3: Metodologia para identificação de “Empresas Suspeitas” e “Contadores Suspeitos”



Fonte: Elaborado pelos autores.

Na etapa 1, o “RIF Modelo”, disponibilizado em formato PDF, não estruturado, teve seus dados tabulados em uma planilha eletrônica.

O resultado da tabulação foi gravado como “OPERACOES.xlsx”, como ilustra a etapa 2.

Na etapa 3, foi criado um arquivo no *Qlikview* denominado “RIF.qvw”. Depois, os dados da planilha “OPERACOES.xlsx” foram carregados pelo *script* “PARTE 1 - OPERAÇÕES” do Apêndice A.

Na etapa 4, os dados societários das empresas foram consultados em um sistema interno e o resultado foi gravado em um arquivo “SOCIETARIO_CNPJ.csv”.

Depois, na etapa 5, o arquivo “SOCIETARIO_CNPJ.csv” foi carregado pelo *script* “PARTE 2 - SOCIETARIO” do Apêndice A.

Na etapa 6, foram pesquisados, em um sistema interno, os dados trabalhistas dos sócios das empresas e o resultado foi gravado em um arquivo “TRABALHISTA_CPF.csv”.

Após isto, conforme a etapa 7, o arquivo “TRABALHISTA_CPF.csv” foi carregado pelo *script* “PARTE 3 - TRABALHISTA” do Apêndice A.

A etapa 8 representa a primeira execução do *script*, conforme as partes 1, 2 e 3 do Apêndice A, para criar a primeira carga de dados.

Nesta primeira carga de dados, a principal preocupação foi o tratamento dos padrões de registros, principalmente os de CPF e de CNPJ.

Posteriormente, pela etapa 9, foram configurados comandos no *script* para identificar as “Empresas Suspeitas”, conforme o trecho “PARTE 4 - EMPRESAS SUSPEITAS” do Apêndice A.

Estes comandos serviram para consultar empresas que constavam tanto nos dados societários quanto nos dados trabalhistas.

Então, como representa a etapa 10, foram cruzados os dados para verificar registros de CNPJs coincidentes nas tabelas “SOCIETARIO” e “TRABALHISTA”.

Na etapa 11, foi criado um objeto para mostrar os CNPJs coincidentes.

Na etapa 12, foram aplicados comandos no script do arquivo “RIF.qvw” para identificar os “Contadores Suspeitos”, conforme o trecho “PARTE 5 - CONTADORES SUSPEITOS” do Apêndice A.

Com base neste cruzamento, na etapa 13, foi criada uma tabela para mostrar os CPFs de pessoas que constaram como contadores, ao mesmo tempo, em mais de uma empresa relacionada no "RIF Modelo".

Finalmente, na etapa 14, foi criada uma tabela para mostrar os CPFs dos "Contadores Suspeitos".

4. Resultados

Pela aplicação do método no "RIF Modelo", foram identificadas 2 (duas) empresas que continham, entre seus empregados ou ex-empregados, pessoas como sócios em outras empresas no mesmo RIF, conforme a Figura 4:

Figura 4: "Empresas Suspeitas" do "RIF Modelo"

CNPJ	NOME EMPRESA
06999555000106	LOGS LOGISTICA E SERVICOS
15151151000109	CONTAN LOGISTICA LTDA

Fonte: Elaborado pelos autores.

E, ainda, foram identificados 3 (três) CPFs de pessoas que constam como contadores em mais de uma empresa, ao mesmo tempo, em um mesmo RIF, como mostra a Figura 5:

Figura 5: "Contadores Suspeitos" do "RIF Modelo"

CPF CONTADOR	QTDE DE EMPRESAS
08765432100	5
85875654353	2
65436564509	2

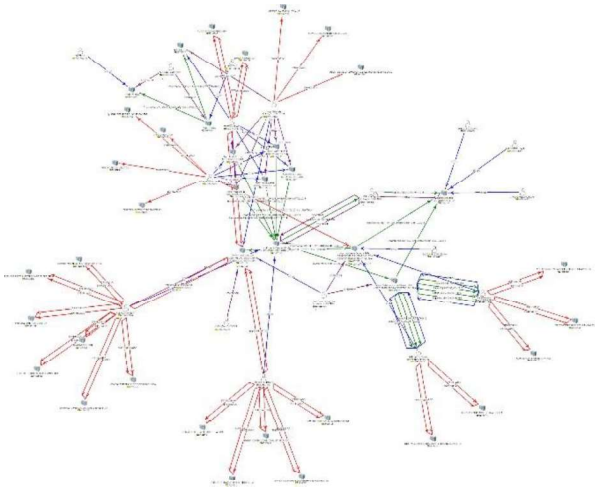
Fonte: Elaborado pelos autores.

Com base nas Figuras 4 e 5, verifica-se que o principal resultado do processamento do arquivo "RIF.qvw" foi a identificação de "empresas suspeitas" e "contadores suspeitos".

Este resultado pode ser usado no programa *Analyst's* para destacar as empresas e contadores suspeitos, gerando um grafo com destaque visual de entidades relevantes.

Na Figura 6 mostra-se o grafo do "RIF Modelo" no seu formato inicial, com as operações do RIF e os dados societários e trabalhistas, por enquanto sem quaisquer ações de destaque de entidades relevantes:

Figura 6: Grafo inicial do "RIF Modelo"

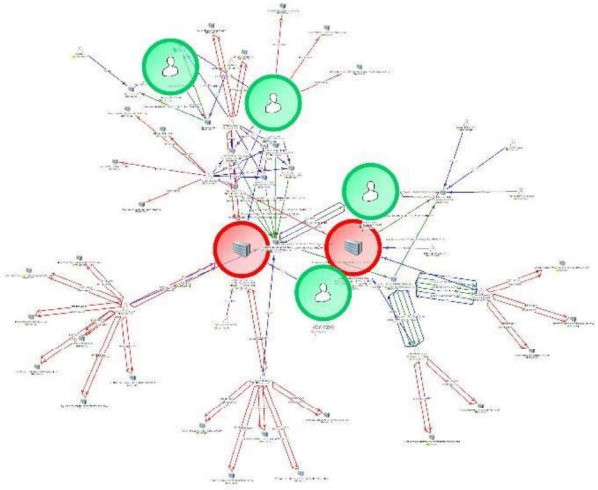


Fonte: Elaborado pelos autores.

Somente pela análise visual da Figura 6 não se consegue perceber rapidamente quais as empresas e contadores suspeitos. Contudo, é possível usar o recurso de formatação condicional do *Analyst's* para destacar tais ocorrências no grafo, para ampliar e destacar com cores as entidades com atributos de "suspeitos".

O resultado da aplicação desta formatação condicional no grafo inicial do "RIF Modelo" é um novo grafo com destaque de entidades, mostrado na Figura 7:

Figura 7: Grafo destacando as empresas e contadores suspeitos



Fonte: Elaborado pelos autores.

Ao compararmos as Figuras 6 e 7 fica evidente que o destaque das entidades melhora a visualização do grafo, automatiza a

detecção de empresas e contadores suspeitos e, assim, facilita a análise do RIF.

Este novo modelo mostrou-se adequado para ajudar a analisar o grande volume de dados e a complexidade das informações contidas no "RIF Modelo".

Diante disto, decidiu-se aplicar esse modelo em outros 20 (vinte) RIF's, que tinham sido analisados e tabulados em planilhas eletrônicas. Tais relatórios receberam uma numeração de 01 a 20.

As planilhas de todos os RIF's foram carregadas e processadas no arquivo "RIF.qvw". Posteriormente, os dados societários e trabalhistas também foram coletados e carregados no mesmo arquivo.

Como resultado do processamento, chegou-se a seguinte relação de existência de "Empresas Suspeitas" e/ou "Contadores Suspeitos", mostrada no Quadro 1:

Quadro 1: RIF's com entidades suspeitas

Nº RIF	Empresas suspeitas?	Contadores suspeitos?
01	**Sim**	**Sim**
02	**Sim**	**Sim**
03	Não	Não
04	Não	Não
05	**Sim**	Não
06	Não	Não
07	**Sim**	Não
08	**Sim**	Não
09	Não	Não
10	**Sim**	Não
11	Não	Não
12	Não	Não
13	**Sim**	Não
14	Não	Não
15	**Sim**	Não
16	**Sim**	**Sim**
17	Não	Não
18	**Sim**	**Sim**
19	Não	**Sim**
20	**Sim**	Não

Fonte: Elaborado pelos autores.

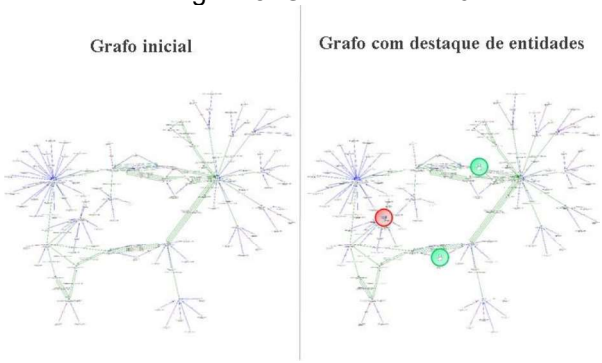
Pelo Quadro 1 constata-se que 12 (doze) dos 20 (vinte) RIF's possuem, pelo menos, uma empresa suspeita ou um contador suspeito.

Os dados desses RIF's foram importados no programa *Analyst's*, gerando um primeiro diagrama "inicial".

Em seguida, pela aplicação da formatação condicional "Entidades Suspeitas" foi gerado um segundo diagrama "com destaque de entidades" para cada RIF, destacando em vermelho as "empresas suspeitas" e em verde os "contadores suspeitos".

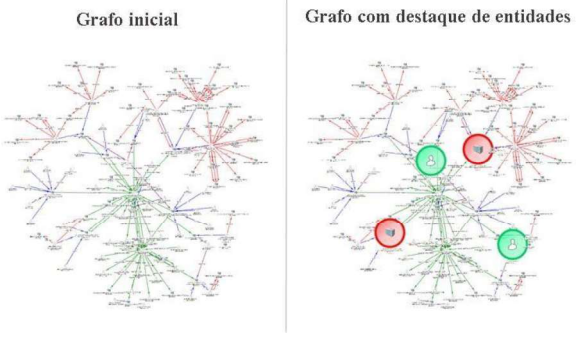
Como amostra dos resultados, nas Figuras 8 a 11 seguem ilustrações comparativas entre grafos de 4 (quatro) RIF's (números 01, 02, 16 e 18), que tinham tanto empresas quanto contadores suspeitos:

Figura 8: Grafos do RIF 01



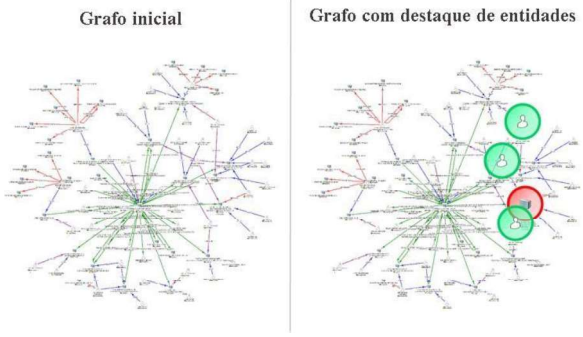
Fonte: Elaborado pelos autores.

Figura 9: Grafos do RIF 02



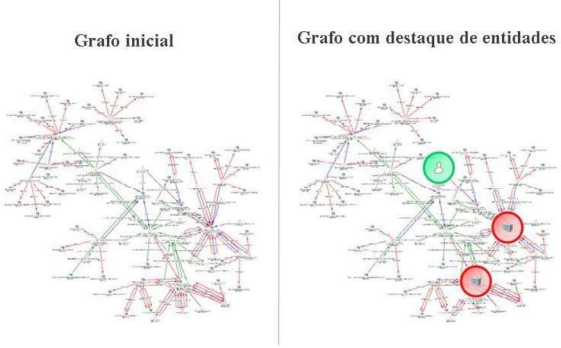
Fonte: Elaborado pelos autores.

Figura 10: Grafos do RIF 16



Fonte: Elaborado pelos autores.

Figura 11: Grafos do RIF 18



Fonte: Elaborado pelos autores.

As Figuras 8 a 11 deixam evidente como o destaque de entidades relevantes facilita o trabalho de análise do investigador que, em um primeiro momento, pode se dedicar a explorar as características das entidades destacadas.

5. Conclusão

Com o objetivo de ajudar na análise de Relatórios de Inteligência Financeira, decidiu-se verificar a possibilidade de usar programas para identificar automaticamente elementos relevantes e destacá-los em grafos.

Um primeiro método foi desenvolvido e apresentado no artigo "Identificação de entidades destaque para a melhoria da Análise de Vínculos" no II WIDaT (2018).

No presente trabalho foram aplicadas novas métricas ao referido método: "Empresas Suspeitas" e "Contadores Suspeitos".

Em seguida, foram processados os dados nos programas *Qlikview* e *Analyst's Notebook*, e os principais resultados foram a detecção automática de elementos suspeitos e a posterior visualização em formato de grafos com destaques de entidades relevantes.

O método desenvolvido mostra que a configuração de métricas em determinadas tecnologias auxilia no processamento do grande volume de dados e ajuda a diminuir a complexidade dos Relatórios de Inteligência Financeira.

A partir deste trabalho, outras métricas de relevância podem ser idealizadas e outras tecnologias podem ser testadas como, por exemplo, de mineração de dados e de inteligência artificial, para demonstrar sua utilidade em investigações de lavagem de dinheiro.

Referências

EDELHAUSER, E.; IONICA, A. A Business Intelligence Software Made in Romania, A Solution for Romanian Companies During the Economic Crisis. COMPUTER SCIENCE AND INFORMATION SYSTEMS, v. 11, n. 2, p. 809–823, jun. 2014.

HERNÁNDEZ QUINTERO, H. A. Aspectos polémicos sobre el objeto material del delito de lavado de activos (delitos fuente). Justicia, v. 22, n. 32, p. 118–138, 2017.

OLIVEIRA, J. C. DE; AGAPITO, L. S.; MIRANDA, M. D. A. E. O modelo de "autorregulação regulada" e a teoria da captura: obstáculos à efetividade no combate à lavagem de dinheiro no Brasil. Revista Quaestio Iuris, v. 10, n. 1, p. 365–388, 2017.

SPARROW, Malcolm K. The application of network analysis to criminal intelligence: An assessment of the prospects. Social networks, v. 13, n. 3, p. 251-274, 1991.

ZAINA, R. ; RAMOS, V. F. C. ; ARAÚJO, G. M. Identificação de entidades destaque para a melhoria da Análise de Vínculos. In: II Workshop de Informação, Dados e Tecnologia (WIDaT), 2018, João Pessoa. Anais do II Workshop de informação, dados e tecnologia (WIDaT). João Pessoa: EDITORA DA UFPB, 2018. v. 1. p. 157-174.

Apêndice A

```
//SCRIPT PARA CARGA DE DADOS DE OPERAÇÕES DE RIF, DADOS SOCIETÁRIOS E DADOS TRABALHISTAS

//CONFIGURAÇÕES GERAIS

SET ThousandSep='.';
SET DecimalSep=',';
SET MoneyThousandSep='.';
SET MoneyDecimalSep=',';
SET MoneyFormat='R$ #.##0,00;-R$ #.##0,00';
SET TimeFormat='hh:mm:ss';
SET DateFormat='DD/MM/YYYY';
SET TimestampFormat='DD/MM/YYYY hh:mm:ss[.fff]';
SET FirstWeekDay=6;
SET BrokenWeeks=1;
SET ReferenceDay=0;
SET FirstMonthOfYear=1;
SET CollationLocale='pt-BR';
SET MonthNames='jan;fev;mar;abr;mai;jun;jul;ago;set;out;nov;dez';
SET LongMonthNames='janeiro;fevereiro;março;abril;maio;junho;julho;agosto;setembro;outubro;novembro;dezembro';
SET DayNames='seg;ter;qua;qui;sex;sáb;dom';
SET LongDayNames='segunda-feira;terça-feira;quarta-feira;quinta-feira;sexta-feira;sábado;domingo';

//PARTE 1 - OPERAÇÕES

//PARTE 1.1 - ETL DOS DADOS DAS PLANILHAS "OPERACOES.xlsx"

ETL_1_OPERACOES:
LOAD
RIF AS ETL_1_OP_NR_RIF,
[ITEM RIF] AS ETL_1_OP_ITEM_RIF,
TRIM(KeepChar([ORIGEM ou DEPOSITANTE CPF/CNPJ],'0123456789')) AS ETL_1_OP_ORIGEM_CPF_CNPJ,
[ORIGEM ou DEPOSITANTE NOME] AS ETL_1_OP_ORIGEM_NOME,
TRIM(KeepChar([DESTINO ou SACADOR CPF/CNPJ],'0123456789')) AS ETL_1_OP_DESTINO_CPF_CNPJ,
[DESTINO ou SACADOR NOME] AS ETL_1_OP_DESTINO_NOME,
[TIPO OPERAÇÃO Selecione] AS ETL_1_OP_TIPO_OPERACAO,
Money#([VALOR (EM R$) somente números],'R$ #.##0,00;-R$ #.##0,00') AS ETL_1_OP_VALOR_OPERACAO,
Date#([DATA/PERIODO],'DD/MM/YYYY') AS ETL_1_OP_DATA_OPERAÇÃO
FROM
OPERACOES\OPERACOES*.xlsx
(ooxml, embedded labels, table is Operações);
LEFT JOIN
LOAD NR_REFERENCIA AS ETL_1_OP_NR_REFERENCIA,
 RIF as ETL_1_OP_NR_RIF
FROM
OPERACOES\Indice_RIFs.xlsx
(ooxml, embedded labels, table is Plan1);

ETL_2_OPERACOES:
LOAD
ETL_1_OP_NR_REFERENCIA AS ETL_2_OP_NR_REFERENCIA,
ETL_1_OP_NR_RIF AS ETL_2_OP_NR_RIF,
ETL_1_OP_ITEM_RIF AS ETL_2_OP_ITEM_RIF,
ETL_1_OP_ORIGEM_CPF_CNPJ AS ETL_2_OP_ORIGEM_CPF_CNPJ,
ETL_1_OP_ORIGEM_NOME AS ETL_2_OP_ORIGEM_NOME,
ETL_1_OP_DESTINO_CPF_CNPJ AS ETL_2_OP_DESTINO_CPF_CNPJ,
ETL_1_OP_DESTINO_NOME AS ETL_2_OP_DESTINO_NOME,
ETL_1_OP_TIPO_OPERACAO AS ETL_2_OP_TIPO_OPERACAO,
ETL_1_OP_VALOR_OPERACAO AS ETL_2_OP_VALOR_OPERACAO,
ETL_1_OP_DATA_OPERAÇÃO AS ETL_2_OP_DATA_OPERAÇÃO,
Left(
Right(
'00000000000000' &
```

```
(IF(LEN(ETL_1_OP_ORIGEM_CPF_CNPJ)>11,ETL_1_OP_ORIGEM_CPF_CNPJ)),14),8) AS
ETL_2_OP_CNPJ_ORIGEM_ID,
Left(
Right(
'00000000000000' &
(IF(LEN(ETL_1_OP_DESTINO_CPF_CNPJ)>11,ETL_1_OP_DESTINO_CPF_CNPJ)),14),8) AS
ETL_2_OP_CNPJ_DESTINO_ID
Resident ETL_1_OPERACOES;

//PARTE 1.2 - CARGA DOS DADOS TRATADOS DAS OPERAÇÕES DE RIFS

OPERACOES:
LOAD
ETL_2_OP_NR_REFERENCIA AS OP_NR_REFERENCIA,
ETL_2_OP_NR_RIF AS OP_NR_RIF,
ETL_2_OP_ITEM_RIF AS OP_ITEM_RIF,
ETL_2_OP_ORIGEM_CPF_CNPJ AS OP_ORIGEM_CPF_CNPJ,
ETL_2_OP_ORIGEM_NOME AS OP_ORIGEM_NOME,
ETL_2_OP_DESTINO_CPF_CNPJ AS OP_DESTINO_CPF_CNPJ,
ETL_2_OP_DESTINO_NOME AS OP_DESTINO_NOME,
ETL_2_OP_TIPO_OPERACAO AS OP_TIPO_OPERACAO,
ETL_2_OP_VALOR_OPERACAO AS OP_VALOR_OPERACAO,
ETL_2_OP_DATA_OPERAÇÃO AS OP_DATA_OPERAÇÃO,
ETL_2_OP_CNPJ_ORIGEM_ID AS CNPJ_ORIGEM_ID,
ETL_2_OP_CNPJ_DESTINO_ID AS CNPJ_DESTINO_ID,
ETL_2_OP_NR_RIF & '-' & ETL_2_OP_ITEM_RIF & '-' & ETL_2_OP_CNPJ_ORIGEM_ID & '-' &
ETL_2_OP_CNPJ_DESTINO_ID AS ID_OPERAÇAO_CNPJ
Resident ETL_2_OPERACOES;

//PARTE 2 - SOCIETARIO

//PARTE 2.1 - ETL DOS DADOS DAS PLANILHAS "SOCIETARIO_CNPJ.csv"

ETL_1_SOCIETARIO:
LOAD
RIGHT('00000000000000' & TRIM(KeepChar(CNPJ,'0123456789')),14) AS ETL_1_SOC_CNPJ,
[NOME DA EMPRESA] AS ETL_1_SOC_NOME_EMPRESA,
RIGHT('00000000000' & TRIM(KeepChar([CPF CONTADOR],'0123456789')),11) AS ETL_1_SOC_CONTADOR,
RIGHT('00000000000' & TRIM(KeepChar([SÓCIO CPF/CNPJ],'0123456789')),11) AS
ETL_1_SOC_CPF_CNPJ_SOCIO,
[NOME DO SÓCIO] AS ETL_1_SOC_NOME_SOCIO
FROM
SOCIETARIO_CNPJ\SOCIETARIO_CNPJ.csv
(txt, codepage is 28591, embedded labels, delimiter is ';', msq, filters(
Top(1, 'CNPJ_2'), Remove(Row, RowCnd(CellValue, 1, StrCnd(equal, 'CNPJ'))),
Remove(Row, RowCnd(CellValue, 1, StrCnd(longer, 2))), Top(1, 'CNPJ')));

ETL_2_SOCIETARIO:
LOAD
ETL_1_SOC_CNPJ as ETL_2_SOC_CNPJ,
ETL_1_SOC_NOME_EMPRESA as ETL_2_SOC_NOME_EMPRESA,
REPLACE(ETL_1_SOC_CONTADOR,'00000000000','') as ETL_2_SOC_CONTADOR,
REPLACE(ETL_1_SOC_CPF_CNPJ_SOCIO,'00000000000','') as ETL_2_SOC_CPF_CNPJ_SOCIO,
ETL_1_SOC_NOME_SOCIO as ETL_2_SOC_NOME_SOCIO,
Left(
Right(
'00000000000000' &
ETL_1_SOC_CNPJ,14),8) AS ETL_2_SOC_CNPJ_ID,
Left(
Right(
'00000000000000' &
ETL_1_SOC_CNPJ,14),8) AS CNPJ
Resident ETL_1_SOCIETARIO;

//PARTE 2.2 - CARGA DOS DADOS SOCIETARIOS TRATADOS
```

```
SOCIETARIO_ORIGEM:
LOAD
ETL_2_SOC_CNPJ AS SOC_ORIGEM_CNPJ,
ETL_2_SOC_NOME_EMPRESA AS SOC_ORIGEM_NOME_EMPRESA,
ETL_2_SOC_CONTADOR AS SOC_ORIGEM_CONTADOR,
ETL_2_SOC_CPF_CNPJ_SOCIO AS SOC_ORIGEM_CPF_CNPJ_SOCIO,
ETL_2_SOC_NOME_SOCIO AS SOC_ORIGEM_NOME_SOCIO,
ETL_2_SOC_CNPJ_ID AS CNPJ_ORIGEM_ID,
ETL_2_SOC_CPF_CNPJ_SOCIO AS CPF_ORIGEM_ID
Resident ETL_2_SOCIETARIO;

SOCIETARIO_DESTINO:
LOAD
ETL_2_SOC_CNPJ AS SOC_DESTINO_CNPJ,
ETL_2_SOC_NOME_EMPRESA AS SOC_DESTINO_NOME_EMPRESA,
ETL_2_SOC_CONTADOR AS SOC_DESTINO_CONTADOR,
ETL_2_SOC_CPF_CNPJ_SOCIO AS SOC_DESTINO_CPF_CNPJ_SOCIO,
ETL_2_SOC_NOME_SOCIO AS SOC_DESTINO_NOME_SOCIO,
ETL_2_SOC_CNPJ_ID AS CNPJ_DESTINO_ID,
ETL_2_SOC_CPF_CNPJ_SOCIO AS CPF_DESTINO_ID
Resident ETL_2_SOCIETARIO;

//PARTE 3 - TRABALHISTA

//PARTE 3.1 - ETL DOS DADOS DAS PLANILHAS "TRABALHISTA_CPF.csv"

ETL_1_TRABALHISTA:
LOAD
TRIM(KeepChar(CPF,'0123456789')) AS ETL_1_TRAB_CPF,
Nome AS ETL_1_TRAB_NOME,
[Razao Social] AS ETL_1_TRAB_RAZAO_SOCIAL,
TRIM(KeepChar(CNPJ,'0123456789')) AS ETL_1_TRAB_CNPJ
FROM
TRABALHISTA_CPF\TRABALHISTA_CPF.csv
(txt, utf8, embedded labels, delimiter is ';', msq, filters(
Top(13, 'OK'),
Remove(Row, RowCnd(CellValue, 13, StrCnd(equal, 'OK', not))),
Remove(Row, RowCnd(CellValue, 8, StrCnd(null))),
Remove(Row, RowCnd(CellValue, 8, StrCnd(longer, 2))),
Top(13, 'Resposta')
));

ETL_2_TRABALHISTA:
LOAD
ETL_1_TRAB_CPF AS ETL_2_TRAB_CPF,
ETL_1_TRAB_NOME AS ETL_2_TRAB_NOME,
ETL_1_TRAB_RAZAO_SOCIAL AS ETL_2_TRAB_RAZAO_SOCIAL,
ETL_1_TRAB_CNPJ AS ETL_2_TRAB_CNPJ,
Left(
Right(
'00000000000000' &
ETL_1_TRAB_CNPJ,14),8) AS ETL_2_TRAB_RAIZ_CNPJ
Resident ETL_1_TRABALHISTA;

//PARTE 3.2: CARGA DOS DADOS TRABALHISTAS TRATADOS

TRABALHISTA_ORIGEM:
LOAD
ETL_2_TRAB_CPF AS TRAB_CPF_ORIGEM,
ETL_2_TRAB_NOME AS TRAB_NOME_ORIGEM,
ETL_2_TRAB_RAZAO_SOCIAL AS TRAB_RAZAO_SOCIAL_ORIGEM,
ETL_2_TRAB_CNPJ AS TRAB_CNPJ_ORIGEM,
ETL_2_TRAB_RAIZ_CNPJ AS TRAB_RAIZ_CNPJ_ORIGEM,
ETL_2_TRAB_CPF AS CPF_ORIGEM_ID
```

```
Resident ETL_2_TRABALHISTA;

TRABALHISTA_DESTINO:
LOAD
ETL_2_TRAB_CPF AS TRAB_CPF_DESTINO,
ETL_2_TRAB_NOME AS TRAB_NOME_DESTINO,
ETL_2_TRAB_RAZAO_SOCIAL AS TRAB_RAZAO_SOCIAL_DESTINO,
ETL_2_TRAB_CNPJ AS TRAB_CNPJ_DESTINO,
ETL_2_TRAB_RAIZ_CNPJ AS TRAB_RAIZ_CNPJ_DESTINO,
ETL_2_TRAB_CPF AS CPF_DESTINO_ID
Resident ETL_2_TRABALHISTA;

//PARTE 4 - EMPRESAS SUSPEITAS

COINCIDENCIA_CNPJ_ORIGEM:
LOAD
CNPJ_ORIGEM_ID,
CNPJ_ORIGEM_ID as CNPJ_COINCIDENTE_ORIGEM
Resident SOCIETARIO_ORIGEM;
Inner Join
LOAD
TRAB_RAIZ_CNPJ_ORIGEM as CNPJ_COINCIDENTE_ORIGEM
Resident TRABALHISTA_ORIGEM;

COINCIDENCIA_CNPJ_DESTINO:
LOAD
CNPJ_DESTINO_ID,
CNPJ_DESTINO_ID as CNPJ_COINCIDENTE_DESTINO
Resident SOCIETARIO_DESTINO;
Inner Join
LOAD
TRAB_RAIZ_CNPJ_DESTINO as CNPJ_COINCIDENTE_DESTINO
Resident TRABALHISTA_DESTINO;

//PARTE 5 - CONTADORES SUSPEITOS

CONTADOR:
LOAD
ID_OPERAÇAO_CNPJ,
CNPJ_ORIGEM_ID as CNPJ_ID
Resident OPERACOES;
LOAD
ID_OPERAÇAO_CNPJ,
CNPJ_DESTINO_ID as CNPJ_ID
Resident OPERACOES;
CONTADOR_EMPRESA:
LOAD
ETL_2_SOC_CNPJ_ID AS CNPJ_ID,
ETL_2_SOC_CONTADOR AS CONTADOR
Resident ETL_2_SOCIETARIO;
```

AUTORIDADE NACIONAL DE PROTEÇÃO DE DADOS E PRIVACIDADE

NATIONAL DATA PROTECTION AUTHORITY AND PRIVACY

Rosilene Paiva Marinho de Sousa[1],
Paulo Henrique Tavares da Silva[2],
Marckson Roberto Ferreira de Sousa[3]

(1) Centro Universitário de João Pessoa (UNIPÊ), adv.rpmarinho@gmail.com
(2) Centro Universitário de João Pessoa (UNIPÊ), paulo.tavares@unipe.edu.br
(3) Universidade Federal da Paraíba (UFPB), marckson.dci.ufpb@gmail.com

Resumo:
Objetiva analisar alterações realizadas na Lei Geral de Proteção de Dados para dispor sobre a proteção de dados pessoais e para criar a Autoridade Nacional de Proteção de Dados. Busca investigar aspectos da criação da referida Autoridade Nacional de Proteção de Dados, na orientação de empresas e órgãos governamentais sobre as situações que envolvem o controle e circulação de dados e informações pessoais. Examinam-se como foram criados seus órgãos, sua respectiva composição e principais características. Quanto à metodologia, trata-se de um estudo bibliográfico e exploratório, com característica descritiva. São ainda abordados elementos normativos que visam a reflexão sobre competências necessárias para garantia da privacidade. A partir da análise, considera-se a necessidade da elaboração de diretrizes para a Política Nacional de Proteção de Dados Pessoais e Privacidade. Verifica-se que a Lei que cria autoridade nacional realiza alterações no âmbito da Lei Geral de Proteção de Dados, buscando adequar aspectos como a inclusão da definição da referida autoridade, e inclusão de competências como a edição de regulamentos e procedimentos, relatórios de impacto à proteção de dados pessoais para os casos em que o tratamento representar alto risco à garantia aos princípios gerais, contribuindo para que essas diretrizes tenham eficácia.

Palavras-chave: Lei Geral de Proteção de Dados; Autoridade Nacional de Proteção de Dados; Política Nacional de Proteção de Dados e Privacidade.

Abstract:
Aims to analyse changes made in the General Data Protection Act to dispose of personal data protection and to create the National Data Protection Authority. It seeks to investigate aspects of the creation of the aforementioned National Data Protection Authority, in the guidance of companies and governmental agencies about the situations involving the control and circulation of data and personal information. They examine how their organs were created, their respective composition and main characteristics. As for the methodology, this is a bibliographic and exploratory study, with a descriptive characteristic. Normative elements are also addressed that aim to reflect on the competencies required to guarantee privacy. Based on the analysis, we consider the need to develop guidelines for the National Policy on Protection of Personal Data and Privacy. It is verified that the Law establishing national authority makes changes in the scope of the General Data Protection Act, seeking to adapt aspects such as the inclusion of the definition of that authority, and inclusion of competencies such as the editing of regulations and procedures, impact reports on the protection of personal data for cases where the treatment poses high risk to the general principles, contributing to these guidelines to be effective.

Keywords: General Data Protection Law; National Data Protection Authority; National Policy on Data Protection and Privacy.

1. Introdução

A necessidade de controle sobre a circulação de dados e informações pessoais tem sido cada vez mais evidenciada, tendo em vista o volume crescente produzido a partir da expansão das tecnologias de informação e comunicação (TIC). Nos últimos anos vários países têm se preocupado com a garantia de direitos de privacidade constitucionalmente assegurados no âmbito dos mesmos, procurando criar marcos regulatórios, que visem à proteção dos referidos dados e informações pessoais.

Com o Brasil não foi diferente, criando com base no Regulamento Geral de Proteção de dados da União Europeia, a Lei

Geral de Proteção de Dados (LGPD), que surgiu com o objetivo de garantir a liberdade, privacidade e o livre desenvolvimento da personalidade da pessoa natural, a partir do controle sobre circulação de dados e informações pessoais.

Para que essas garantias sejam asseguradas, a Lei nº 13.853/2019, que altera a LGPD, surge criando a Autoridade Nacional de Proteção de Dados (ANPD), bem como seus respectivos órgãos, para que se possa elaborar uma Política Nacional de Proteção de Dados Pessoais e da Privacidade, definindo, assim, suas principais competências.

Por Autoridade Nacional, compreende-se "[...] órgão da administração pública responsável por zelar, implementar e fiscalizar o cumprimento desta Lei em todo o território nacional" (BRASIL, 2019, *online*).

A proteção à privacidade está intimamente ligada ao controle sobre dados pessoais, compreendido estes, por "[...] informação relacionada a pessoa natural identificada ou identificável", conforme previsão do art. 5º, inciso I da LGPD (BRASIL, 2018, *online*).

O controle sobre dados pessoais visa garantir segurança para circulação destes, que conforme o a artigo 6º, inciso VII da LGPD, constitui:

> [...] utilização de medidas técnicas e administrativas aptas a proteger os dados pessoais de acessos não autorizados e de situações acidentais ou ilícitas de destruição, perda, alteração, comunicação ou difusão (BRASIL, 2018, *online*).

A segurança está relacionada às técnicas utilizadas para proteção dos dados de acesso ou disseminação não autorizados.

Nesse sentido, a relevância deste trabalho surge considerando a necessidade de uma padronização regulatória na orientação de empresas e órgãos governamentais sobre as situações que envolvem o controle e circulação de dados e informações pessoais.

2. Objetivos

O escopo principal desse trabalho consiste em analisar alterações realizadas na LGPD para dispor sobre a proteção de dados pessoais e para criar a ANPD. Como objetivos específicos, busca-se investigar aspectos da criação da ANPD, responsável pela orientação de empresas e órgãos governamentais na proteção de dados e informações pessoais; examinar como foram criados e estruturados seus órgãos; e refletir sobre a respectiva competência para garantia da privacidade.

3. Procedimentos Metodológicos

Quanto ao aspecto metodológico, trata-se de uma pesquisa bibliográfica e exploratória, com característica descritiva. A pesquisa exploratória tem como finalidade examinar o conhecimento sobre o tema pesquisado, sobre o qual se pretende estudar em outras perspectivas (SAMPIERI; COLLADO; LUCIO, 2013). Já a pesquisa bibliográfica, segundo Gil (2008), evidencia-se pela necessidade de se verificar material já elaborado, constituído, sobretudo de livros, artigos científicos, leis, dentre outros. Foram analisadas as legislações mais recentes envolvendo a LGPD de 2018 e a criação da ANPD de 2019, além de normas correlatas, bem como artigos sobre o tema. Em relação à pesquisa descritiva, Gil (2008) destaca que estas possuem como objetivo primordial a descrição das características de determinada população ou fenômeno.

4. Resultados

Nos resultados são apresentadas discussões sobre as alterações na LGPD, para a criação da ANPD, advindas com a Lei nº 13.853/2019. Examinam-se como se deu a criação de seus órgãos, como foram estruturados e suas principais características, bem como sua competência na orientação de empresas e órgãos governamentais sobre as situações que envolvem o controle e circulação de dados e informações pessoais.

4.1. A Lei Geral de Proteção de Dados e suas Principais Alterações

A Lei nº 13.709, de 14 de agosto de 2018, criada para dispor sobre a proteção de dados pessoais, surgiu com força principiológica constitucional na garantia de direitos fundamentais, como o direito à privacidade, a liberdade e o livre

desenvolvimento da personalidade da pessoa natural.

Visando o controle da circulação sobre dados e informações pessoais, a referida lei foi criada no contexto do Regulamento Geral de Proteção de dados da União Europeia, como uma necessidade de se criar um marco regulatório cujo escopo funda-se em fortalecer o papel fiscalizatório dos órgãos de controle. Busca-se entregar às pessoas naturais o poder efetivo sobre seus próprios dados, detalhando os conceitos de transparência e de consentimento explícito.

Com o intuito de apresentar os principais aspectos da LGPD, o Serviço Federal de Processamento de Dados (SERPRO), sintetizou os principais pontos da lei, destacando que serão afetados diferentes setores e serviços e a toda a sociedade, seja no âmbito individual, empresarial ou governamental, conforme exposto na Figura 1:

Figura 1 – Principais aspectos da LGPD

Fonte: SERPRO (2019)

Com a aprovação da Lei nº 13.853/2019 (BRASIL, 2019), que altera a Lei nº 13.709, de 14 de agosto de 2018, para dispor sobre a proteção de dados pessoais e para criar a ANPD, foi possível perceber um relevante número de modificações, entre as quais se podem considerar algumas mais importantes, conforme apresentado no Quadro 1:

Quadro 1 – Principais alterações na LGPD pela Lei nº 13.853/2019

A Lei destaca no parágrafo único do artigo 1º, que as normas gerais são de interesse nacional e devem ser observadas pela União, Estados, Distrito Federal e Municípios.
O artigo 5º, inciso VIII, altera o conceito de encarregado permitindo que o encarregado seja também pessoa jurídica e não apenas pessoa natural.
O inciso XIX do artigo 5º, também altera a definição de autoridade nacional para classifica-la como órgão da administração pública direta, responsável por zelar, implementar e fiscalizar o cumprimento da Lei em todo o território nacional.
Inclusão de competências para a ANPD, por meio do artigo 55-J, incluído no capítulo IX da LGPD.

Fonte: Dados da Pesquisa (2019)

As competências previstas no artigo 55-J serão analisadas na sequência.

4.2. Competências da Autoridade Nacional de Proteção de Dados para Garantia da Privacidade

A implementação da LGPD, para que haja proteção de dados e informações pessoais, constitui uma ampliação da proteção à privacidade e exige normas que possam auxiliar tecnicamente essa proteção para que esta possa ter eficiência. Nesse sentido, torna-se necessário verificar a forma como deve ocorrer o processamento dos dados e informações.

Segundo Alves (2019, *online*), em entrevista ao SERPRO, a autodeterminação informativa como um direito do cidadão fez da Alemanha o berço da proteção de dados, e assim:

> [...] Essa vivência já gerou uma certa harmonia, na sociedade europeia, de que a autoridade nacional de proteção de dados deve atuar para punir excessos e criar realmente um ambiente de conformação legal, mas ela não pode impedir o desenvolvimento econômico. Isso é algo que também foi consagrado na LGPD, na medida em que a inovação, o desenvolvimento econômico e a livre iniciativa estão em seus fundamentos e devem nortear o intérprete, o doutrinador, o juiz e o regulador. Ou seja, não basta a autoridade de dados vir com um viés punitivo. Ela tem que construir antes de punir. Essa é a proposta e a

> forma mais adequada de encarar a questão.

Nesse contexto, a Lei nº 13.853/2019, que altera a LGPD e cria a ANPD (BRASIL, 2019), surge como norma técnica que permite por em prática o controle sobre a circulação de dados e informações pessoais.

Pela referida lei, criou-se, sem aumento de despesas, a ANPD, como órgão da administração pública federal, integrante da presidência da república. A ANPD tem natureza jurídica transitória podendo ser, de acordo com o artigo 55-A, §1º e §2º, transformada pelo Poder Executivo em entidade da administração pública federal indireta, submetida a regime autárquico especial e vinculada à Presidência da República, devendo ocorrer em até dois anos da entrada em vigor da estrutura regimental da ANPD.

Segundo exposto no artigo 55-C, a ANPD fica composta pelo conselho diretor, órgão máximo de direção, composto por cinco membros, incluído o Diretor-Presidente; o Conselho Nacional de Proteção de Dados Pessoais e da Privacidade; Corregedoria; Ouvidoria; órgão de assessoramento jurídico próprio; e, unidades administrativas e unidades especializadas. Há de se salientar que esta estrutura poderá atuar a partir de suas competências na proteção à privacidade, inclusive atribuindo sanções nos casos de violação, que hodiernamente ainda não estão interligadas com o Código Penal Brasileiro (BRASIL, 1940).

Entre as principais competências previstas no artigo 55-J, da referida lei, estão, zelar pela proteção dos dados pessoais, nos termos da legislação, e pela observância dos segredos comercial e industrial, observada a proteção de dados pessoais e do sigilo das informações quando protegido por lei ou quando a quebra do sigilo violar os fundamentos previstos no art. 2º da LGPD; elaborar diretrizes para a Política Nacional de Proteção de Dados Pessoais e da Privacidade; fiscalizar e aplicar sanções em caso de tratamento de dados realizado em descumprimento à legislação, mediante processo administrativo que assegure o contraditório, a ampla defesa e o direito de recurso; apreciar petições de titular contra controlador após comprovada pelo titular a apresentação de reclamação ao controlador não solucionada no prazo estabelecido em regulamentação; promover na população o conhecimento das normas e das políticas públicas sobre proteção de dados pessoais e das medidas de segurança; promover e elaborar estudos sobre as práticas nacionais e internacionais de proteção de dados pessoais e privacidade; promover ações de cooperação com autoridades de proteção de dados pessoais de outros países, de natureza internacional ou transnacional. Essas competências visam o controle sobre dados pessoais visando garantir proteção à privacidade e segurança para circulação destes de forma correta e transparente.

Observa-se que para que haja uma política nacional de proteção de dados e privacidade, torna-se necessário cumprir outras competências de cunho mais técnico, como, editar regulamentos e procedimentos sobre proteção de dados pessoais e privacidade, bem como sobre relatórios de impacto à proteção de dados pessoais para os casos em que o tratamento representar alto risco à garantia dos princípios gerais de proteção de dados pessoais previstos na Lei; editar normas, orientações e procedimentos simplificados e diferenciados, inclusive quanto aos prazos, para que microempresas e empresas de pequeno porte, bem como iniciativas empresariais de caráter incremental ou disruptivo, que se autodeclarem startups ou empresas de inovação, possam adequar-se a esta Lei; deliberar, na esfera administrativa, em caráter terminativo, sobre a sua interpretação, as suas competências e os casos omissos; e, implementar mecanismos simplificados, inclusive por meio eletrônico, para o registro de reclamações sobre o tratamento de dados pessoais em desconformidade com esta Lei.

Nesse sentido, percebe-se a relevância dessas competências estabelecidas, sendo necessário da forma mais enérgica possível definir e elaborar as diretrizes para a Política Nacional de Proteção de Dados Pessoais e da Privacidade, sobretudo, regulamentos, normas, procedimentos e iniciativas de modo que possa haver um entendimento e uma prática simplificada entre empresas, governo

e cidadãos sobre o controle e proteção na circulação de dados e informações pessoais.

5. Considerações Finais

A necessidade de estabelecer uma proteção e controle na circulação de dados e informações pessoais tem como arcabouço a proteção aos direitos fundamentais, entre eles considerado o direito a privacidade, liberdade e o livre desenvolvimento da personalidade da pessoa natural.

Buscando atender a uma necessidade que tem se expandido a nível mundial, em face da expansão das TIC, a proteção de dados e informações pessoais tem sido considerada, inclusive, um requisito para que países possam negociar com segurança sem ferir as garantias fundamentais estabelecidas nos seus próprios textos constitucionais. Assim, visando se adequar as necessidades para realização de negócios internacionais, o Brasil aprovou a LGPD, buscando regular essa seara de proteção.

A LGPD foi alterada posteriormente, para implementação da parte técnica dessa proteção, criando a ANPD, que será responsável pelo controle sobre o tratamento de dados e informações para permitir que estes, possam circular com segurança, garantindo assim, a proteção de direitos fundamentais já mencionados.

Resta claro, de forma conclusiva, que até a entrada em vigor da LGPD, esforços devem ser ampliados no sentido de por em prática uma cultura de proteção, que dependerá da implementação de uma Política Nacional de Proteção de Dados bem articulada. Inclusive para que possa envolver o cidadão, permitindo de forma simplificada, clara e transparente, que o mesmo possa consentir e conhecer o que empresas e o poder público fazem com seus dados e informações pessoais. Os relatórios disponibilizados deverão possibilitar a percepção de situações de risco à garantia aos princípios gerais, propiciando que as diretrizes tenham eficácia.

Referências

ALVES, Fabrício. Proteção de dados pessoais é a evolução da privacidade. In: SERPRO. **Notícias e Artigo**. 2019. Disponível em: https://www.serpro.gov.br/lgpd/noticias/protecao-dados-evolucao-privacidade. Acesso em: 21 set. 2019.

BRASIL. **Decreto-Lei nº 2.848**, de 7 de dezembro de 1940. Código Penal. Disponível em:http://www.planalto.gov.br/ccivil_03/decreto-lei/del2848compilado.htm. Acesso em: 15 out. 2019.

BRASIL. **Lei nº 13.709**, de 14 de agosto de 2018. Dispõe sobre a proteção de dados pessoais e altera a Lei nº 12.965, de 23 de abril de 2014 (Marco Civil da Internet). Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/L13709.htm. Acesso em: 21 set. 2019.

BRASIL. **Lei nº 13.853**, de 8 de julho de 2019. Altera a Lei nº 13.709, de 14 de agosto de 2018, para dispor sobre a proteção de dados pessoais e para criar a Autoridade Nacional de Proteção de Dados; e dá outras providências. Disponível em: http://www.planalto.gov.br/ccivil_03/_Ato2019-2022/2019/Lei/L13853.htm. Acesso em: 21 set. 2019.

GIL, Antônio Carlos. **Métodos e Técnicas de Pesquisa Social**. 6. ed. São Paulo: Atlas, 2008.

SAMPIERI, Roberto Hernández; COLLADO, Carlos Fernández; LUCIO, Maria del Pilar Baptista. **Metodologia de Pesquisa**. 5. ed. Porto Alegre: Penso, 2013.

SERPRO. O que é a Lei Geral de Proteção de Dados Pessoais? Dê um "giro" pela lei e conheça desde já as principais transformações que ela traz para o país. 2019. Disponível em: https://www.serpro.gov.br/lgpd/menu/a-lgpd/o-que-muda-com-a-lgpd. Acesso em: 21 set. 2019.

CARACTERIZAÇÃO DA PRODUÇÃO CIENTÍFICA E TECNOLÓGICA DAS DOUTORAS NO BRASIL

CHARACTERIZATION OF SCIENTIFIC AND TECHNOLOGICAL PRODUCTION OF PHD IN BRAZIL

Monique de Oliveira Santiago[1], Thiago Magela Rodrigues Dias[2], Felipe Affonso[3]
Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG), Av. Amazonas 7675 - Nova Gameleira - Belo Horizonte, (1) moniqueosantiago@gmail.com, (2) thiagomagela@gmail.com
(3) felipe-affonso@hotmail.com

Resumo:
A crescente participação das mulheres nas carreiras científica e tecnológica tem sido foco de diversos estudos que buscam traçar um perfil da sua trajetória e desempenho acadêmico na ciência. Neste contexto, este trabalho objetivou analisar a participação do conjunto de doutores que possuem currículos cadastrados na Plataforma Lattes e cujo gênero registrado seja do sexo feminino. Após a coleta dos dados, foi realizada a etapa de seleção dos currículos pelo critério de gênero, e tratamento dos dados obtendo um conjunto de 149.850 currículos cadastrados com gênero feminino e titulação máxima concluída doutorado distribuídas nas suas diversas áreas do conhecimento científico. Os dados das doutoras foram agrupados quanto a formação acadêmica, publicações, produções, orientações e grandes áreas de atuação, sendo possível analisar a evolução da produção científica e tecnológica do conjunto de forma temporal. Estudar os diversos aspectos da diferença de gênero em geral e particularmente na ciência e tecnologia, além de ser relevante, pode ser fonte de inspiração para políticas e programas de governo que buscam promover mudanças, incentivar e valorizar a participação das mulheres na ciência.
Palavras-chave: Mulheres na ciência; Plataforma Lattes; Produção científica e tecnológica; Gênero feminino; Bibliometria.

Abstract:
The increasing participation of women in scientific and technological careers has been the focus of several studies that seek to outline their trajectory and academic performance in science. In this context, this work aimed to analyze the participation of the group of doctors who have curricula registered in the Lattes Platform and whose registered gender is female. After data collection, the stage of selection of curricula by gender criteria was performed, and data processing obtained a set of 149,850 curricula registered with female gender and maximum doctorate degree completed distributed in its various areas of scientific knowledge. The data of the PhDs were grouped in terms of academic background, publications, productions, orientations and major areas of activity. It is possible to analyze the evolution of scientific and technological production of the set in a temporal way. Studying the various aspects of gender difference in general and particularly in science and technology, as well as being relevant, can be a source of inspiration for government policies and programs that seek to promote change, encourage and value women's participation in science.
Keywords: Women in science. Lattes Platform. Scientific and technological production. Feminine gender. Bibliometrics.

1. Introdução

Dados referente a produção científica tem propiciado estudos que buscam compreender tanto a evolução da ciência, quanto a colaboração científica. Dentre os diversos estudos aplicados a esse conjunto, um que tem ganhado destaque são os relacionados ao gênero. Gênero corresponde a um campo interdisciplinar que tem como temática a identidade e a representação de homens e mulheres na sociedade. Este campo inclui o subcampo Estudo da Mulher que abrange, entre vários temas, a mulher e suas variadas relações com a ciência (LETA, 2014).

Por ser um tema interdisciplinar e abrangente, estudos que focam a mulher na ciência, possuem diversas abordagens que buscam traçar um perfil da trajetória e desempenho feminino na carreira. Apesar do progresso na participação feminina em vários segmentos na carreira acadêmica e científica, ainda percebe-se uma lacuna de gênero na ciência em todo o mundo que precisa ser melhor compreendida. Assim, realizar um estudo para averiguar a participação científica das mulheres no Brasil, é um passo para entender o cenário atual, e através dessa compreensão contribuir para a adoção de

medidas que promovam a igualdade entre os gêneros. Refletir e buscar novos meios de investigar a produtividade das mulheres na ciência é um passo para minimizar as desigualdades existentes nas carreiras.

Uma das maiores dificuldades ao analisar a produção científica de um país pode estar relacionada com a aquisição dos dados, que normalmente estão presentes em diversos repositórios. Entretanto, esse processo pode ser facilitado pela utilização do repositório de dados curriculares da Plataforma Lattes. Segundo Lane (2010), a Plataforma Lattes é considerada um importante conjunto de dados científicos brasileiro, no qual fornece informações de alta qualidade e possibilita pesquisar dados dos indivíduos que estão ali cadastrados, como formação acadêmica e produção científica, dentre outros.

Estudar grandes repositórios de dados torna-se uma tarefa complexa, pois a quantidade de dados a serem analisados e as características de cada repositório são únicas e em sua maioria não possuem padrão definido. Assim, utiliza-se da bibliometria que busca quantificar os processos de comunicação escrita, utilizando métodos para análises estatísticas sobre a produção e disseminação do conhecimento aplicadas a fontes de dados científicos (ARAÚJO, 2006). Logo, usa-se de análises bibliométricas como indicadores da produção científica com o objetivo de dispor os indicadores para o planejamento nacional para evolução das pesquisas científicas.

Portanto, utilizando os dados de acesso aberto disponíveis nos currículos cadastrados na Plataforma Lattes como principal fonte de dados para análises sobre a produção científica, surge uma excelente alternativa para estudos bibliométricos sobre a participação das mulheres na comunidade científica brasileira.

2. Objetivos

Este estudo tem como objetivo analisar a participação científica e tecnológica das mulheres brasileiras, investigando como suas pesquisas têm sido realizadas e como tem evoluído ao longo dos anos a partir de análises bibliométricas realizadas sobre dados curriculares disponíveis na Plataforma Lattes. Além de apresentar uma visão das mulheres que têm realizado pesquisas no Brasil, esse estudo visa contribuir para a geração de indicadores científicos nacionais e para a gestão das informações na área científica e tecnológica.

3. Procedimentos Metodológicos

Todas as informações curriculares da Plataforma Lattes são incluídas pelo próprio indivíduo e estão disponíveis livremente na internet. Esse amplo conjunto de dados contém todo o registro da trajetória profissional, acadêmica e produções científicas do pesquisador. Pela sua riqueza de informações e como esta fonte de dados não foi amplamente analisada, justifica-se assim a escolha da Plataforma Lattes como fonte de dados para medir e avaliar o desempenho científico nacional das mulheres. O processo de extração e seleção dos dados curriculares da Plataforma Lattes foi realizado através do arcabouço proposto por Dias (2016), denominado LattesDataXplorer, que possui um conjunto de técnicas e métodos responsáveis por coletar, selecionar, tratar e analisar os dados. Assim, o módulo de coleta do arcabouço foi responsável por coletar todos os currículos em março de 2019, ultrapassando 6.126.000 registros.

Para realizar uma análise detalhada da participação científica nacional das mulheres, optou-se por limitar os dados através do nível de formação acadêmica/titulação, reduzindo o conjunto para indivíduos que possuem o nível de formação doutorado concluído. Apesar deste conjunto não ser o mais significativo entre os níveis de formação, conforme enunciado por Dias (2016), eles são responsáveis por 74,51% dos artigos publicados em periódicos e 64,67% dos artigos publicados em anais de congresso, além de possuir em geral data de atualização de seus currículos recente e notadamente são responsáveis pelo mais alto nível de formação, a saber, mestrado e doutorado. Logo, após a aquisição de todos os currículos cadastrados na Plataforma Lattes, foi utilizado o módulo de seleção do LattesDataXplorer para selecionar, dentre estes, os currículos que possuem a formação acadêmica/titulação doutorado concluído, totalizando assim um conjunto com 307.780 currículos.

Depois das etapas de coleta dos dados curriculares e seleção dos doutores, foi realizado a seleção pelo critério de gênero. Nesta etapa foi solicitado junto ao CNPq uma listagem contendo todos os identificadores e seus gêneros respectivos, pois alguns campos coletados pela Plataforma Lattes não são exibidos na consulta pública, como exemplos: CPF, sexo, cor ou raça, dados do nascimento, identidade, passaporte, filiação, endereço residencial, entre outros. De posse da lista disponibilizada pelo CNPq, foi possível selecionar os identificadores que possuem o gênero feminino e utilizá-los para filtrar os que possuem identificadores iguais aos do arquivo, definindo assim o conjunto final de arquivos XML das doutoras com total de 149.838 currículos.

Os arquivos XML extraídos da Plataforma Lattes apresentam uma estrutura bem delimitada, com informações estruturadas e diversidade dos dados como, nível de formação acadêmica/titulação, grandes áreas de atuação, projetos de pesquisa e extensão, produções bibliográficas e técnicas como artigos publicados em anais de congresso e periódico, apresentação de trabalhos, participação em bancas, eventos, orientações, dentre outros. Como cada currículo possui uma quantidade específica dessas informações, foi realizado um tratamento com intuito de melhorar o entendimento dos dados. Portanto, a etapa de pré-processamento foi realizada acessando cada currículo XML, obtendo a informação específica daquele currículo e armazenando em uma coleção de arquivos estruturados. Ao final, esses dados foram agrupados quanto a formação acadêmica, publicações, produções, orientações, grandes áreas, entre outros. Após agrupar os dados, foi realizada uma caracterização para facilitar as análises dos dados e nesta etapa, para cada coleção de arquivos estruturados, foram gerados gráficos, tabelas, mapas, entre outros.

4. Resultados

Os dados coletados da Plataforma Lattes, utilizando o arcabouço LattesDataXplorer em março de 2019, totalizaram mais de 6.126.000 registros. Desse total, foram selecionados os registros com o nível de formação acadêmica/titulação doutorado concluído, totalizando em 307.780 (5,02%) currículos das diversas áreas do conhecimento científico. Esses mesmos dados foram selecionados pelo critério de gênero, no qual 149.838 (48,68%) correspondem a currículos das doutoras e 157.942 (51,31%) são currículos dos doutores.

Esses dados apresentam um crescimento significativo referente aos dados disponibilizados pelo Painel Lattes. A última extração realizada pelo Portal Lattes ocorreu em 30 de novembro de 2016, sendo a mesma disponibilizada para consulta pública, no qual forneceu dados referente ao painel Distribuição por sexo, faixa etária e área correspondendo a 63.853 (47,50%) currículos das doutoras e 70.567 (52,49%) currículos dos doutores. Logo, ao confrontar esses dados fornecidos pelo Painel Lattes com os dados atuais coletados da Plataforma Lattes, podemos perceber o crescimento de 128,97% dos pesquisadores que concluíram o doutorado em um período de mais de dois anos. Outro aspecto relevante refere-se ao percentual da formação de doutores, no qual o gênero feminino apresentou aumento de 1,18% e do gênero masculino diminuiu em -1,38%. Mesmo com esse considerável percentual, não foi suficiente para as doutoras ultrapassarem os doutores.

Como os doutores são responsáveis pela formação dos alunos em diferentes níveis de escolaridade, uma informação relevante para análise corresponde às orientações concluídas e em andamento das mulheres. Logo, quantificando esses dados temos um total de 3.577.801 orientações concluídas referentes a todas as orientações realizadas pelos doutores desde o início de sua carreira e que já estão finalizadas, e um total de 341.607 orientações em andamento que ainda não estão concluídas.

Um aspecto relevante na quantificação das orientações refere-se aos níveis de formação da pós-graduação: mestrado, doutorado e pós-doutorado. Nestes três níveis, a soma das porcentagens para as orientações concluídas corresponde a 17%, enquanto que para as orientações em andamento essa soma corresponde a 47%. Uma das hipóteses para essa diferença de percentual diz respeito ao fato de que, como

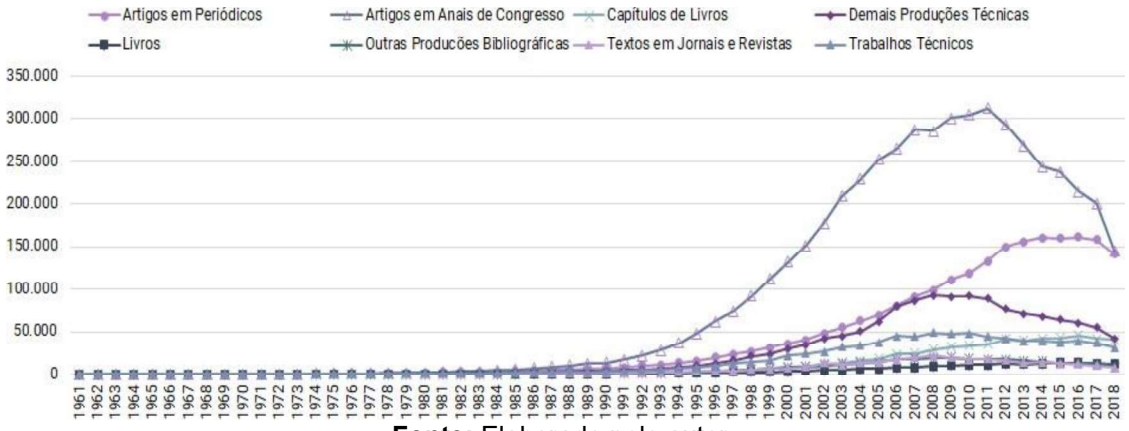
os doutores são responsáveis pela formação dos alunos nos principais programas de pós-graduação *stricto sensu* no Brasil, eles tendem a orientar mais alunos de graduação e menos alunos de pós-graduação no início de sua carreira, e com o passar dos anos após a conclusão do doutorado o número de orientações da graduação diminui e as orientações da pós-graduação aumentam. No entanto, como são consideradas todas as suas orientações em todo o seu histórico, as orientações em nível mais baixo de formação acaba por ser mais representativas.

Como o conjunto de doutores são responsáveis pela maioria dos trabalhos cadastrados na Plataforma Lattes (DIAS, 2016), realizou-se um levantamento das produções bibliográficas, técnicas e outras produções artísticas/culturais das mulheres. Com a quantificação desses dados e levando em consideração a produtividade dos doutores por ano, foi exibido na Figura 1, de forma temporal os tipos de produções mais relevantes. Nesta análise é possível perceber que, nos primeiros vinte anos (1961 à 1981) a produção científica permaneceu constante para todos os tipos de produções. Uma hipótese para explicar esse fato pode ser relacionado ao lançamento e padronização do currículo da Plataforma Lattes que ocorreu em agosto de 1999, no qual os dados referente as publicações anteriores a este período podem não ter sido divulgados pelos doutores, assim como, outro fator que precisa ser levado em consideração corresponde ao pequeno número de doutores na época.

A partir do início da década de 1980, ocorreu um aumento em todos os tipos de produções científicas e tecnológicas. As produções que se evidenciaram das demais correspondem aos artigos em periódicos e em anais de congressos. Com relação aos artigos em periódicos, o mesmo apresentou um crescimento significativo até 2013, permanecendo constante após esse ano e queda em 2017. Já as produções dos artigos em anais e congressos tiveram um aumento considerável com ápice no ano de 2011 e queda significativa após esse ano. Esse mesmo comportamento é apresentado no estudo de Dias (2016), para o conjunto de todos os doutores com currículos cadastrados na Plataforma Lattes. Esse declive acentuado referente aos artigos em anais de congresso foi tão expressivo que desde o ápice até o ano de 2018, apresentou uma queda de 53,62%, chegando ao final do último ano com valor total de artigos próximo aos total dos artigos em periódicos.

São diferentes hipóteses que podem estar relacionadas ao declive acentuado dos artigos em anais de congresso a partir de 2011. Uma delas refere-se a classificação da produção científica utilizado pela CAPES, se considerarmos que o sistema de avaliação da produção influencia as ações dos indivíduos. Fato esse impulsionado pela não consideração dos artigos em anais de congresso nas avaliações do programas de pós-graduação. Assim, os artigos em periódicos tornam-se mais interessantes para a publicação, pois influenciam nos conceitos dos programas no qual os doutores

Figura 1: Quantitativo das produções científicas por ano



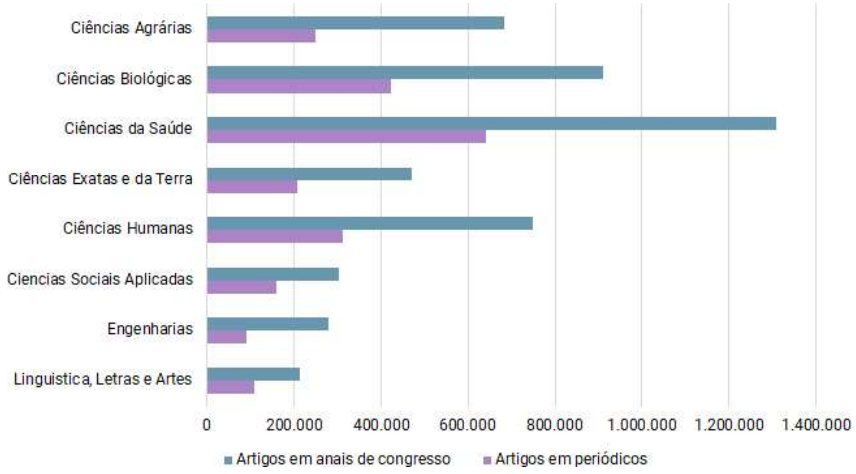
Fonte: Elaborado pelo autor.

participam, e os mesmos tendem a direcionar seus esforços para esse tipo de publicação.

Ao cadastrar as informações no módulo formação acadêmica/titulação na Plataforma Lattes referente ao mestrado e/ou doutorado, é possível escolher dentre nove opções que correspondem as grandes áreas de atuação e selecionar as áreas que se relacionam com a pós-graduação cursada. Assim, utilizando os dois meios de produção mais significativos para agrupar os currículos das doutoras por grandes áreas de atuação (Figura 2), como era esperado, obteve-se como grandes áreas de atuação mais significativas relacionadas a humanas e menos significativas correspondendo as engenharias. Esses dados corroboram com o estudo de Olinto (2011), comprovando que os homens predominam nas carreiras exatas e as mulheres tem predominância nas áreas biológicas e saúde.

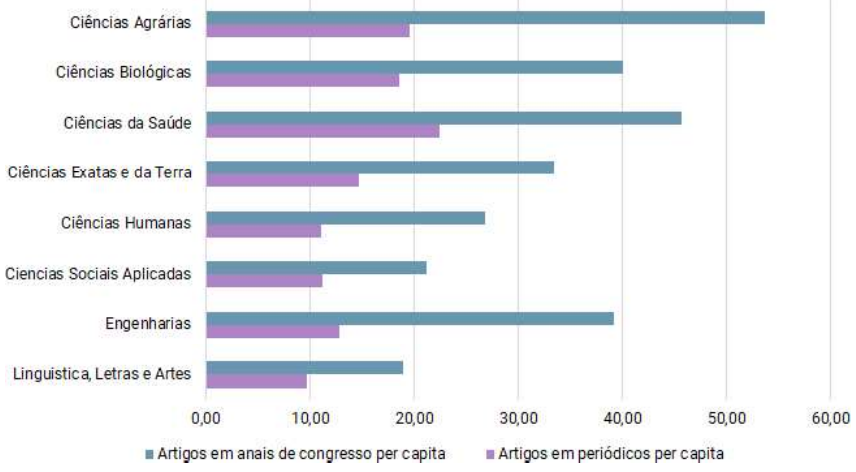
Dando continuidade a análise por grandes áreas, dividindo o total de artigos pelo total de indivíduos de acordo com a grande área de atuação correspondente, temos um número de produções per capita (Figura 3). Esse estudo torna-se mais interessante, pois é possível verificar a produção sem influência do grande número de autores de algumas áreas. Ao confrontarmos os dados das duas figuras anteriores, percebe-se uma alteração considerável de posição entre algumas áreas de atuação na produção de artigos em periódicos e anais de congresso. Para os artigos em anais de congresso, a primeira área que se destaca em relação as outras é a

Figura 2: Artigos em periódicos e anais de congresso por grande área de atuação



Fonte: Elaborado pelo autor.

Figura 3: Artigos em periódicos e anais de congresso per capita por grande área de atuação



Ciências Agrárias, no qual na figura 2 a mesma apresenta-se como a quarta maior produtora de artigos e na figura 3 aparece em primeiro lugar. Ciências da Saúde e Ciências Biológicas que estavam em primeiro e segundo lugar na figura 2, continuam bem significativas na figura 3, segundo e terceiro lugar respectivamente. A área de Engenharias teve uma alteração surpreendente e inesperada do penúltimo (Figura 2) para o quarto lugar (Figura 3). A Ciências Humanas que estava em terceiro, deslocou-se para o sexto lugar. As áreas de Ciências Exatas e da Terra e Linguística, Letras e Artes continuam na mesma posição para as duas figuras, enquanto que Ciências Sociais cai uma posição e aparece em sétimo lugar.

Para os artigos em periódicos a alteração das posições aparece menos significativa que para os artigos em anais de congresso. As áreas de Ciências da Saúde e Ciências Biológicas continuam em primeiro e terceiro, respectivamente. Ciências Agrárias do quarto lugar (Figura 2) surge em segundo lugar na figura 3. A área de Engenharias teve uma alteração surpreendente e inesperada do último (Figura 2) para o quinto lugar (Figura 3). Ciências Exatas e da Terra sobe uma posição e aparece em quarto lugar, enquanto que Linguística, Letras e Artes cai uma posição e aparece em oitavo lugar. Ciências Sociais Aplicadas continua na mesma posição para as duas figuras, enquanto que Ciências Humanas cai do terceiro para o penúltimo lugar na figura 3.

Esses dados nos revelam a importância da produção científica das doutoras nas diversas grandes áreas de atuação. Foi possível identificar o cenário de produções científica atual e verificar que mesmo as áreas que possuem menos participação feminina, como as engenharias, a produção per capita apresenta o cenário oposto.

5. Considerações Finais

Estudos que focalizam os diversos aspectos da diferença de gênero no trabalho em geral e particularmente na ciência e tecnologia, são relevantes e podem ser fonte de inspiração para políticas e programas de governo que promovam mudanças e levem a uma participação igualitária para mulheres e homens. Como este estudo apresenta todo o potencial oferecido pela Plataforma Lattes para a compreensão da participação científica das mulheres no Brasil, todos os resultados apresentados objetivaram apresentar uma visão da participação feminina de forma temporal sobre o conjunto de dados. Como trabalhos futuros espera-se realizar uma análise da colaboração científica das doutoras, das bolsistas de produtividade em pesquisa do CNPq e da produção científica das docentes que atuam em programas de pós-graduação.

Agradecimentos: Os autores agradecem ao CEFET-MG e CAPES pelo auxílio a pesquisa.

Referências

ARAÚJO, Carlos Alberto. Bibliometria: evolução histórica e questões atuais. **Em questão**, Universidade Federal do Rio Grande do Sul, v. 12, n. 1, p. 11-32, 2006.

Classificação intelectual: http://www.capes.gov.br/pt/avaliacao/instrumentos-de-apoio/classificacao-da-producao-intelectual

DIAS, Thiago Magela Rodrigues. **Um Estudo da Produção Científica Brasileira a partir de Dados da Plataforma Lattes**. 181 p. Tese (Doutorado em Modelagem Matemática e Computacional) — Centro Federal de Educação Tecnológica de Minas Gerais, Setembro 2016.

LANE, Julia. Let's make science metrics more scientific. **Nature**, Nature Publishing Group, v. 464, n. 7288, p. 488, 2010.

LETA, Jacqueline. Mulheres na ciência brasileira: desempenho inferior? **Revista Feminismos**, v. 2, n. 3, p. 139–152, Set.–Dez. 2014.

OLINTO, Gilda. A inclusão das mulheres nas carreiras de ciência e tecnologia no Brasil. **Inclusão Social**, Brasília, DF, v. 5, n. 1, p. 68–77, jul./dez. 2011.

Painel Lattes - Distribuição por Sexo, Faixa Etária e Grande Área de Atuação: http://estatico.cnpq.br/painelLattes/sexofaixaetaria

CLASSIFICAÇÃO AUTOMÁTICA DE TESES E DISSERTAÇÕES DA ÁREA DA CIÊNCIA DA INFORMAÇÃO SOB A ÓTICA DOS GRUPOS DE TRABALHO DA ANCIB AUTOMATIC

CLASSIFICATION OF INFORMATION SCIENCE AREA THESIS AND DISSERTATIONS ON ANCIB WORK GROUPS OPTICS

André Fabiano Dyck, Moisés Lima Dutra, Angel Freddy Godoy Viera
Universidade Federal de Santa Catarina (UFSC)
Florianópolis, SC - Brasil
andre.dyck@ufsc.br, moises.dutra@ufsc.br, a.godoy@ufsc.br

Resumo:
A recuperação de documentos pode ser melhorada com a aplicação da classificação textual. Um modelo simples e popular para extrair características de texto para classificação é o modelo *Bag-of-Words*. Bibliotecas de textos de teses e dissertações formam ricas coleções de documentos sobre a produção acadêmica no país. Instituições públicas de ensino superior disponibilizam sua produção científica em Repositórios Institucionais. Considerando que os eixos temáticos de pesquisa dos Programas de Pós- Graduação em Ciência da Informação têm um alinhamento com os Grupos de Trabalho da Associação Nacional de Pesquisa e Pós-Graduação em Ciência da Informação, a questão de pesquisa deste trabalho é dar subsídios para saber quais são as temáticas mais trabalhadas. Este estudo propõe uma arquitetura para classificação automática de teses e dissertações da área da Ciência da Informação sob a ótica dos Grupos de Trabalho da Associação Nacional de Pesquisa e Pós-Graduação em Ciência da Informação, utilizando a comparação de n-gramas, que são combinações de n-palavras que ocorrem no texto, não considerando questões semânticas. Um resultado esperado é o desenvolvimento de um protótipo para validar a arquitetura e identificar quais as temáticas mais trabalhadas no Programa de Pós-Graduação em Ciência da Informação, da Universidade Federal de Santa Catarina.
Palavras-chave: Ciência da Informação; Classificação textual; *Bag-of-*Words; N-Gramas

Abstract:
Document retrieval can be improved by applying textual classification. A simple and popular model for extracting text characteristics for classification is the Bag-of-Words model. Libraries of thesis and dissertation texts form rich collections of documents about academic production in the country. Public higher education institutions make their scientific production available in Institutional Repositories. Considering that the research thematic axes of research of the Graduate Programs in Information Science are aligned with the Working Groups of the National Association for Research and Postgraduate Information Science, the research question of this work is to give subsidies to know which are the most worked themes. This study proposes an architecture for automatic classification of thesis and dissertations in the field of Information Science on the optics of the Working Groups of the National Association for Research and Postgraduate Information Science, using the comparison of n-grams, which are combinations of n-words that occur in the text, not considering semantic issues. An expected result is the application of a prototype to validate the architecture and identify which themes are most worked on in the Postgraduate Program in Information Science, of the Federal University of Santa Catarina.
Keywords: Information Science; Text Classification; Bag-of-Words; N-Grams;

1. Introdução

Agrupar documentos em categorias é uma das soluções adotadas para agilizar o processo de recuperação de informação, cada vez mais relevante devido à inundação de oferta de informação dos dias atuais.

Estas categorias, ou rótulos, podem ser geradas por meio de intervenção humana, geralmente associando uma semântica que facilitaria a recuperação, ou usando apenas algoritmos computadorizados que utilizam outras características dos textos para agrupá-los, num processo que é um tipo de classificação.

A classificação é uma capacidade inerente do ser humano, que utiliza as categorias como ferramenta para entender o mundo, e este processo envolve uma série de etapas. Segundo Piaget, no construtivismo, o sujeito aprende com base na assimilação, na integração e na reorganização de estruturas que lhe permitem interpretar o mundo e interagir com ele. Ainda longe de mapear e simular este processo complexo, a classificação de texto

apenas atua na organização de informação por meio de atribuição de rótulos.

Em um sentido computacional, classificar é atribuir rótulos aos dados, que no caso da classificação textual são as palavras de um documento. A categorização, por outro lado, trata de agrupar documentos semelhantes, não rotulados, com base em alguma medida de similaridade. (INGERSOLL; MORTON, 2013). Neste trabalho concordamos com a distinção feita por Ingersoll e Morton (2013), de que a classificação textual (*text classification*) e a categorização textual (*text clustering*) são visões diferentes sobre os dados. Enquanto a primeira distingue - de forma que um dado é de uma categoria e não é de outra, de modo absoluto, a segunda considera a semelhança entre os dados dentro de uma mesma categoria, atribuindo níveis de especialização.

A oferta de repositórios de documentos, onde podemos fazer pesquisas livres para encontrar os mais variados temas, está aumentando. A localização manual de uma tese ou dissertação, de uma determinada temática, disponível em um repositório de documentos de uma instituição de ensino superior (Repositório Institucional - RI), passa pela leitura do título, resumo e palavras-chave e posterior avaliação para identificar se esta publicação é do eixo temático[1] desejado. Considerando o número de publicações existentes num RI, a localização manual de todas as teses e dissertações de uma determinada temática pode ser trabalhosa e demorada.

Os eixos temáticos de pesquisa dos Programas de Pós-Graduação (PPG) em Ciência da Informação (CI) têm um alinhamento com os Grupos de Trabalho (GT) da Associação Nacional de Pesquisa e Pós-Graduação em Ciência da Informação (Ancib). Esta pesquisa tenta responder as seguintes questões: Quais são as temáticas mais trabalhadas pelos PPG em CI? Qual é o alinhamento das teses e dissertações da área da CI com os temas dos GTs da Ancib? Como localizar teses e dissertações alinhadas com eixo temático específico, relacionado a um GT da Ancib?

Esta pesquisa faz parte de uma investigação mais ampla, em andamento na forma de uma tese de doutorado, referente à identificação da produção científica de uma instituição pública de ensino superior, através de suas publicações em seus portais Web de livre acesso.

A estrutura do artigo segue com: (i) a apresentação dos objetivos; (ii) os procedimentos metodológicos para alcançá-los; (iii) a arquitetura proposta e sua fundamentação teórica e; (iv) as considerações finais.

2. Objetivos

Este artigo tem o objetivo de propor uma arquitetura para classificação automática de teses e dissertações da área da CI sob a ótica dos GTs da Ancib.

3. Procedimentos Metodológicos

A realização deste trabalho baseia-se fundamentalmente no modelo de representação dos documentos chamado saco de palavras (*bag-of-words - BoW*). O *BoW* é um modelo de representação simplificado usado no processamento de linguagem natural[2], uma subárea da ciência da computação, para representar os documentos como um conjunto de palavras, sem considerar sua semântica original (HARRIS, 1954, GOLDBERG, 2017).

Ao trabalhar as coleções de textos como *BoW*, utilizando apenas as palavras e suas combinações presentes no texto, sem considerar questões semânticas, chamamos de n-gramas essas palavras ou suas combinações. Por exemplo, o unigrama "informação", onde o n-grama=1; o bigrama "big data", onde o n-grama=2 e; o trigrama "ciência da informação", onde o n-grama=3 (MOURA et. al., 2010, JURAFSKY; MARTIN, 2018). As coleções de textos, trabalhadas neste artigo, se referem aos resumos de teses e dissertações do Programa de Pós-Graduação em Ciência da Informação, da Universidade Federal de Santa Catarina (PGCIN/UFSC[3]), extraídos do RI/UFSC[4], e

1 Eixo temático no contexto deste trabalho significa um suporte ou guia para limitar os conteúdos de um assunto principal.

2 Linguagem natural no contexto deste trabalho é o que usamos na comunicação entre os humanos.

3 http://pgcin.paginas.ufsc.br/

às ementas dos GTs da Ancib, extraídas do site do “Fórum de Coordenadores de Grupo de Trabalho da Ancib”[5].

No presente estudo realizamos uma pesquisa aplicada, de cunho exploratório, cuja coleta e tabulação de dados se deram entre os dias 26/08/2019 e 09/09/2019, e que toma como cenário de aplicação o conjunto de 223 documentos das teses e dissertações em CI que estão disponíveis no RI/UFSC. Para obter os resumos, ao acessar o RI/UFSC, selecionamos a comunidade "Teses e Dissertações" e então a coleção "Programa de Pós-Graduação em Ciência da Informação". Em seguida, exportamos os metadados[6] dos 223 documentos. Estes metadados foram exportados para um arquivo CSV[7] e, então, lidos pela biblioteca Python Pandas. Com o arquivo CSV em memória, extraímos apenas a coluna de resumos.

De posse dos resumos e ementas, o passo seguinte é criar dicionários de termos que tomam por base o modelo *BoW*. A redução dos textos nas palavras que os constituem, forma os unigramas dos textos. Uma possibilidade de preservar um mínimo do significado do texto, usando ainda *BoW*, é utilizar também bigramas e trigramas (GOLDBERG, 2017). Assim, mantém-se a proximidade de duas e três palavras do texto original. Por esse motivo, como decisão de projeto, esta pesquisa criará unigramas, bigramas e trigramas para cada resumo e para cada ementa coletados, para então compará-los.

Para a criação dos três dicionários de termos, aplica-se as seguintes técnicas de limpeza dos dados: (i) Remoção de pontuação; e (ii) Transformação de todas as letras em minúsculas. Apenas para o dicionário de unigramas, aplicamos também a técnica de limpeza das palavras sem valor semântico (*stop words)*. Esta limpeza não é feita para os dicionários de bigramas e trigramas porque precisamos da semântica existente na proximidade de duas e três palavras no texto original. Com os três dicionários de termos prontos, contabiliza-se a frequência das ocorrências de cada termo.

Para que a comparação dos termos dos resumos com os termos das categorias / ementas dos GTs resulte em números expressivos, serão armazenados os termos com as cinco maiores frequências. Esta estratégia será avaliada na aplicação do experimento piloto sobre as publicações do PGCIN/UFSC, a ser realizado como próxima etapa da pesquisa no desenvolvimento da tese de doutorado.

Com os três dicionários de termos criados e limpos, (i) o dicionário de termos com unigramas; (ii) o dicionário de termos com bigramas; e (iii) o dicionário de termos com trigramas, tanto para os resumos das publicações do PGCIN/UFSC, como para as ementas dos GTs da Ancib, parte-se, então, para as comparações entre esses dicionários. Compara-se o dicionário de termos de cada tipo de n-grama que foi gerado a partir de cada resumo com o dicionário de termos das ementas dos GTs.

Como os trigramas possuem um valor semântico maior do que os bigramas, que por sua vez possuem uma semântica maior do que os unigramas, será atribuído um peso maior para as combinações de trigramas, depois para as combinações de bigramas e por último, um peso menor, para as combinações de unigramas.

Como resultado destas comparações e pesos, será criado um índice da combinação entre os três dicionários dos resumos de publicações do PGCIN/UFSC com os três dicionários das ementas dos GTs da Ancib.

4. Arquitetura Proposta

Aprendizagem de máquina é uma área de inteligência artificial que está preocupada em desenvolver algoritmos que aprendem padrões presentes em uma massa de dados (chamada de massa de dados de aprendizagem). Estes padrões aprendidos podem ser usados para predizer informações sobre dados novos, por isso a importância da massa de dados de aprendizagem ser diversa o suficiente para ampliar as chances

[4] https://repositorio.ufsc.br/

[5] http://gtancib.fci.unb.br/

[6] Metadados no contexto deste trabalho são dados sobre as dissertações e teses, como por exemplo: título, resumo, tipo de publicação, palavras-chave etc.

[7] CSV é um formato de arquivo onde seus dados são normalmente separados por uma vírgula. Esta sigla significa "***C**omma-**S**eparated-**V**alues*".

de predições (BAEZA-YATES, R.; RIBEIRO-NETO, 2013).

O uso deste tipo de algoritmos é extensivo em diagnóstico médico, detecção de fraudes a cartões de crédito, análise de mercado de ações, e recuperação de informação. Na recuperação de informação a classificação de textos é chave para o sucesso (BAEZA-YATES, R.; RIBEIRO-NETO, 2013).

Para automatizar a classificação de texto podemos fazer uso de várias técnicas e conceitos. Há principalmente três tipos de técnicas de aprendizagem: (i) Aprendizagem de máquina supervisionada, quando há intervenção humana na fase de treinamento; (ii) Aprendizagem de máquina não supervisionada, quando não há intervenção humana no treinamento, como, por exemplo, a técnica chamada de clusterização (*clustering*); e (iii) Aprendizagem semi-supervisionada, onde o conjunto inicial de dados é composto apenas por uma pequena entrada rotulada e grande parte dos dados da entrada não está rotulada, i.e., a categoria associada a eles é desconhecida. Neste caso, o objetivo é similar à classificação supervisionada, que é gerar uma relação binária mapeando a entrada para saída (BAEZA-YATES, R.; RIBEIRO-NETO, 2013).

Nesta pesquisa vamos usar um tipo de aprendizagem supervisionada, a classificação textual (*text classification*), uma vez que as categorias, ou seja, os GTs da Ancib, foram humana e previamente definidas.

Nosso objetivo de classificar documentos de textos, resumos das teses e dissertações, de acordo com categorias pré-definidas, pode descrever a tarefa como uma função: DxC{T,F}, onde D={d_1,d_2,...,d_{223}} é o conjunto que representa o domínio de documentos (*corpus*), no nosso caso 223 resumos de teses e dissertações, e C={c_1,c_2,...,c_{11}} é o conjunto pré-definido de categorias que são os 11 GTs da Ancib. O valor T atribuído a <d_j,c_i> indica uma decisão de classificar d_j como c_i, e F indica que d_j não é classificado como c_i (BAEZA-YATES, R.; RIBEIRO- NETO, 2013).

A Figura 1 esquematiza os módulos que constituem o modelo de arquitetura de classificação automática de teses e dissertações utilizando *BoW*.

1. Repositório Institucional: Ambiente de acesso livre e irrestrito à literatura científica e acadêmica da instituição.

2. Resumos de Teses e Dissertações: Os resumos extraídos do Repositório Institucional submetidos ao processamento de linguagem natural.

3. Classificador: Onde ocorre o treinamento e a obtenção da função de classificação.

4. Categorias / Ementas dos Grupos de Trabalho da Ancib: As ementas dos GTs da Ancib extraídos do site do "Fórum de Coordenadores de Grupo de Trabalho da Ancib", submetidos a processamento de linguagem natural utilizando a técnica *BoW*.

5. Fórum de Coordenadores de Grupo de Trabalho da Ancib: Site com a apresentação da Ancib e a descrição e ementas de cada um do seus GTs.

6. Resumos classificados em Grupo de Trabalho da Ancib: Resultados da classificação automática com os resumos classificados em uma categoria (um GT da Ancib).

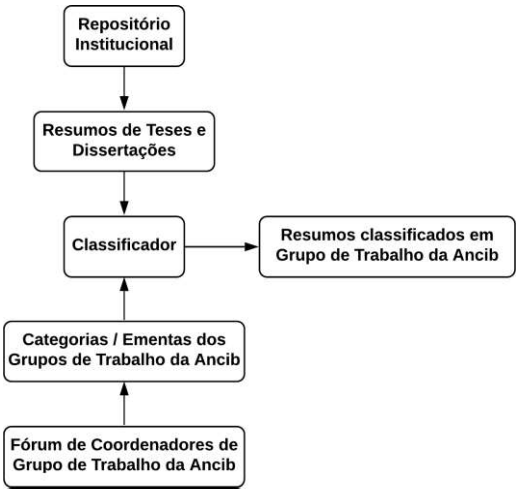


Figura 1 - Modelo de arquitetura de classificação automática utilizando *BoW*.
Elaborado pelos autores.

A partir da interação destes seis módulos chega-se à classificação automática dos resumos das teses e dissertações com relação aos GTs da Ancib.

O módulo 1 representa o RI, ambiente de acesso livre e irrestrito à literatura científica e acadêmica da instituição, que é a

fonte dos resumos das publicações de teses e dissertações.

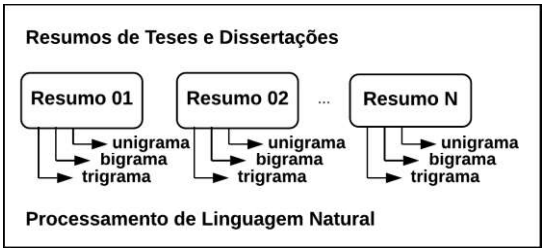


Figura 2 - Processamento de linguagem natural sobre os Resumos de teses e dissertações.
Elaborado pelos autores.

O módulo 2, detalhado na Figura 2, apresenta o processamento de linguagem natural, utilizando o modelo *BoW*, para a criação e limpeza de dados dos unigramas, bigramas e trigramas, para cada um dos resumos de teses e dissertações extraídos do RI.

O módulo 3 representa a classificação automática, onde ocorre o treinamento e a obtenção da função de classificação.

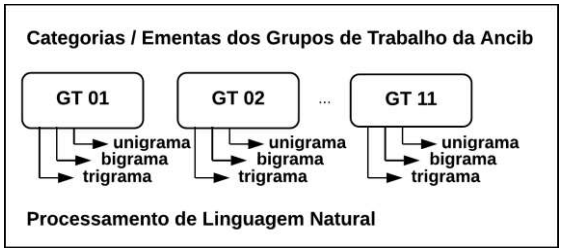


Figura 3 - Processamento de linguagem natural sobre as Categorias / Ementas dos GTs da Ancib.
Elaborado pelos autores.

O módulo 4, detalhando na Figura 3, apresenta o processamento de linguagem natural, semelhante ao módulo 2, utilizando o modelo *BoW*, para a criação e limpeza de dados dos unigramas, bigramas e trigramas, para cada uma das categorias / ementas dos 11 GTs da Ancib.

O módulo 5 representa o “Fórum de Coordenadores de Grupo de Trabalho da Ancib”, de onde foram extraídas as categorias / ementas de cada um dos GTs.

O módulo 6, detalhado na Figura 4, apresenta o resultado da classificação automática, onde cada um dos resumos de teses e dissertações foi classificado com uma probabilidade de pertencimento a determinado GT da Ancib.

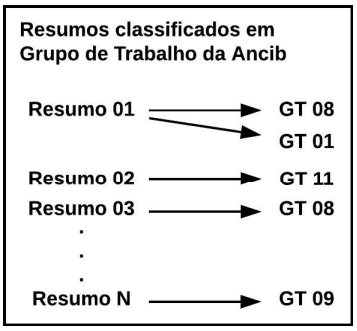


Figura 4 - Resumos classificados em GTs da Ancib.
Elaborado pelos autores.

O processo da classificação automática engloba:

(i) A comparação de cada resumo de tese e dissertação com cada uma das categorias / ementas dos GTs da Ancib. A comparação é feita entre unigramas, bigramas e trigramas, ou seja, o unigrama do resumo 01 é comparado com o unigrama da ementa do GT 01; o bigrama do resumo 01 é comparado com o bigrama da ementa do GT 01; o trigrama do resumo 01 é comparado com o trigrama da ementa do GT 01. Em seguida, é feita a comparação dos unigramas, bigramas e trigramas do resumo 02 com os unigramas, bigramas e trigramas da ementa do GT 01. Depois que todos os resumos de teses e dissertações foram comparados com a categoria / ementa do GT 01, inicia-se a comparação dos resumos com a categoria / ementa do GT 02. E assim, a comparação segue até a comparação dos resumos com a categoria / ementas do GT 11.

(ii) A geração de índices resultantes da comparação entre os unigramas, entre os bigramas e entre os trigramas de um resumo com uma categoria / ementa de um GT.

(iii) A atribuição de pesos para estes índices, sendo que os pesos para os índices dos trigramas serão maiores do que os pesos para os índices dos bigramas, que, por sua vez, serão maiores do que os pesos para os índices do unigramas.

(iv) A criação futura de uma fórmula com os pesos e os índices dos trigramas, bigramas e unigramas. A aplicação desta

fórmula nos resultados da comparação entre resumos e ementas definirá a probabilidade de pertencimento de um resumo a determinado GT. A construção desta fórmula se dará após a execução de diversos testes com diferentes corpora de documentos, numa fase posterior do trabalho.

5. Considerações Finais

Nesta pesquisa foi utilizado um tipo de aprendizagem supervisionada, com o objetivo de classificar teses e dissertações, de acordo com categorias pré-definidas. Utilizando o modelo *BoW*, propusemos uma arquitetura de classificação automática.

A definição do código, que compõe os módulos e os testes dos pesos a serem utilizados para diferenciar a relevância das combinações entre os n-gramas, será realizada na continuidade deste estudo, por meio de um piloto sobre as publicações de teses e dissertações do PGCIN/UFSC e, assim, validar a arquitetura e responder as perguntas levantadas neste trabalho.

Futuros trabalhos incluem a utilização de n-gramas maiores, para avaliar a quantidade adequada deles bem como, testar o modelo com um método não-supervisionado, a Clusterização, usando volumes maiores de dados. E, também, posteriormente, substituir o método *BoW* por uma técnica que preserva a semântica, como, por exemplo, vetores de palavras (*WordEmbeddings).*

Os resultados e a discussão da aplicação da arquitetura serão apresentados em artigo posterior.

Referências

BAEZA-YATES, R. e RIBEIRO-NETO, B. **Recuperação de Informação: conceitos e tecnologia das máquinas de busca.** 2. Ed. Porto Alegre: Bookman, 2013.

BAGGIO, Claudia Carmem. **Análise das políticas de informação dos repositórios institucionais das Universidades Federais do Brasil.** (Dissertação) Programa de Pós-Graduação em Ciência da Informação – Universidade Federal de Santa Catarina, 2016.

FAN, Weiguo, et. Al. **Tapping the power of text mining.** Commun. ACM 49, 9 (September 2006), 76-82. DOI: https://doi.org/10.1145/1151030.1151032

GOLDBERG, Yoav. **Neural Network Methods in Natural Language Processing.** Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers. 310p. Vol. 10, No. 1, Pages 1-309. 2017.

HARRIS, Zellig S. **Distributional Structure**. WORD, 10:2-3, 146-162, 1954. Publicado online em 04 dez. 2015. Disponível em:<https://www.tandfonline.com/doi/pdf/10.1080/00437956.1954.11659520> Acesso em: 11 ago. 2019.

INGERSOLL, Grant S.; MORTON, Thomas S.; FARRIS, Andrew L. 2013. **Taming Text: How to find, organize and manipulate it.** Shelter Island, NY (USA): Manning Publications Co., 2013. 298p.

JURAFSKY, Daniel; MARTIN, James H. **Speech and Language Processing An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.** Stanford University. Third Edition draft. 2018. Disponível em: <https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf> Acesso em: 04 set. 2019.

KUMAR, Prachi. An Introduction to N-grams: What Are They and Why Do We Need Them? **XRDS Crossroads The ACM Magazine for Students.** October 21, 2017. Disponível em: <https://blog.xrds.acm.org/2017/10/introduction-n-grams-need/> Acesso em: 31 ago. 2019.

MEIRELES, Magali Rezende Gouvêa; CENDÓN, Beatriz Valadares. **Categorização e Classificação de documentos a partir de suas citações: uma proposta baseada em Redes Neurais Artificiais.** Pesquisa Brasileira em Ciência da Informação e Biblioteconomia. Vol. 7, No 1. 2012.

MOURA, M. F.; NOGUEIRA, B. M.; CONRADO, M. S.; SANTOS, F. F.; REZENDE, S. O. **Um modelo para seleção de n-gramas significativos e não redundantes em tarefas de mineração de textos.** Boletim de pesquisa e desenvolvimento / Embrapa Informática Agropecuária, ISSN 1677- 9274; 23. Campinas, 2010.

DADOS ABERTOS E SUAS APLICAÇÕES EM CIDADES INTELIGENTES

OPEN DATA AND ITS APPLICATIONS IN SMART CITIES

Izabella Bauer de Assis Cunha[1], Frederico Cesar Mafra Pereira[2], Renata Maria Abrantes Baracho[3]
(1) UFMG, Av. Antônio Carlos, 6627, Pampulha - Belo Horizonte - MG, bellabauer89@gmail.com.
(2) UNA, Av. João Pinheiro, 565 - Centro, Belo Horizonte - MG, professorfrederico@yahoo.com.br.
(3) UFMG, Av. Antônio Carlos, 6627, Pampulha - Belo Horizonte - MG, renatambaracho@gmail.com.

Resumo:
Com a crescente urbanização, os estudos relacionados às Cidades Inteligentes (Smart Cities) buscam desenvolver soluções inovadoras para minimizar problemas urbanos, e consequentemente proporcionar melhor qualidade de vida para o cidadão e a sociedade. Esta pesquisa aplicada teve como objetivo propor um processo de Modelagem da Informação, necessário para subsidiar parâmetros indicativos para a concepção de cenários de Cidades Inteligentes, com o uso dos dados abertos em acidentes de trânsito, disponíveis no site da prefeitura de Belo Horizonte (PBH). A fundamentação teórica baseou-se em Modelagem da Informação, Cidades Inteligentes e administração pública com foco em mobilidade urbana. A abordagem da pesquisa é de caráter misto, combinando métodos quantitativos e qualitativos. Foi construída uma visualização analítica dos dados, em painéis de suporte à tomada de decisão, com intuito de oferecer uma proposta para subsidiar a construção de cenários para Cidades Inteligentes. A pesquisa mostrou a relevância das informações extraídas, gerenciadas e analisadas para a sociedade, provenientes de dados abertos, bem como as oportunidades para a contribuição das diferentes especialidades do campo da administração pública.
Palavras-chave: Cidades Inteligentes; Administração Pública; Dados Abertos; Modelagem da Informação.

Abstract:
Due the growth of urbanization, studies related to Smart Cities seek to develop innovative solutions to minimize urban problems, and consequently provide a better quality of life for citizens and society. This applied research had the objective of proposing an Information Modeling process, necessary to support indicative parameters for the design of Smart Cities scenarios, with extraction open data in traffic accidents, available on the website of the city of Belo Horizonte (PBH). The theoretical foundation was based on Information Modeling, Smart Cities and public administration focusing on urban mobility. The research approach is mixed in nature, combining quantitative and qualitative methods. An analytical visualization of the data was built in decision support panels, in order to offer a proposal to support the construction of scenarios for Smart Cities. The research showed the relevance of information extracted, managed and analyzed for society from open data, as well as opportunities for different specialties of the field of public administration.
Keywords: Smart Cities; Public Administration; Open Data; Information Modelling.

1. Introdução

A exploração das novas tecnologias da informação, a intensa concorrência do mercado, e as crises econômicas, proporcionam um desafio constante para a sobrevivência das organizações e cidades. Para auxiliar a tomada de decisões, principalmente em grandes centros urbanos, é eminente modelar, recuperar e gerenciar as informações.

Na busca de soluções inovadoras para enfrentar os desafios do aumento da disponibilização da informação e do crescimento das cidades, surge o conceito Cidades Inteligentes (Smart Cities). Refere-se a uma nova abordagem para minimizar problemas urbanos, desenvolvendo uma cidade mais sustentável e melhor para se viver, onde o conceito destaca-se como um ícone de qualidade de vida e sustentabilidade (ALAWADHI et al., 2012; CHOURABI et al., 2012).

No cenário da administração pública, estas soluções estão vinculadas à modernização por meio do uso de Tecnologias de Informação e Comunicação (TIC) e proporcionam a melhoria da eficiência dos processos operacionais e administrativos, bem como dos serviços públicos oferecidos aos cidadãos (DINIZ, 2009).

Dada a dificuldade em recuperar informação centrada na administração pública, foi desenvolvida a Modelagem da Informação permitindo a análise de um grande volume de dados abertos, e o agrupamento dos temas pela proximidade conceitual. Zandbergen (2017) ressalta que o objetivo principal dos projetos de Cidade Inteligente é a maior eficiência da administração pública, da comunicação e da descentralização política.

Considerou-se como objeto de estudo uma das áreas temáticas da administração pública a mobilidade urbana. Trata-se de dados abertos da prefeitura, relacionados a acidentes de trânsito, que, com o crescimento populacional e avanços tecnológicos, vem gerando sinais de alerta e necessidade contínua de prevenção. Os dados abertos visam garantir e facilitar aos cidadãos, à sociedade e às esferas públicas da federação, o acesso aos dados e informações produzidas ou custodiadas pelos órgãos, sobre o dia a dia da cidade, com o intuito de promover a interlocução com o governo, para construção de uma cidade melhor para se viver, trabalhar e visitar.

2. Objetivos

Neste contexto, o objetivo geral desta pesquisa aplicada é propor um processo de Modelagem da Informação, necessário para subsidiar parâmetros indicativos para concepção de cenários de Cidades Inteligentes, com o uso dos dados abertos em acidentes de trânsito, disponíveis no site da prefeitura de Belo Horizonte (PBH).

Os objetivos específicos da pesquisa são: compreender como se dá a estruturação de informações públicas para tomada de decisão, no departamento de acidentes de trânsito; identificar e analisar os dados abertos de acidentes de trânsito, disponíveis no site da PBH; propor a Modelagem da Informação de acidentes de trânsito; viabilizar o uso dos dados abertos aplicado a PBH; propor uma visualização de dados de acidentes de trânsito, para dar visibilidade aos cidadãos e aos órgãos públicos, através de painéis analíticos.

3. Procedimentos Metodológicos

Na busca de analisar os dados abertos da cidade de Belo Horizonte, no estado de Minas Gerais, foram identificados quais são os pontos principais a serem utilizados em uma Cidade Inteligente, na área temática de mobilidade urbana. Com o objetivo de propor um processo de modelagem de informações em acidentes de trânsito e criar soluções para a cidade, optou-se por uma pesquisa aplicada, que gera conhecimento para aplicações práticas dirigidas à solução de problemas específicos (GIL, 1994).

A abordagem da pesquisa utilizada é de caráter misto, empregando a combinação de métodos quantitativos e qualitativos para realizar seu processo investigativo. Creswell (2010) aborda esta abordagem como continuidade e evolução da metodologia de pesquisa, utilizando os pontos fortes dos dois métodos. A natureza interdisciplinar da pesquisa contribui para estas abordagens metodológicas diferentes. O uso combinado destes métodos proporciona mais insights e uma maior compreensão dos problemas de pesquisa.

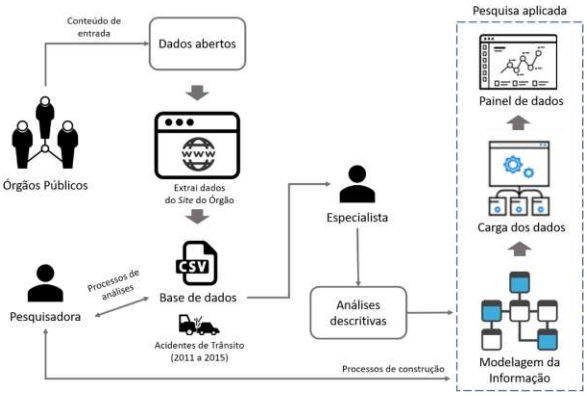
A pesquisa foi realizada em fases sequenciais, com concepção pragmática. Para a extração e tratamento dos dados abertos sobre acidentes de trânsito da cidade de Belo Horizonte, optou-se por utilizar o método quantitativo, que considera que tudo é quantificável, o que significa traduzir opiniões e números em informações as quais serão classificadas e analisadas (GIL, 1994). Para propor a Modelagem da Informação e analisar os dados levantados, optou-se pelo método qualitativo, que é mais indicado para investigações críticas e interpretativas, e se preocupa com a compreensão aprofundada de um grupo social ou de uma organização.

A estratégia utilizada para responder ao problema de pesquisa deste trabalho foi o estudo de caso, o qual, segundo Jung (2004, p. 158), “é um procedimento de pesquisa que investiga um fenômeno dentro do contexto local, real e especialmente quando os limites entre fenômeno e o contexto não estão claramente definidos”. Na concepção de Gil (1996), o estudo de caso é caracterizado pelo estudo profundo e exaustivo de um ou de poucos objetos, de maneira que permita o seu amplo e detalhado conhecimento.

O procedimento de coleta de dados foi por meio da extração dos dados abertos disponíveis no site da prefeitura de Belo Horizonte, relacionados ao tema de acidentes de trânsito, e a entrevista com especialista em mobilidade urbana, através de meio eletrônico. A entrevista qualitativa utilizada nesta pesquisa envolve poucas questões não estruturadas e em geral abertas, com objetivo de levantar concepções e opiniões do participante (CRESWELL, 2010).

Foram analisados os dados necessários para a modelagem de informações para Cidades Inteligentes, e os dados não relevantes foram descartados. O resumo do método proposto é mostrado na Figura 1:

Figura 1. Sistemas de Cidades e os Inter-relacionamentos.



Fonte: Elaborado pela autora (2018).

Os órgãos públicos do município possuem diversos sistemas e informações descentralizadas, acerca de suas diferentes áreas de competência. Estes órgãos disponibilizam informações de acordo com a Lei de Acesso à Informação, para a Prodabel (Empresa de Informática e Informação do município de Belo Horizonte), que processa os dados e publicam no site de dados abertos da PBH. A pesquisadora extraiu os dados relacionados a área temática de mobilidade urbana com foco em acidentes de trânsito, referentes aos anos que estavam disponíveis no site (2011 a 2015), totalizando 467.365 registros. Através das informações extraídas, desenvolveu um processo de análise, onde segmentou e correlacionou os dados. O especialista de mobilidade urbana da PMMG (POLÍCIA MILITAR DE MINAS GERAIS) teve acesso a estas informações da base, a partir da entrevista realizada pela pesquisadora, e propôs análises descritivas para os dados de acidentes de trânsito. Em sequência, a pesquisadora gerou processos de construção da pesquisa aplicada, propondo uma Modelagem da Informação para estruturação de uma visão consolidada dos dados, para futuras análises estratégicas. Em seguida, utilizou-se uma ferramenta de extração, transformação e carga, chamada SQL Server Integration Services v2016, para fazer a carga dos dados em uma base estruturada. Por fim, utilizou a ferramenta Power BI v2.65 da Microsoft, para propor painéis de visualização de dados analíticos das informações de acidentes de trânsito.

4. Resultados

Os métodos apresentados anteriormente na Figura 1 contribuíram para o resultado do processo de Modelagem da Informação dos dados de acidentes de trânsito. Foram analisadas em profundidade os arquivos coletados, com o intuito de relacionar os campos em comum, identificar possíveis dimensões por proximidade de temas e selecionar quais dados seriam relevantes para a pesquisa. Com isto várias informações redundantes entre os arquivos ou que não apresentavam dados, como “Não Informados” ou zerados, foram desconsiderados. Foram demonstrados em uma visão macro os subtemas de acidentes de trânsito como Boletim, Veículo, Logradouro e Envolvidos. No subtema Boletim, foram encontrados campos em comum (número, data e hora e origem) em todos os arquivos, para ligação destes dados. As dimensões agrupadas foram Acidente e Regional, que contém dados sobre o tipo de acidente e em qual regional municipal ocorreu. No subtema Veículo foram demonstrados os campos da espécie do veículo (ex: automóvel, motocicleta) e as dimensões agrupadas foram Situação, Socorro e Categoria, que contém os dados se o veículo estava em movimento ou parado, qual foi o tipo de socorro e se a categoria do veículo era particular ou alugado. No subtema Logradouro não

existem dimensões vinculadas, apenas os campos descritivos relacionados ao endereço do acidente. E por fim, no subtema Envolvido foram demonstrados os campos relacionados ao perfil do condutor ou vítima, se estava com cinto de segurança e apresentava sinais de embriaguez. As dimensões agrupadas foram Habilitação do envolvido e Severidade do acidente, conforme representado na Figura 2.

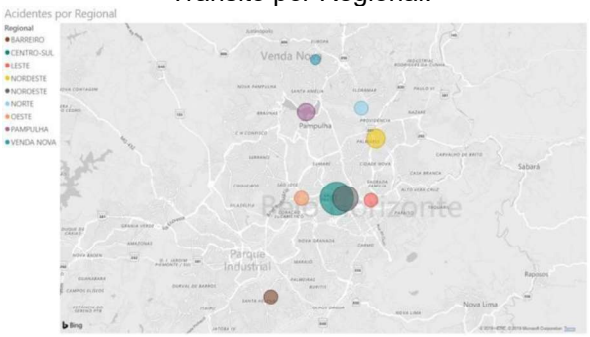
Figura 2. Modelagem da Informação.

Fonte: Elaborado pela autora (2018).

Após a proposta da Modelagem da Informação, os dados extraídos em arquivos CSV foram carregados em tabelas de bancos de dados estruturados, por subtemas e tópicos, utilizando ferramenta SQL Server Integration Services v2016 para extração, transformação e carga de dados, para facilitar o manuseio na construção das análises, realizadas através dos painéis de visualização de dados.

O conhecimento do especialista da PMMG foi utilizado para definir análises descritivas que seriam propostas nos painéis de visualização de dados. Os questionamentos realizados permitiram que a pesquisadora analisasse o cenário em diversas perspectivas,em busca de respostas assertivas, para evolução das cidades e prevenção de acidentes. As figuras seguir representam os painéis de visualização construídos com as correlações dos dados extraídos de acidentes de trânsito. Na Figura 3, para construção do painel, foram utilizadas como fonte de origem o subtema Boletim e a dimensão Regional. A quantidade de Acidentes de Trânsito por Regional mostra que o índice maior de ocorrências se concentra na região do Centro-Sul, com 20,4% de casos, e a segunda região é Noroeste, com 16%. A região com menor índice é Venda Nova, com 6,4% de acidentes, seguido da região do Barreiro com 8,1%.

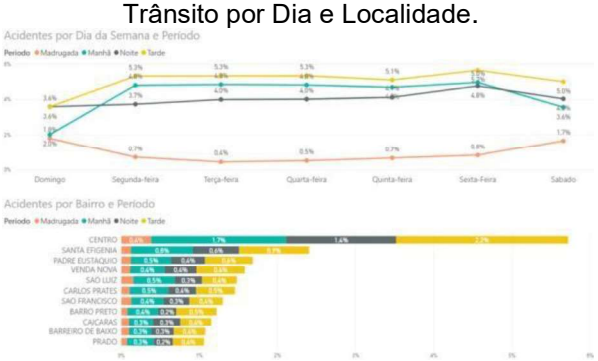
Figura 3. Painel de dados: Acidentes de Trânsito por Regional.



Fonte: Elaborado pela autora (2018).

O gráfico da Figura 4 comprova a concentração de acidentes no Centro de Belo Horizonte (5,7%), onde se tem um alto volume de veículos trafegando, e em segundo local, o bairro Santa Efigênia (2,4%). O período com mais ocorrências é na parte da tarde, de segunda a quarta-feira (5,3%). A fonte de origem do painel foram os subtemas Boletim e Logradouro. Estas análises respondem os dois primeiros questionamentos do especialista da PMMG, e demonstra que os acidentes podem ter relação com volume de tráfego, pela maior ocorrência ser durante a semana, e na parte da tarde, onde várias pessoas estão transitando, principalmente na região central da cidade.

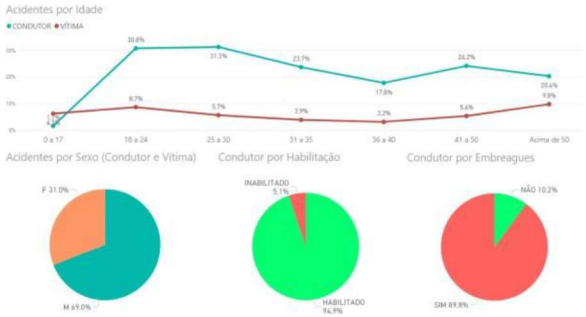
Figura 4. Painel de dados: Acidentes de Trânsito por Dia e Localidade.



Fonte: Elaborado pela autora (2018).

Na perspectiva do subtema Envolvido (vítima ou condutor), representado na Figura 5, o perfil se concentra nas faixas de idade entre 18 a 24 anos e 25 a 30, que juntas somam 62% dos acidentes registrados, predominantemente o sexo masculino, representando 69%, comparado a 31% do sexo feminino. Este resultado é proporcional à quantidade de homens habilitados, que representa aproximadamente 70% dos condutores, segundo a Associação Nacional dos Detrans (AND). Sobre a visão do condutor do veículo, a maioria é habilitado (94,9%), com sinais de embriaguez (89,8%), o que demonstra que álcool/drogas não combinam com direção. Existe ainda uma parcela de condutores sem habilitação (5,1%) que se envolvem em acidentes. Com relação aos questionamentos do especialista sobre tempo de habilitação do condutor e se a vítima tem dificuldade de locomoção/deficiência, seria possível responder caso houvesse mais informações nos dados abertos sobre o perfil dos envolvidos no acidente.

Figura 5. Painel de dados: Acidentes de Trânsito por Perfil do Envolvido.



Fonte: Elaborado pela autora (2018).

Diante da amostra dos dados abertos de acidentes de trânsito de Belo Horizonte, foi possível comparar os resultados apresentados com as perguntas do especialista em mobilidade urbana. Em relação aos locais com alto índice de acidentes de trânsito, a regional municipal Centro-Sul (20,4%) e o bairro Centro (5,7%) apresentaram a maior ocorrência, de segunda a quarta-feira (5,3%), no período da tarde, constatando que há uma possível relação entre o volume médio de tráfego e o número de acidentes. As características dos locais de acidentes de trânsito são normalmente vias pavimentadas, com limite de velocidade entre 40 a 60km/h, com fiscalização eletrônica em alguns pontos de velocidade e detecção de avanço de sinais.

Em análise ao perfil do condutor envolvido no acidente de trânsito, predomina o sexo masculino (69%), com a faixa etária entre 18 a 24 anos (29,6%), embriagado em sua maior parte (81,9%). Já para o perfil da vítima, predomina o sexo feminino (16,9%), com idade acima de 50 anos (9,8%).

5. Considerações Finais

De acordo com os objetivos apresentados, a pesquisadora conseguiu propor um processo de Modelagem da Informação, com os passos bem definidos e com a possibilidade de replicação de todo método proposto, desde a extração dos dados abertos de um órgão público, até a construção dos painéis de visualização dos dados, para subsidiar parâmetros indicativos para concepção de cenários de Cidades Inteligentes.

As informações coletadas e as análises realizadas trouxeram uma perspectiva e conhecimento sobre o que acontece na cidade, quais são as regiões, dias e horários que tem o maior índice de acidentes, tipo de veículos e severidade, entre outros. A prefeitura de Belo Horizonte pode utilizar este método para criar suas próprias soluções de monitoramento e prevenção de Acidentes de Trânsito e expandir estas análises para outras áreas temáticas da administração pública.

De forma abrangente, esta pesquisa contribuiu para o entendimento de como começar a abordar o assunto Cidades Inteligentes e sua implantação. Baseado nos trabalhos correlatos pesquisados, constatou-se que o estado da arte no tema Cidades Inteligentes deve-se iniciar no conceito de dados, através da modelagem da informação e da análise dos dados, e não na sua aplicação. O aprendizado proporcionado pela pesquisa foi da construção de um recurso metodológico, caracterizado por um modelo conceitual que permitisse a consolidação de fontes de origem distintas, para simplificar o entendimento pelos usuários finais (gestores públicos), permitindo a visualização das

possíveis aplicações em Cidades Inteligentes.

Como contribuição prática, a pesquisa demonstrou a viabilidade de ter acesso aos dados estratégicos de uma cidade, através da disponibilidade dos dados abertos, por áreas temáticas da administração pública. Além da modelagem destas informações, para facilitar a visão geral do tema de acidentes de trânsito, a visualização dos dados e análises, foi criado um banco de dados estruturado, com tabelas segmentadas por subtemas e dimensões.

Para subsidiar parâmetros de uma Cidade Inteligente, é necessário evoluir a pesquisa, em relação ao acesso aos dados atualizados de acidentes de trânsito de Belo Horizonte (2016 a 2018), que atualmente não se encontram disponíveis no site dos dados abertos da PBH. A limitação presente nesta pesquisa foi a análise dos resultados estar atrelada aos dados históricos de 2011 a 2015, podendo não retratar a realidade atual dos acidentes de trânsito em Belo Horizonte.

Como proposta de trabalhos futuros, sugere-se cruzar informações externas complementares, para ter como benefício uma análise completa, através do mapeamento dos hospitais e pontos de saúde mais próximos na região, rotas exclusivas para veículos de emergência, delegacias e corpo de bombeiro, helipontos, subsidiando a implantação de Cidades Inteligentes. Outra sugestão é a continuidade da parceria da Prodabel com a Prefeitura de Belo Horizonte, que poderão auxiliar na melhoria da disponibilização dos dados abertos nas diversas áreas temáticas, e replicação por outras cidades e órgãos do processo da Modelagem da Informação, proposto nesta pesquisa.

Outro trabalho indicado é o processamento e análise inteligente do volume de dados não estruturados (Big Data), gerados pelos sensores e dispositivos nas Cidades Inteligentes, através do desenvolvimento de aplicativos de Machine Learning, com o uso da Inteligência Artificial, possibilitando ações preventivas relacionadas à segurança, maior eficiência na tomada de decisões e perspectivas nunca antes exploradas.

Referências

ALAWADHI, S. et al. **Building Understanding of Smart City Initiatives.** In: SCHOLL, H. J., et al. (Eds.). Electronic Government: EGOV 2012. Lecture Notes in Computer Science. Berlin: Heidelberg, 2012. v. 7443. p. 40-53.

CHOURABI, H. et al. **Understanding Smart Cities:** An Integrative Framework. In: HAWAII INTERNATIONAL CONFERENCE ON SYSTEM SCIENCES. 45., 2012, Proceedings […]. Washington: IEEE Computer Society, 2012. Jan. p. 2289-2297. Disponível em: http://observgo.uquebec.ca/observgo/fichiers/78979_B.pdf. Acesso em: 20 jun. 2018.

CRESWELL, Jonh. W. **Projeto de Pesquisa:** métodos qualitativo, quantitativo e misto. 3. ed. Porto Alegre: Artmed, 2010.

DINIZ. **O governo eletrônico no Brasil:** perspectiva histórica a partir de um modelo estruturado de análise. Revista de Administração Pública-RAP, Rio de janeiro, n. 43, v. 1, p. 23-48, jan./fev. 2009. Disponível em: http://www.scielo.br/scielo.php?pid=S0034-76122009000100003&script=sci_abstract&tlng=pt. Acesso em: 12 jun. 2018.

GIL, A. C. **Como Elaborar Projetos de Pesquisas.** 3. ed. São Paulo: Atlas, 1996. p. 159.

GIL, A. C. **Métodos e técnicas de pesquisa social.** 4 ed. São Paulo: Atlas, 1994. p. 207.

PREFEITURA DE BELO HORIZONTE (PBH). **Dados Abertos.** Políticas de Dados Abertos. jun. 2018. Disponível em: https://prefeitura.pbh.gov.br/bhtrans/informacoes/dados/dados-abertos. Acesso em: 23 mar. 2018.

ZANDBERGEN, Dorien. **"We Are Sensemakers":** The (Anti-)politics of Smart City Cocreation. Public Culture, Durham, v. 29, n. 3, set. p. 539-562, set. 2017. Disponível em: https://doi-org.ez27.periodicos.capes.gov.br/10.1215/08992363-3869596. Acesso em: 3 out. 2017.

DADOS E METADADOS: REFLEXÕES CONCEITUAIS

Felipe Augusto Arakaki[1] , Ana Carolina Simionato Arakaki[2]
(1) Universidade de Brasília (UnB), Faculdade de Ciência da Informação, felipe.arakaki@unb.br; (2) Universidade Federal de São Carlos (UFSCar), Departamento de Ciência da Informação, acsimionato@ufscar.br

Resumo:
Diante da ordem epistemológica dos termos dado e metadado, o objetivo deste trabalho consiste em discutir e relacionar esses conceitos na Ciência da Informação. A pesquisa é caracterizada por uma metodologia de análise exploratória, sendo possível identificar elementos conceituais a partir da literatura científica, analisando as informações a partir do Perspectivismo. Como resultados, é apresentado as definições e relações entre 'dado' e 'metadado'. Considera-se que o levantamento e as discussões apresentadas no texto, demonstram diversas relações entre os conceitos, principalmente atrelados a *Web* Semântica, Ciência dos dados, *Big data*, entre outros, levantando a necessidade de aprofundamento nas definições e reflexões na área de Ciência da Informação, devido a divergências de perspectivas conceituais.
Palavras-chave: Dados; Metadados; Ciência da Informação; Dado e metadado.

Abstract:
On the epistemological order of the terms data and metadata, the objective of this paper is to discuss and relate these concepts in Information Science. The research is characterized by an exploratory analysis methodology, being possible to identify conceptual elements from the scientific literature, analyzing the information from the Perspectivism. As a result, the definitions and relationships between data and metadata are presented. It is considered that the survey and discussions presented in the text demonstrate several relationships between the concepts, mainly linked to the Semantic Web, Data Science, Big Data, among others, raising the need for deepening the definitions and reflections in the area of Science. Information due to divergences in conceptual perspectives.
Keywords: Data; Metadata; Information Science; Data and metadata.

1. Introdução

O termo dado é usado em diversos contextos, principalmente na perspectiva do atual cenário e das discussões envoltas nas questões do *Linked Data*, *Big Data*, *e-Science*, entre outros conceitos, como também nas tendências da Ciência de dados e da *Web* de dados.

Na área de Ciência da Informação, o termo dado obteve um novo contexto, influenciado ao conjunto de atualizações terminológicas e conceituais do termo metadado. Diante das diversas definições de dado e metadados, e da própria evolução dos conceitos relacionados a esses dois termos, Furner (2019) aponta a necessidade de discutir e comparar o conceito de metadados com outros termos, como exemplo: dados, documento, informação e os dados do próprio registro. Além disso, Baker (2011) ressalta a necessidade de uma discussão conceitual dos dois termos que são utilizados e aproximados entre as áreas de Ciência da Informação e Ciência da Computação.

Portanto, este trabalho busca compartilhar as reflexões do Grupo de Pesquisa "Dados e Metadados" sobre a importância em contextualizar e discutir sobre esses conceitos na literatura.

2. Objetivos

O objetivo deste trabalho consiste em discutir e relacionar os conceitos dos termos dado e metadado, no intuito de debater a proximidade dos dois conceitos na Ciência da Informação.

3. Procedimentos Metodológicos

Caracterizada por uma metodologia de análise exploratória para identificar os elementos conceituais, a partir da literatura científica da área de Ciência da Informação, os conceitos de dados e metadados.

Para a elucidação dos termos, utilizou-se o aporte teórico do Perspectivismo definido por Peterson (1996) e contextualizado para Ciência da Informação por Santos e Vidotti (2009). De acordo com as autoras "[...] cada um destes componentes que se pretende desenvolver no estudo dos processos que atuam nas diretrizes, modelagens e estruturas de sistemas para atendimento de necessidades de sujeitos em ambientes informacionais específicos." (SANTOS; VIDOTTI, 2009, p. 03).

4. Resultados e discussões

Ao longo dos anos, autores como Alves (2010); Alves e Santos (2013); Joudrey, Taylor e Wisser (2018); Méndez Rodríguez (2002); Pomerantz (2015); Zeng e Qin (2008, 2016), têm discutido e estruturado definições sobre o termo metadados na perspectiva da Ciência da Informação. Dentre esses autores, o que é possível de identificar como ponto em comum é que o conceito de metadados está atribuído a uma informação estruturada para as ações de identificação, descoberta, seleção, uso, acesso e gerenciamento.

O termo metadado é conceituado por meio das "[...] informações de valor agregado que criam para organizar, descrever, rastrear e melhorar o acesso a objetos de informação e itens físicos e coleções, relacionados a esses objetos". (GILLILAND, 2016, p. 02, tradução nossa).

Os metadados inicialmente foram identificados pela expressão 'dados sobre dados', cunhada na década de 60 para se referir a um conjunto de declarações sobre os dados (POMERANTZ, 2015). Mas é perceptível que os metadados fazem parte da rotina de diversas comunidades profissionais que projetam, criam, descrevem, preservam e usam sistemas e recursos informacionais (GILLILAND, 2016). Como afirma Haynes (2004) que a complexidade de compreensão dos metadados e as suas funções são essenciais para as atividades dos setores que envolvem conhecimento, informação, cultura e aprendizagem.

As funções dos metadados está direcionada a sua tipologia. A partir da análise terminológica realizada por Arakaki (2019, p. 80-81), as tipologias são identificadas como:

- **Metadados administrativos**: usados para gerenciar e administrar coleções e recursos informacionais, para auxiliar na tomada de decisão e manutenção dos registros e recursos informacionais. Fornecem informações sobre a origem e a manutenção de um objeto;
- **Metadados de autenticação**: são informações que possibilitam a identificação, integridade, legitimidade de um recurso informacional;
- **Metadados de preservação**: estão relacionados com informações de preservação e conservação dos recursos informacionais;
- **Metadados de proveniência**: estão relacionadas às informações de procedência, fornece dados sobre entidades, criação e modificações e seus relacionamentos;
- **Metadados técnicos**: estão relacionados a como um sistema funciona, fornecendo informações do sistema ou do recurso;
- ***Metametadata***: corresponde à informações sobre o registro criado, ou informações da criação de um conjunto de dados;
- **Metadados descritivos:** identificam características identificadoras e os contextos intelectuais dos recursos de informação para fins de descoberta, identificação, seleção, aquisição, contexto e compreensão;
- **Metadados de direitos**: estão relacionados às informações sobre propriedade, e direitos autorais;
- **Metadados de acesso e uso**: são informações de como um recurso informacional foi acessado e usado, como restrições de circulação e acesso, registros de exposições, entre outros;

- **Metadados estruturais**: está relacionado à composição e organização do recurso informacional;
- ***Markup languages:*** integra metadados e sinalizações para outros recursos estruturais ou semânticos.

As tipologias dos metadados estão presentes na seleção do padrão de metadados a ser utilizado no sistema informacional, conforme apontado por Zeng e Qin (2008) há uma intrínseca ligação no estabelecimento de metadados e formatos de metadados. A construção de um padrão de metadados exige a adoção de procedimentos metodológicos para a definição dos metadados, assim como eles, precisam estar em uma estrutura de descrição padronizada. Santos, Simionato e Arakaki (2014) apontam que a definição dos metadados deve ser uma ação consensual para que o sistema contenha interoperabilidade de seus dados, e ainda quando há dados ambíguos e que necessitam de um elevado detalhamento do recurso informacional.

Ainda, os metadados comprovam a autenticidade e o grau de completude do recurso, estabelecem o seu contexto, identificam suas relações estruturais com outros recursos, provêm diversos pontos acesso para diferentes tipos de usuários e podem fornecer informações que são geralmente obtidas por meios tradicionais. (GILLILAND, 2016).

Os dados são destacados inicialmente pelo conceito atribuído por Santos e Sant'Ana (2013, p. 205) que definem que dado como "[...] uma unidade de conteúdo necessariamente relacionada a determinado contexto e composta pela tríade entidade, atributo e valor, de tal forma que, mesmo que não esteja explícito o detalhamento sobre contexto do conteúdo, ele deverá estar disponível de modo implícito no utilizador, permitindo, portanto, sua plena interpretação." A partir da definição de Santos e Sant'Ana (2014), observa-se que o dado é composto pela tríade entidade, atributo e valor e apesar de implícito, o dado sempre está atrelado a um contexto.

Com o intuito de traçar uma evolução do conceito do termo dado, Furner (2016) faz um levantamento do conceito de dados ao longo dos séculos e discute a partir da perspectiva histórica a mudança do conceito. De acordo com o autor, o conceito dos dados podem possuir diversas perspectivas como

- **Abordagem extensional**: busca caracterizar coisas ou tipos de coisas que se enquadram no conceito de "dados";
- **Abordagem intencional**: identifica as propriedades que algo deve ter se for ser tratado como dados;
- **Abordagem classificatória**: reconhece um conceito individual como "dados" pode ter, ou possuir múltiplos sentidos, e que esses sentidos podem ser categorizados de acordo com similaridades em função e contexto.
- **Abordagem histórica:** que, ao invés de ou além delas, conduzir análises lógicas e / ou computacionais das propriedades necessárias dos conceitos, permitir que os autores considerem o desenvolvimento culturalmente específico dos significados de termos como dados ao longo do tempo.

A partir dessas abordagens (extensional, intensional, classificatória e histórica), Furner (2016) destaca as diversas interpretações que os dados podem assumir.

- **A interpretação clássica**: origem do termo do latim *dătŭm*, como verbo 'dar'. Furner (2016) discute a origem do termo em latim dătŭm, por volta do ano 100 ac, sendo utilizado muitas vezes como verbo (dar);
- **A interpretação documental**: dados como metadados. De acordo com Furner (2016), ainda por volta do ano de 100 ac, o termo 'dado' (*datum*) começou a ser utilizado como substantivo, como informação sobre algo, ou seja, como metadado;

- **A interpretação eclesiástica**: dados como dons de Deus. Furner (2016) relata que por volta de 1614, o termo dados começou a ser utilizado em sermões, com o significado de 'com a graça de Deus';
- **A interpretação geométrica**: dados como premissas geométricas. No contexto da geometria, o termo dado começou a ser utilizado por volta de 1645, para representar os valores dos lados e ângulos. (FURNER, 2016);
- **A interpretação matemática**: dados como premissas matemáticas. Após a interpretação geométrica, por volta de 1704, o conceito de dados foi ampliado para qualquer aplicação matemática, independente da área de aplicação. (FURNER, 2016);
- **A interpretação epistêmica**: dados como evidência. Por volta de 1648, o termo dado foi incorporado em alguns dicionários atribuindo o conceito de dados como fatos;
- **A interpretação informacional**: dados como valores de atributo. Na segunda metade do século XIX, há uma mudança na interpretação dominante do conceito de dados, sendo uma das principais alterações que o termo dados não apenas refere-se a informações numéricas;
- **A interpretação computacional**: dados como *bits*. O termo dado, na computação foi utilizado pela primeira vez em 1953, com a publicação do IBM701 *Electronic Data Processing Machine*, mas o termo só foi constar nos dicionários terminológicos da computação em 1980. A princípio o termo dado, foi utilizado para definir o valor do atributo em um banco de dados. Posteriormente, começou a ser utilizado como sinônimo de *bits*.
- **A interpretação diafórica**: por volta dos anos 2000, o termo relaciona-se a sua pluralidade, os dados são atribuídos como realidade objetiva, aparências subjetivas, observações, idéias, significados, ou mesmo, expressões linguísticas de observações individuais.

Corroborando com essa discussão focada em ambientes de bibliotecas, Baker *et al.* (2011) atribuem aos dados que são produzidos por meio de uma informação ou curadoria, são nomeados como 'dados de biblioteca", descrevem recursos ou ajudam a sua descoberta. Os dados de bibliotecas de dividem em: *datasets* (conjuntos de dados), *metadata element set* (conjuntos de elementos) e *value vocabularies* (vocabulários de valor) (BAKER et al., 2011).

Os *datasets* são coleções estruturadas de metadados para descrição de recursos, como livros, isto é, um conjunto de registros de metadados. Os *datasets* equivalem aos registros de bibliotecas que consistem em declarações, elementos da entidade e seus valores. Os elementos são definidos a partir de padrões, como MARC21 ou *Dublin Core* e os valores por vocabulários de valores, como a *Library of Congress Subject Headings* (LCSH) (ISAAC et al., 2011). Pode-se considerar exemplos como a *British National Bibliography*, o catálogo da *Hungarian National Library*, o *CrossRef* e a *Europeana* (BAKER et al., 2011).

Os *value vocabularies* definem os valores dos elementos para descrição de um recurso. Eles não definem informações de um recurso, e sim conceitos relacionados a um recurso como pessoas, assuntos, idiomas, países etc. (ISAAC et al., 2011), como exemplos: o *Library of Congress Subject Headings*, o *Virtual International Authority File (VIAF)*, a *Classificação Decimal de Dewey* (CDD) e o *GeoNames*.

Metadata element set é definido como "[...] um conjunto de elementos de metadados que definem as classes e atributos utilizados para descrever entidades de interesse." (ISAAC et al., 2011, não paginado, tradução nossa). O conjunto de elementos também designa a um conjunto completo de elementos de metadados como também a codificação dos elementos e estrutura em uma linguagem de marcação.

(ZENG; QIN, 2008). Pode ser citados como exemplos, o *Dublin Core Metadata Terms*, os elementos do *Resource Description and Access* (RDA), do *Simple Knowledge Organization System* (SKOS) e o *Friend of a Friend vocabulary* (FOAF) (BAKER et al., 2011).

Nesse sentido, a literatura aponta que em muitos casos, dados e metadados são tratados como sinônimos. Segundo Wickett et al. (2013, não paginado, tradução nossa) "[...] os componentes de dados e metadados estão entrelaçados: nenhuma distinção estrutural permite uma discriminação imediata entre dados e metadados."

Exemplificando uma situação no contexto das bibliotecas, Jeffery et al. (2014, não paginado, tradução nossa) distinguem que "[...] para o pesquisador, o registro da biblioteca são metadados para descobrir um livro ou artigo de interesse. Para o bibliotecário, o registro pode ser utilizado como dados para analisar a completude relativa das coleções por assunto, por editora, por ano, etc.". Ou seja, o mesmo objeto (dado/metadado) pode ser considerado ora metadado, como forma de descoberta e busca de um recurso informacional no catálogo, por exemplo. Ora poderá ser considerado como dado, no momento que o bibliotecário começar analisar os registros bibliográficos contidos na biblioteca como um todo, para realizar análises do acervo, de empréstimos, entre outras possibilidades.

De acordo com Hyvönen (2012, p. 10, tradução nossa), "[...] em torno de 2005, as ideias sobre *Linked Data* e *Web* de Dados começaram a ganhar impulso como uma abordagem simples para *Web* Semântica, focada na publicação de grandes conjuntos de dados existentes e usando apenas ontologias RDF simples e leves." Esse processo foi um dos movimentos para ressaltar a importância dos dados e ampliação das pesquisas na área.

Paralelo a esse movimento, tecnologias e aos novos paradigmas tanto da Ciência, como a *e-Science*, quanto da *Web* de documentos para *Web* de dados, impulsionaram a importância dos dados para os processos do dia-a-dia. Nesse contexto, novos conceitos relacionados aos dados têm surgido como *Big data* e até as discussões de um campo específico para tratar de dados como a Ciência de Dados.

5. Considerações Finais

Diante do levantamento e das discussões apresentadas no texto, observa-se que há diversas relações entre os conceitos de dados e metadados. O desenvolvimento tecnológico, atrelado às discussões e formalização da *Web* Semântica, Ciência dos dados, *Big data*, entre outros conceitos, levantaram a importância dos dados e dos metadados.

Entretanto, muitos autores discordam com esses conceitos e destaca-se a necessidade de aprofundamento em trabalhos futuros, para fomentar as discussões sobre a temática e contextualizar e delinear posicionamentos e perspectivas divergentes.

Referências

ALVES, R. C. V. **Metadados como elementos do processo de catalogação**. 2010. 132 f. f. Tese (Doutorado em Ciência da Informação) – Universidade Estadual Paulista, Faculdade de Filosofia e Ciências, Marília/SP, 2010. Disponível em: http://repositorio.unesp.br/handle/11449/103361. Acesso em: 8 set. 2019.

ALVES, R. C. V.; SANTOS, P. L. V. A. da C. **Metadados no domínio bibliográfico**. Rio de Janeiro: Intertexto, 2013.

ARAKAKI, F. A. **Metadados administrativos e a proveniência dos dados**: modelo baseado na família PROV. 2019. 139 f. Tese (Doutorado) - Doutorado em Ciência da Informação, Universidade Estadual Paulista "Júlio Mesquita Filho", Marília, 2019. Disponível em: https://repositorio.unesp.br/handle/11449/180490. Acesso em: 8 set. 2019.

BAKER, T. et al. **Library Linked Data Incubator Group Final Report**. W3C Incubator Group Report, 2011. Disponível em: http://www.w3.org/2005/Incubator/lld/XGR-lld-20111025/. Acesso em: 8 set. 2019.

FURNER J. "Data": The data. In: Kelly M., Bielby J. (eds) **Information Cultures in the Digital Age**. Springer VS, Wiesbaden, 2016.

FURNER, J. Definitions of "Metadata": A Brief Survey of International Standards. **Journal of the Association for Information Science and Technology**, 2019. doi:10.1002/asi.24295. Acesso em: 8 set. 2019.

GILLILAND, A. J. Setting the Stage. In: BACA, Murtha (Org.). **Introd. Metadata**. 3. ed. Los Angeles: Getty Research Institute, 2016. . Disponível em: http://www.getty.edu/publications/intrometadata/. Acesso em: 8 set. 2019.

HAYNES, D. **Metadata for information management and retrieval**. [S.l.]: Facet Publishing, 2004.

HYVÖNEN, E. **Publishing and Using Cultural Heritage Linked Data on the Semantic Web**. EUA: Morgan & Claypool Publishers, 2012.

ISAAC, A. et al. **Library Linked Data Incubator Group**: Datasets, Value Vocabularies, and Metadata Element Sets: W3C Incubator Group Report 25 October 2011. W3C, 2011. Disponível em: http://www.w3.org/2005/Incubator/lld/XGR-lld-vocabdataset20111025/#Published_Datasets. Acesso em: 8 set. 2019.

JEFFERY, K. et al. A 3-Layer Model for Metadata. INTERNATIONAL CONFERENCE ON DUBLIN CORE AND METADATA APPLICATION, 13., Portugal, **Anais**… DCMI, EUA. 2014. Disponível em: http://dcevents.dublincore.org/IntConf/dc-2013/paper/view/199/199. Acesso em: 8 set. 2019.

JOUDREY, D. N.; TAYLOR, A. G.; WISSER, K. M. **The organization of information**. 4. ed. Santa Barbara, California: Libraries Unlimited, 2018.

MÉNDEZ RODRÍGUEZ, E. **Metadatos y recuperación de información**. Gijón, Asturias: Ediciones Trea, 2002.

PETERSON. D. (Org.). **Forms of representation**: an interdisciplinary theme for cognitive science. Wiltshire: Cromwell Press, 1996. 208 p.

POMERANTZ, J. **Metadata**. Cambridge, Massachusetts ; London, England: The MIT Press, 2015.

SANTOS, P. L. V. A. da C.; SANTANA, R. C. G. Dado e Granularidade na perspectiva da Informação e Tecnologia: uma interpretação pela Ciência da Informação. **Ciência da Informação**, [S.l.], v. 42, n. 2, jan. 2013. ISSN 1518-8353. Disponível em: http://revista.ibict.br/index.php/ciinf/article/view/228. Acesso em: 8 set. 2019.

SANTOS, P. L. V. A. da C.; SIMIONATO, A. C.; ARAKAKI, F. A. Definição de metadados para recursos informacionais: apresentação da metodologia BEAM. **Informação & Informação**, Londrina, v. 19, n. 1, p. 146-163, fev. 2014. ISSN 1981-8920. Disponível em: http://repositorio.unesp.br/handle/11449/114736. Acesso em: 8 set. 2019.

SANTOS, P. L. A. C.; VIDOTTI, S. A. B. G. Perspectivismo e tecnologias de informação e comunicação: acréscimos à Ciência da Informação. **DataGramaZero**: revista de Ciência da Informação, Rio de Janeiro, v. 10, n. 3, 2009.

WICKETT, K. M. et al. Identifying content and levels of representation in scientific data. **Proceedings Of The American Society For Information Science And Technology**, [s.l.], v. 49, n. 1, p.1-10, 2013. Wiley. http://dx.doi.org/10.1002/meet.14504901199. Acesso em: 8 set. 2019.

ZENG, M. L. QIN, J. **Metadata**. New York: Neal-Schuman Publishers, 2008.

ZENG, M. L. QIN, J. **Metadata**. 2. ed. London: facet publishing, 2016.

EDUCAÇÃO A DISTÂNCIA E CIÊNCIA DE DADOS: DESENVOLVIMENTO DE MODELOS PREDITIVOS NO RECONHECIMENTO DA EVASÃO ESTUDANTIL

DISTANCE LEARNING AND DATA SCIENCE: DEVELOPING PREDICTIVE MODELS TO THE STUDENT EVASION RECOGNITION

Paulo R. V. do Carmo[1], Alan H. Costa[1], Sandro Rautenberg[1], Maria A. C. Knüppel[1], Marta C. R. Anciutti[1]

(1) Universidade Estadual do Centro-Oeste, Alameda Élio Antonio Dalla Vecchia, 838, Vila Carli, Guarapuava-PR, CEP 85040-167,
{pauloviviurka4, alanhenschel2, sandro.rautenberg, knuppelc, martanciutti}@gmail.com

Resumo: Na Educação a Distância têm-se a preocupação com a Evasão Estudantil, um problema que acomete o Núcleo de Educação a Distância da Universidade Estadual do Centro-Oeste. Com a utilização de algoritmos da Ciência de Dados pode-se reconhecer características discriminatórias acerca do fenômeno da desistência. Como uma pesquisa aplicada, visa-se formalizar o desenvolvimento de modelos preditivos para abstrair informação útil a partir da base de dados de Ambientes Virtuais de Aprendizagem, especificamente, do MOODLE. Como resultado, um *workflow* baseado no algoritmo de agrupamento de dados *k-means* é estabelecido. Este *workflow* permite o reconhecimento dos alunos em processo de evasão, auxiliando os tomadores de decisão no dimensionamento de atividades afirmativas, em se tratando a retenção de alunos.

Palavras-chave: Evasão Estudantil; Ensino a Distância; Modelos Preditivos; Agrupamento de Dados; Ciência de Dados.

Abstract: In Distance Education there is concern about student evasion, a problem that affects the Núcleo de Educação a Distância of the Universidade Estadual do Centro-Oeste. Using predictive algorithms, tech- niques covered by the Data Science, it is possible to recognize discriminatory features about the phenomena of withdrawal. In this sense, as an applied research, this work aims to formalize the development of predic- tive models for getting useful information from virtual learning environments, specifically, from MOODLE. As a result, a workflow based on the k-means data clustering algorithm is established. This workflow enables the recognition of students in the dropout process, assisting decision-makers, when formulating affirmative activities for retaining students.
Keywords: Student Evasion; Distance Education; Predictive Models; Data Clustering; Data Science.

1. Introdução

O uso inovador da Internet vem ampliando as plataformas e os serviços digitais, consequentemente, incrementando exponencialmente a produção de dados. Diversos aplicativos e dispositivos interconectados (computadores, *smartphones*, etc) relacionam uma série de eventos (van der AALST, 2014), armazenando enormes quantidades de registros, sinais, imagens, vídeos e *posts*. Ou seja, os dados são abundantes e veloz- mente produzidos, podendo servir como ma- téria-prima em processos decisórios (THE ECONOMIST, 2017). Por isso, o desenvolvimento de soluções computacionais que obtém informações de volumes de dados torna-se foco de investimento das organizações. Atualmente, a esse contexto relaciona-se a Ciência de Dados e a Tomada de Decisão, conforme exposto a seguir.

Primeiramente, tem-se que as fontes de dados mantidas pelas organizações tornaram-se enormes, dificultando a captura, o armazenamento, o gerenciamento, a análise e a exploração de dados por parte de ferramentas computacionais tradicionais (GARTNER, 2019; MANYIKA et al., 2011). Para auxiliar a análise e a exploração de dados, pode-se recorrer à Ciência de Dados (*Data Science*). A Ciência de Dados é caracterizada como uma camada de métodos devotados à extração de informação útil a partir de bases de dados cada vez mais complexas e dinâmicas (BUGNION; MANIVANNAN; NICOLAS, 2017). Diante essa perspectiva, ao

recuperar informação útil, pode-se auxiliar os gestores em suas atividades decisórias.

Pontualmente, este trabalho parte do pressuposto que as organizações voltadas ao Ensino a Distância (EaD) podem se beneficiar dos métodos e tecnologias da Ciência de Dados para melhorar seus processos. Principalmente, em face destas organizações adotarem Tecnologias da Informação e Comunicação (TICs), produzindo conjuntos de dados a serem explorados, a partir: **(i)** do gerenciamento da vida acadêmica de seus discentes; e **(ii)** do uso de meios digitais para o compartilhamento de material educacional.

No tocante às instituições de EaD, a Evasão Estudantil é um dos problemas principais enfrentados, que nacionalmente alcança a taxa média anual de 25% dos alunos matriculados (TAMARIZ; SOUZA, 2015). Esse problema também é enfrentado pelo Núcleo de Educação à Distância da Universidade Estadual do Centro-Oeste (NEaD/UNICENTRO). Atualmente, o índice de Evasão Estudantil da referida instituição está em torno de 54%.

2. Objetivo

Tecnologicamente, para minimizar os índices de Evasão Estudantil, Silva (2017) pontua a construção de Modelos Preditivos[1]. No contexto do EaD, modelos dessa natureza deveriam identificar preventivamente os indícios de um aluno em processo de evasão e suportar o processo de Tomada de Decisão dos gestores, diminuindo os índices de desistência discente. No tocante deste trabalho, pressupõe-se que o emprego de métodos e tecnologias da Ciência de Dados é pertinente ao estabelecimento de Modelos Preditivos da Evasão Estudantil.

Diante o exposto, neste trabalho investiga-se o desenvolvimento de Modelos Preditivos aplicados ao auxílio dos processos decisórios de minimizar a Evasão Estudantil no NEaD/UNICENTRO.

[1] Modelos preditivos implementam funções matemáticas identificam padrões de ocorrência de fatos futuros, dada uma base de dados históricos (TAURION, 2014).

3. Materiais e Métodos

Para subsidiar as atividades de EaD, o NEaD/UNICENTRO utiliza o Ambiente Virtual de Aprendizagem MOODLE (*Modular Object Oriented Distance LEarning*) como plataforma de suporte. O MOODLE oferece um ambiente dinâmico que oportuniza o aprendizado a qualquer momento e em qualquer lugar, ao utilizar a Internet como plataforma de comunicação (MOODLE, 2018). Atualmente, a base de dados do MOODLE do NEaD/UNICENTRO comporta os registros de 5.785 alunos de EaD (formados, desistentes ou em formação). De acordo com seu modelo relacional, os registros da interação de usuários para com os objetos disponibilizados são armazenados minuciosamente em 300 tabelas na referida base de dados. Diante dessa riqueza de detalhes, a base de dados do MOODLE é um insumo pertinente ao discernimento da Evasão Estudantil inerente ao NEaD/UNICENTRO.

Neste sentido, para promover a análise e a exploração, adota-se o procedimento metodológico proposto de Bugnion; Manivannan e Nicolas (2017). Os referidos autores sugerem sete passos pertinentes à Ciência de Dados, conforme segue:

- **Obtenção de Dados**. Preconiza as tarefas de avaliação e seleção de dados primários e seus metadados.
- **Ingestão de Dados**. Trata da transformação e da carga dos dados primários advindos de fontes diferentes e formatos diversificados em uma base de dados centralizada.
- **Exploração de Dados**. Privilegia a execução de estudos preliminares de modo a estabelecer conjecturas iniciais dos dados em relação à informação requisitada.
- **Definição dos Parâmetros**. Está intimamente ligado às escolhas necessárias para o emprego do(s) algoritmo(s) de Aprendizado de Máquina.
- **Implementação do Modelo**. Prima pela utilização dos algoritmos de Aprendizado de Máquina para estabelecer o modelo que melhor represente as características dos dados em análise.
- **Utilização do Modelo**. Avalia-se o poder de generalização do modelo perante situações do mundo real.

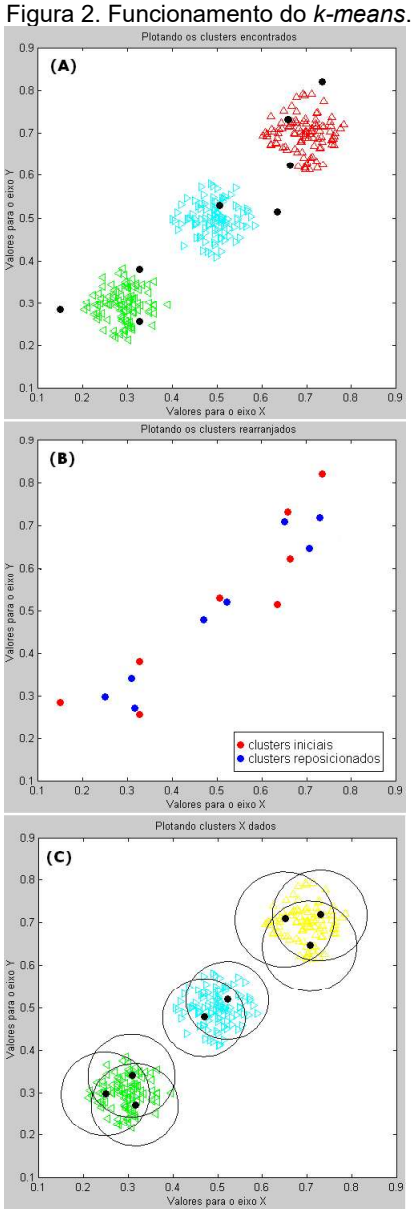
- **Tomada de Decisão**. Em situações reais, combinando o resultado gerado pelo modelo com o conhecimento particular do domínio, os gestores tomam suas decisões. Isso envolve a customização da apresentação/visualização das informações mediante relatórios, gráficos ou figuras, por exemplo.

Como suporte tecnológico ao trabalho, salienta-se o uso da Linguagem de Programação Python (PYTHON.ORG, 2018) e algumas ferramentas e bibliotecas. Neste sentido, utiliza-se:

- **PostgreSQL**[2]. Implementa um Sistema Gerenciador de Banco de Dados Objeto Relacional de código aberto.
- **JSON (JavaScript Object Notation)**[3]. Especifica um formato leve de intercâmbio de dados de fácil leitura e escrita.
- **MongoDB**[4]. Implementa um Banco de Dados de Documentos, com escalabilidade, flexibilidade e um amplo suporte a *queries*.
- **Pandas**[5]**.** Facilita o uso de estruturas e ferramentas de análise de dados.
- **Psycopg2**[6]**.** Permite o acesso rápido e seguro a bases de dados mantidas no Sistema Gerenciador de Banco de Dados PostgreSQL.
- **Scikit-learn**[7]**.** Disponibiliza vários algoritmos de Aprendizado de Máquina desenvolvidos em Python.

Ademais, vale ressaltar que na implementação do modelo preditivo é utilizado o algoritmo *k-means*. Conforme a literatura, o referido algoritmo é simples, eficiente e apresenta bons resultados ao ser empregado em tarefas de agrupamento em volumosas bases de dados (JAIN, 2010). Computacionalmente, o *k-means* é um algoritmo iterativo que separa um conjunto de dados em *k* subconjuntos representativos. Em poucas palavras, em seu processamento, o *k-means* estabelece uma medida de verossimilhança de *n* elementos de dados, considerando a minimização itera- tiva da distância destes elementos em rela- ção à um número definido de *k* centroides (do inglês, *cluster*).

Figura 2. Funcionamento do *k-means*.



Fonte: Dados da Pesquisa.

A Figura 1 ilustra o funcionamento do *k-means*. Na figura, é possível observar: **(A)** o conjunto de dados que será submetido ao processo de agrupamento para a divisão em três grupos e oito centroides previamente inicializados; **(B)** após a execução de *n* iterações do algoritmo, os centroides são reposicionados para melhor representar o conjunto

[2] Acesso: https://www.postgresql.org/.
[3] Acesso: https://www.json.org/.
[4] Acesso: https://www.mongodb.com/.
[5] Acesso: https://pandas.pydata.org/.
[6] Acesso: http://initd.org/psycopg/docs/.
[7] Acesso: http://scikit-learn.org.

de dados; e **(C)** - o conjunto de dados submetido e representado pela região de verossimilhança dos centroides reposicionados.

4. Resultado e Discussão

O desenvolvimento do trabalho é discutido em consonância às atividades propostas por Bugnion; Manivannan e Nicolas (2017).

4.1 Obtenção de Dados e Ingestão de Dados

Inicialmente, obteve-se um arquivo de *backup* dos registros da base de dados do MOODLE junto à Coordenadoria de Tecnolo- gia da Informação da UNICENTRO (COOR- TI/UNICENTRO). Então, procedeu-se a im- portação dos dados acerca de discentes para uma instância do PostgreSQL.

4.2 Exploração de Dados

Alguns *scripts* e *queries* foram desenvolvidos para melhor entender o modelo de dados e os registros da base de dados de MOODLE. Em face disso, alguns cenários de análise foram criados, com o intuito de encontrar parâmetros para o desenvolvimento de um modelo preditivo.

Destaca-se que, inicialmente, foram exploradas as análises da nota final do aluno em disciplinas e da quantidade de tarefas não entregues. Essas abordagens se mostra- ram pouco eficientes para estabelecer um modelo preditivo. Diante disso, mediante al- gumas reuniões com gestores do NEaD/UNICENTRO, decidiu-se que os dados utilizados seriam aqueles armazenados na tabela de *log* do MOODLE. Na referida tabela estão concentrados, principalmente, os campos de iteração do usuário na plataforma de EAD, contendo: **(i)** *actionid* - o código da ação realizada; e **(ii)** *targetid* - a tela em que a ação foi executada. Adicionalmente a es- ses dados, foram considerados o *userid* e *timecreated*, que contém respectivamente, a identificação do aluno que realizou a ação e o instante no qual a ação foi disparada.

Como resultado desse passo, uma tabela de *log* customizada foi criada. Neste sentido, outros dados ou rótulos pertinentes aos identificadores *userid*, *actionid e targetid* foram agregados em um arquivo no formato JSON. Posteriormente, o arquivo resultante for importado para uma instância do MongoDB

4.3 Definição de Parâmetros

Com os dados organizados, duas *queries* MongoDB foram desenvolvidas com o intuito de definir alguns parâmetros de representação dos dados, sendo:

- ***Query* A**. Para cada aluno, sumariza a execução trimestral de determinada ação e tela de atuação; e
- **Query B**. Para cada aluno, agrega todas as sumarizações em um vetor.

Ao executar as *queries*, os dados foram reorganizados, sumarizando as 57 ações possíveis de serem executadas quando da utilização do MOODLE/NEaD/UNICENTRO.

Quadro 1. Ações relevantes de discentes de EAD.

#	Ação	Descrição
1	*Created Discussion Subscription*	O total de respostas em postagens de fóruns
2	*Created Post*	O total de criações de postagens em um fórum
3	*Created Submission*	O total de submissões de tarefas
4	*Total Action*	O total de ações
5	*Total Created*	O total de ações de envio
6	*Total Viewed*	O total de ações de visualização
7	*Viewed Attempt*	O total de visualizações em tentativas de *quizz*
8	*Viewed Course*	O total de visualizações do curso
9	*Viewed Course Module*	O total de visualizações de tarefas
10	*Viewed Discussion*	O total de visualizações de postagens do fórum
11	*Viewed Grade Report*	O total de visualizações da grade de notas
12	*Viewed Message*	O total de visualizações de mensagens
13	*Viewed Submission Form*	O total de visualizações dos dados de envio de um *post* ou tarefa
14	*Viewed Submission Status*	O total de visualizações da situação de uma tarefa enviada

Fonte: Dados da Pesquisa.

Ao visualizar as 57 ações com um dendograma[8], observou-se quais são as ações de usuários preponderantes. Este grupo é formado por 14 ações que descrevem o comportamento típico de alunos no uso do MOODLE. O referido grupo de ações é enumerado no Quadro 1, sendo usado como

[8] Um dendograma é um diagrama que evidencia a relação hierárquica entre objetos (BOCK, 2019).

dados de entrada no modelo preditivo em desenvolvimento.

4.4 Implementação do Modelo

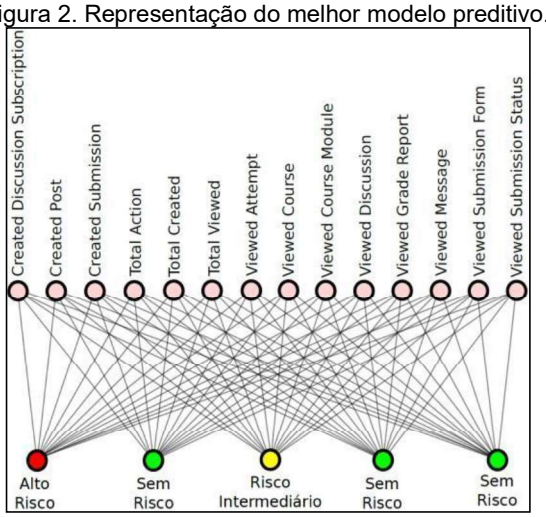
Para a implementação do modelo preditivo, foi desenvolvido um *script* em Python que utiliza o algoritmo *k-means* da biblioteca ScikitLearn. Este *script* foi executado 50 vezes para cada configuração testada. A Tabela 1 evidencia o número de agrupamentos considerado nos testes, a taxa média de acerto aferida e a variância média dos dados alcançada em relação aos agrupamentos formados.

Tabela 1. Modelos preditivos implementados e suas taxa de acerto *versus* variância nos agrupamentos.

#	Taxa de Acerto (%)	Variância
02	62.483	5.0390e-08
03	78.626	4.6148e-08
04	89.382	1.0582e-06
05	95.822	3.8399e-07
06	89.067	7.4852e-06
07	84.964	8.6147e-05
08	84.650	5.6599e-05
09	85.076	1.6210e-04
10	87.509	2.7443e-04
11	89.490	1.1840e-03
12	95.687	5.3677e-05
13	93.973	2.2468e-05
14	93.918	8.9078e-06
15	91.284	2.3249e-04
16	91.643	2.2567e-04
17	94.153	3.8435e-04
18	94.933	3.1329e-04
19	95.829	1.0323e-04
20	95.783	3.3889e-05

Fonte: Dados da Pesquisa.

Figura 2. Representação do melhor modelo preditivo.



Fonte: Dados da Pesquisa.

A partir dos testes realizados, selecionou-se a configuração com cinco agrupamentos (em destaque na tabela). Ressalta-se que, embora a configuração de 19 agrupamentos obtivesse a melhor média de acerto, sua variância (uniformidade do agrupamento em relação aos dados representados) é mais destoante. Isto significa que a configuração com cinco agrupamentos melhor se adequa para representar os dados utilizados. Essa configuração é representada na Figura 2, na qual o grau de evasão de um discente é categorizado como: alto risco, risco intermediário e baixo risco.

4.5 Tomada de Decisão

Considera-se que a atividade de Tomada de Decisão é reservada aos gestores do NEaD/UNICENTRO no tocante a agir proativamente na redução do índice de evasão estudantil.

Figura 3. Exemplo de relatório (parcial) enviado aos gestores do NEaD/UNICENTRO.

	Nome	Sobrenome	Evasão
1	[redacted]	CARPI	Alto Risco
2	[redacted]	RIBAS	Alto Risco
3	[redacted]	PEREIRA	Alto Risco
4	[redacted]	ZANDONA	Risco Intermediário
5	[redacted]	GIROTTO	Risco Intermediário
6	[redacted]	SOUZA	Risco Intermediário
7	[redacted]	QUADROS	Sem Risco
8	[redacted]	SILVA	Sem Risco
9	[redacted]	NASCIMENTO	Sem Risco

Fonte: Dados da Pesquisa.

Neste sentido, com a utilização do modelo preditivo implementado, é possível gerar um relatório, classificando o modo de utiliza- ção da plataforma MOODLE como sendo de um aluno com alto risco, risco intermediário ou sem risco de evasão. A Figura 3 exempli- fica parte um relatório gerado. Mediante o relatório, os gestores do NEaD/UNICENTRO podem contatar os discentes, preventivamente, questionando os motivos da desistência ou até dissuadi-los quanto à uma decisão repentina, por exemplo.

5. Considerações Finais

Este artigo disserta sobre um estudo aplicado, interdisciplinar aos conceitos da Ciência de Dados e EaD. Pontualmente, foi desenvolvido um modelo preditivo para a tomada de decisão quanto à evasão estudantil no NEaD/UNICENTRO. Seguindo o ciclo de vida de Ciência de Dados (BUGNION; MANIVANNAN; NICOLAS, 2017), implementou-se um *workflow* para relacionar alunos e seus respectivos riscos de evasão. Cabe ressaltar que modelo preditivo é baseado no algoritmo *k-means*, agrupando os alunos em três categorias de evasão: **(i)** alto risco; **(ii)** risco intermediário; e **(iii)** sem risco.

Em decorrência da experiência adquirida, são traçados como trabalhos futuros:

- estudos e aplicação de outros algoritmos de agrupamento ou classificação para comparação de resultados; e
- implementação de uma interface customizada do *workflow* desenvolvido de fácil utilização, possibilitando aos gestores do NEaD/UNICENTRO maior independência no estabelecimento de novos modelos preditivos.

Agradecimentos

À Secretaria da Ciência, Tecnologia e Ensino Superior (SETI/PR) pelo suporte financeiro (Projeto - Implementação da Universidade Virtual do Paraná – UVPR/SETI, Termo de Cooperação nº 145/2017, vinculado a unidade gestora do Fundo Paraná).

Referências

BOCK, T. **What is a Dendogram?** 2019. Disponível em: <https://www.displayr.com/what-is-dendrogram> . Acesso em: 11 jun 2019.

BUGNION, P.; MANIVANNAN, A.; NICOLAS, P. R. **Scala: Guide for Data Science Professionals**. Birmingham: Packt Publishing, 2017

ECONOMIST. **The world's most valuable resource is no longer oil, but data**. Disponível em: <https://goo.gl/AW4XsF>. Acesso em: 09 jun 2019.

GARTNER. **What is Big Data? – Gartner IT Glossary – Big Data**. Disponível em: <https://goo.gl/GwQWLA>. Acesso em: 23 mai 2019.

JAIN, A. K. Data clustering: 50 years beyond k-means. **Pattern Recognition Letters**, v. 31, n. 8, p. 651–666, 2010.

MANYIKA, J.; CHUI, M.; BROWN, B.; BUGHIN, J.; DOBBS, R.; ROXBURGH, C. B.; HUNG, A. **Big data: The next frontier for innovation, competition, and productivity**. Disponível em: <https://goo.gl/Vg2G2U>. Acesso em: 23 mai 2019.

MOODLE. Features – MoodleDocs. Disponível em: <https://docs.moodle.org/35/en/Features>. Acesso em: 23 mai 2019.

PYTHON.ORG. **What is Python? Executive Summary | Python.org**. Disponível em: <https://goo.gl/QYghos>. Acesso em: 28 mai 2019.

RAUTENBERG, S. et al. Evasão Estudantil e Ciência de Dados: primeiros passos de uma pesquisa aplicada no contexto da Educação a Distância da Universidade Estadual do Centro-Oeste. In: DIAS, G. A.; DUTRA, M. L. (Ed.) **Anais do WIDAT'2018 - II Workshop de Informação, Dados e Tecnologia**. João Pessoa: PPGCI-UFPB, 2018. p. 368.

SILVA, F. C. da. **Gestão da Evasão na EaD: Modelo Estatístico Preditivo para os Cursos de Graduação a Distância da Universidade Federal de Santa Catarina**. Florianópolis, 2017. 137 f. Dissertação (Mestrado) - Universidade Federal de Santa Catarina, Centro Sócio-Econômico. Programa de Pós-Graduação em Administração.

TAMARIZ, A. D. R.; SOUZA, M. de. Educação a distância no Brasil: perspectivas para redução na evasão de alunos matriculados. **Revista Científica Linkania Master**, v. 5, n. 1, 2015.

TAURION, C. **O que é um modelo preditivo?** Disponível em: https://cio.com.br/ o-que-e-um-modelo-preditivo. Acesso em: 09 jun 2019.

van der AALST, W. Data Scientist: The Engineer of the Future. In: Interoperability of Enterprises Systems and Applications Conference (I-ESA'2014), 2014, Albi-France, **Proceedings…** Heidelberg: Springer, 2014.

E-SCIENCE: DADOS GOVERNAMENTAIS ABERTOS À LUZ DA CIÊNCIA DA INFORMAÇÃO

E-SCIENCE: GOVERNMENT DATA OPEN IN THE LIGHT OF INFORMATION SCIENCE

Luiz Gustavo de Sena Brandão Pessoa[1], Tereza Ludimila de Castro Cardoso [(2)], Marckson Roberto Ferreira de Sousa [(3)]

(1) Universidade Federal da Paraíba - Cidade Universitária João Pessoa-PB gustavobrandao@bol.com.br
(2) Universidade Federal da Paraíba, - Cidade Universitária João Pessoa-PB luddyjampa@gmail.com
(3) Universidade Federal da Paraíba, - Cidade Universitária João Pessoa-PB marckson.dci.ufpb@gmail.com

Resumo:
A tecnologia da informação trouxe um paradigma que se preocupa com a questão da disseminação dos dados que estão disponíveis nos diversos ambientes informacionais. É nesse contexto que a ciência de dados traz uma problemática para a Ciência da Informação: analisar os aspectos emergentes de tratamento, uso e reuso dos dados abertos a partir das necessidades informacionais dos usuários. Esse artigo propõe uma reflexão do paradigma dos dados e do tratamento que a Ciência da Informação pode fazer com a disponibilidade dos dados governamentais abertos. A proposta buscou verificar se os municípios que compõem a microrregião do litoral norte do Estado da Paraíba estão disponibilizando os dados para a sociedade, como preceitua a Lei de Acesso à Informação. A metodologia utilizada corresponde a uma pesquisa documental e descritiva, com tratamento de dados realizado por estatística simples através do aplicativo *LibreOffice Calc*. Foi utilizado o modelo de indicadores disponibilizados pelos relatórios da Controladoria Geral da União. Os resultados demonstram que os municípios estudados iniciaram o processo de implantação, mas que em alguns itens ainda precisam de atenção do gestor, principalmente no que se refere a disponibilizar o dado em tempo real.

Palavras-chave: Paradigma dos dados; Dados Abertos; Ciência da Informação.

Abstract:
The information technology has brought a paradigm that is concerned with the issue of the dissemination of data that are available in the various informational environments. It is in this context that data science brings a problem to the information science: to analyze the emerging aspects of treatment, use and reuse of data opened from the informational needs of users. This article proposes a reflection of the data and treatment paradigm that information science can do with the availability of open governmental data. The proposal sought to verify whether the municipalities that comprise the microregion of the northern coast of the state of Paraíba are making available the data to society, as preceded by the law of Access to information. The methodology used corresponds to a documental and descriptive research, with data processing performed by simple statistics through the LibreOffice CALC application. The model of indicators made available by the reports of the Comptroller General of the Union. The results show that the municipalities studied started the implantation process, but in some items still need the manager's attention, especially in terms of providing the data in real time.

Keywords: Paradigm of the data; Open data; Information Science.

1. Introdução

No momento atual de complexidade econômica, política e social, a Ciência da Informação (CI) como uma área que estuda o fenômeno informação e seus aspectos funcionais e de tratamento de dados, nos faz refletir sobre as necessidades de se pensar a transparência das contas públicas à luz das mudanças tecnológicas e informacionais.

A informação surge como uma fonte inesgotável de evolução voltada para uma sociedade cada vez mais tecnológica. Para Capurro e Hjorland (2007) este conceito, enquanto conhecimento surge no contexto de explosão tecnológica no período pós Segunda Guerra, onde a informação desempenha um papel central para a sociedade. Para Capurro (2003), do ponto de vista epistemológico a CI apresenta três tipos de paradigmas: o físico, o cognitivo e o social. O paradigma físico remete aos sistemas informatizados, sendo este fortemente influenciado pela questão tecnológica (ALMEIDA et. al, 2007). Dessa forma, visa prioritariamente uma "gestão de dados" mais eficiente desenvolvendo e aperfeiçoando seus métodos. Os paradigmas da CI se relacionam com os paradigmas científicos, trazendo para a atual conjuntura a

significação necessária para a ciência, pois sem os sistemas informatizados, sem o usuário e sem a necessidade de informação de uma comunidade científica não haveria pesquisa de fato**.** Paralelamente a esse momento na CI, Jim Gray realiza através de ferramentas computacionais, experimentos com o tratamento de grandes quantidades de dados disponibilizados a partir de outros cientistas das mais diferentes áreas, como observam Hey, Tansley e Tolle (2009), para eles seria um momento novo para a história da ciência.

A partir dessa reflexão surge o termo e-Science que foi introduzido por Jhon Taylor em 2001, de acordo com Sales, Souza e Sayão (2014). Este autor definiu e-Science como algo que mudaria a forma de fazer ciência, através de uma colaboração global em áreas chave da ciência e de uma subsequente geração de infraestrutura que possibilitaria esta colaboração. Dessa forma, o e-Science apresenta-se como o quarto paradigma científico. A obra de Oliveira e Silva (2016), destaca também que o quarto paradigma científico trata de uma nova abordagem de comunicação científica, gerenciamento, curadoria, preservação com finalidade de colaboração mútua e acesso livre a publicação dos dados científicos. Esse paradigma é alicerçado na grande quantidade de bases de dados disponíveis através das ferramentas da tecnologia da informação que possibilita novas adequações frente às necessidades de usabilidade, organização, preservação e recuperação a partir das pesquisas que estão disponíveis pelas diversas áreas de conhecimento.

Para Saracevic (1996, p.58), a recuperação da informação é tida como o item mais importante da CI. Dessa forma, pode ser estudado o comportamento desse novo paradigma e suas implicações num contexto social, impulsionado pelos avanços relacionados às ferramentas da tecnologia da informação e comunicação. Com efeito, esse paradigma não vem para substituir os anteriores, mas traz em seu arcabouço uma forma de tratar os dados à luz de um novo momento científico que emerge. O conceito de dados abertos, na visão de Sayão e Sales (2013), está associado à livre disponibilidade para o reuso em outras investigações científicas, possibilitando outros tratamentos, aplicações e resultados. Essa possibilidade requer uma reflexão com relação aos direitos de autoria, patentes e outros mecanismos de controle de autoria intelectual. Dados para serem abertos devem estar disponíveis para *downloads* gratuitos, com livre permissão para cópias, verificações, aplicações e demais tratamentos que gerem novas descobertas e possibilidades de uso.

Considerando esse aspecto o Brasil vem adotando algumas práticas de transparência, governança e *accountability* – termo que implica na responsabilidade do gestor em prestar contas à sociedade de suas decisões de aplicação de recursos públicos ou privados -, participação nas contas públicas nacionais, através de ferramentas de disponibilização de dados à sociedade, para que se possa acompanhar a gestão que é materializada pelas tomadas de decisões e condutas do gestor público. Assim, uma dessas ferramentas foi a implantação de uma política de abertura de dados públicos à sociedade, que pode ser considerado uma quebra de paradigma para os gestores.

Assim, desde a implantação da Lei de Acesso à Informação – LAI, o governo federal vem buscando uma série de medidas visando a disponibilização dos dados abertos governamentais, favorecendo políticas de transparência e *accoutability*. Neste sentido, os Tribunais de Contas têm um papel preponderante no acompanhamento dessa implantação, uma vez que esses órgãos fazem todo o acompanhamento dos dados disponibilizados das contas públicas da União, dos Estados e dos Municípios. É nesse contexto que os dados abertos governamentais devem estar disponibilizados em seus portais em tempo real, assim como preceitua a LAI. Assim, o presente artigo busca realizar uma pesquisa com os municípios do estado da Paraíba com relação ao atendimento desse preceito legal a partir da seguinte questão de pesquisa: de que forma os municípios do estado da Paraíba estão disponibilizando à sociedade os dados para o acompanhamento da gestão municipal?

2. Objetivos

Como objetivo geral pretende-se investigar se os municípios que compõem o

Estado da Paraíba estão disponibilizando à sociedade, os dados para o acompanhamento da gestão pública municipal.

1.1 Como objetivos específicos, temos:

- realizar uma busca nos portais institucionais dos municípios pesquisados;
- verificar a partir de indicadores se os dados pesquisados estão disponibilizados de acordo com a legislação de acesso à informação;
- tratar os dados pesquisados, analisando-os estatisticamente.

3. Procedimentos Metodológicos

A presente pesquisa está em andamento e é fruto de um projeto de Extensão Universitária vinculado ao Campus IV da Universidade Federal da Paraíba. A proposta busca investigar se os 223 municípios que compõem o Estado da Paraíba estão disponibilizando as informações em seus portais institucionais. O Estado da Paraíba possui 4 mesorregiões subdivididas em 23 microrregiões. Para este trabalho serão apresentados resultados parciais a partir dos dados dos 11 municípios que compõem a microrregião do litoral paraibano, composta pelos municípios de Rio Tinto, Mamanguape, Mataraca, Marcação, Curral de Cima, Pedro Régis, Itapororoca, Jacaraú, Curral de Cima, Baia da Traição e Capim. Esta pesquisa é caracterizada como documental e quanti-qualitativa. Com a finalidade de alcançar o objetivo proposto, foi considerado o acesso digital aos portais institucionais dos municípios que compõem o litoral norte do Estado da Paraíba. O critério de escolha dos resultados parciais desses municípios foi por conveniência, em função de estarem localizados na microrregião geográfica do litoral norte, onde também está localizado o Campus IV da Universidade Federal da Paraíba. Quanto aos dados que foram verificados, utilizou-se como referência, o Modelo de Indicadores de Verificação para Avaliação dos Portais da Transparência, que são disponibilizados na apresentação dos relatórios da Controladoria Geral da União. Neste, estão contidos os indicadores com as categorias de pesquisa mais relevantes no processo de adequação de transparência e *accoutability*. O modelo categoriza o assunto de item e questiona se o dado está disponível ou não, no portal institucional do município.

Os dez indicadores utilizados na pesquisa foram elencados a partir da base legal da LAI, que buscava informação de dados quanto a:

1. regulamentação da LAI;
2. implantação de serviço de informação ao cidadão (SIC);
3. alternativa de enviar pedido de forma eletrônica ao SIC;
4. previsão e arrecadação de receitas;
5. empenho, liquidação e pagamento de despesas;
6. classificação orçamentária da unidade que financiou o gasto;
7. pessoa física ou jurídica beneficiada com o pagamento;
8. indicação de procedimento licitatório;
9. informação da despesa do bem ou serviço prestado;
10. atendimento do requisito “tempo real” (resultados específicos são apresentados na Tabela 2, em virtude de sua especificidade com relação aos indicadores 4 à 9).

O tratamento dos dados foi feito através de planilha eletrônica do *LibreOffice* e os campos foram preenchidos de forma codificada ao atendimento ou não da LAI, onde 1 corresponde a resposta afirmativa e 2 corresponde a resposta negativa. Foi feito uma estatística simples e descritiva, uma vez que os resultados são parciais. Entretanto, de posse da totalidade dos dados dos demais municípios, pretende-se tratá-los no sistema PSPP, que trata estatisticamente dados quantitativos (https://www.gnu.org/software/pspp/).

4. Resultados

A partir dos dados obtidos na pesquisa realizada nos portais dos municípios listados anteriormente, foi verificado que todos os municípios pesquisados possuem aspectos de transparência em seus portais, o que viabilizou a pesquisa, que se encontra em andamento.

Posteriormente, em todos os portais, foram realizadas buscas nos *links* que tratavam dos itens dos dados abertos, verificando-se que a maioria possuía o *link* “transparência fiscal” como metadados. O

acesso a esse *link* possibilitou verificar, a partir de cada indicador, se o portal disponibilizava a informação. O tratamento dos dados está demonstrado na Tabela 1 do Apêndice A.

Inicialmente foi verificado se os municípios regulamentaram a LAI. A pesquisa demonstrou que 100% dos municípios pesquisados haviam feito a regulamentação, que normalmente é feita através de decreto municipal. Neste sentido, o acompanhamento dos Tribunais de Contas vem sendo realizado de forma constante, e uma vez regulamentada, obriga os gestores sucessores a observarem essa norma.

O próximo item questionava se o município implantou o Sistema de Informação ao Cidadão. Observou-se também que todos os municípios haviam implantado esse serviço, que possibilita ao cidadão buscar o dado que deseja. Essa busca poderia ser feita pelo cidadão de forma presencial, entretanto o próximo item questionava se esse serviço de busca também poderia ser feito de forma eletrônica. Assim, foi constatado que todos os municípios disponibilizam o SIC também em formato eletrônico.

O indicador que tratava da apresentação de previsão e arrecadação de receitas demonstrou que todos os municípios também se adequaram a essa norma. A previsão das receitas ocorre em momento anterior à sua arrecadação. Já a arrecadação é realizada à medida que os ingressos acontecem. Como esses ingressos não ocorrem de maneira única, esse item precisa ser observado em consonância com o atendimento "em tempo real", conforme mostrado na Tabela 2.

Em contraponto aos ingressos, os indicadores que tratam da realização dos empenhos e pagamentos das despesas também foram avaliados. Observou-se que os municípios demonstram que estão realizando os empenhos e pagamentos das despesas.

Quanto à informação da unidade que realizou o gasto, os resultados demonstram que os municípios pesquisados estão disponibilizando essas unidades que estão executando gastos. Tais unidades podem ser em função dos recursos vinculados ao governo federal, como Saúde e Educação ou outros entes financiadores do gasto público. Em seguida, ainda em função dos pagamentosrealizados, o indicador aponta se os recursos são recebidos de pessoa física ou pessoa jurídica. Neste quesito, observa-se que 45% dos municípios pesquisados não apresentam os dados (Tabela 1), correspondendo aos municípios de Mamanguape, Itapororoca, Pedro Régis, Marcação e Cuité de Mamanguape. Esse dado é imprescindível para que a sociedade acompanhe quem está recebendo os recursos, interferindo na análise da *accountability* na gestão municipal, o que fica mais grave quando analisado em "tempo real", demonstrando que 64% dos municípios não apresentam dessa forma.

A indicador se há ou não procedimento licitatório também merece atenção dos interessados no acompanhamento da gestão. É nesse momento que está demonstrada a modalidade e as demais regras do processo de escolha de quem vai prestar o serviço ou vender o produto. Assim, foi demonstrado também que em 10% dos casos, os municípios não apresentam em seus portais os dados de procedimentos licitatórios realizados no âmbito da gestão. Quando verificados o quesito "tempo real", observou-se que o indicador fica em 36%, relacionados aos municípios de Cuité de Mamanguape, Marcação, Rio Tinto e Mamanguape. Em seguida o indicador que trata de informações sobre a prestação de serviço ou entrega do bem não foi apresentado em 27% dos casos, e quando se analisa sob o critério "tempo real", o percentual fica em 55%, referentes aos municípios de Marcação, Jacaraú, Itapororoca, Mataraca, Rio Tinto e Mamanguape.

Observa-se que a Tabela 2 atualiza a disponibilidade dos indicadores 4 a 9 em tempo real. Para isso, o município precisa ter uma equipe trabalhando esses dados de forma que estejam disponíveis no momento de ocorrência do fato gerador. Assim, esse indicador demonstrou uma limitação de informação principalmente nos municípios de Mamanguape e Marcação. O destaque da pesquisa ficou para os municípios de Baía da Traição, Curral de Cima e Capim, uma vez que na análise de seus portais institucionais, observa-se que a gestão está disponibilizando os dados para que a sociedade possa acompanhar a gestão.

5. Considerações Finais

A situação paradigmática do e-Science, demonstra que a Ciência da Informação está diante de um novo cenário que gera novos desafios. A pesquisa em tela demonstrou através dos dados parciais obtidos que não basta apenas uma legislação para que o gestor disponibilize os dados. Além de ser necessário o acompanhamento dos órgãos de controle e da sociedade, são necessárias também competências específicas para a geração da informação, a partir dos dados disponibilizados. Para o que esta pesquisa se propôs a fazer, seus objetivos foram atingidos, e assim foi possível afirmar que o maior problema encontrado até esse momento da pesquisa foi quanto ao atendimento da disponibilização do dado em tempo real. Essa indisponibilidade limita o trabalho do pesquisador e daquele que pretende acompanhar a gestão. Outro ponto, foi com relação aos dados abertos dos procedimentos licitatórios e das pessoas físicas ou jurídicas que recebem recursos públicos através da prestação de serviço ou venda de produtos. Esses pontos levam o pesquisador a refletir também sobre o alcance dos propósitos da abertura dos dados no Brasil.

Referências

ALMEIDA, D. P. R. et al. Paradigmas Contemporâneos da Ciência da Informação: a recuperação da informação como ponto focal. **Revista Eletrônica Informação e Cognição,** Marília, v.6, p.16-27, 2007. Disponível em: http://www.brapci.inf.br/_repositorio/2010/03/pdf_fc4f01292e_0008415.pdf. Acesso em: 28 ago. 2019

CAPURRO, R. Epistemologia e ciência da informação. In: ENANCIB, 5., 2003. Belo Horizonte. **Anais**... Belo Horizonte: UFMG, 2003. Disponível em: http://www.capurro.de/enancib_p.htm. Acesso em: 28 ago. 2019.

CAPURRO, R.; HJORLAND, B. O. O conceito de informação. **Perspectivas em Ciência da Informação**, Belo Horizonte, v.1, n.1, p.148-207, 2007. Disponível em: http://portaldeperiodicos.eci.ufmg.br/index.php/pci/article/view/54. Acesso em: 28 ago. 2019.

HEY, T.; TANSLEY, S.; TOLLE, K. (Ed.). Jim Gray on eScience: a transformed scientific method. In:_____. (Ed.). **The fourth paradigm: dataintensive scientific discovery.** Redmond:Microsoft Research, 2009. p. xvii-xxxi. Disponível em: http://digital.library.unt.edu/ark:/67531/metadc31516/. Acesso em: 31 ago. 2019

OLIVEIRA, A. C. S; SILVA, E. M. Ciência aberta: dimensões para um novo fazer científico. **Informação & Informação**. Londrina, v. 21, n. 2, p. 5 – 39, maio/ago. 2016.

SAYÃO, L. F.; SALES, L. F. Dados de pesquisa: contribuição para o estabelecimento de um modelo de curadoria digital para o país. Tendências da Pesquisa Brasileira em Ciência da Informação, Belo Horizonte, v. 6, n. 1, 2013. Disponível em: http://inseer.ibict.br/ancib/index.php/tpbci/article/viewArticle/102. Acesso em: 04 set. 2019

SALES, L. F.; SOUZA, R.F.; SAYÃO, L.F. Publicação Ampliada: um novo modelo de publicação científica voltada para os desafios de uma ciência orientada por dados. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 15., 2014, Belo Horizonte. **Anais**... Belo Horizonte: ECI/UFMG, 2014. p.3471-3492. Disponível em:http://enancib2014.eci.ufmg.br/documentos/anais/anais-gt7/view. Acesso em: 28 ago. 2019.

SARACEVIC, T. Ciência da Informação: origem, evolução e relações. **Perspectivas em Ciência da Informação**. Belo Horizonte, v. 1, n. 1, p. 41-62, jan./jun. 1996.

Apêndice A – Tabelas referentes ao tratamento dos dados de pesquisa nos municípios

Tabela 1 - Dados abertos nos portais dos municípios da microrregião do litoral norte do Estado da Paraíba

Relação dos municípios da microrregião do litoral norte do Estado da Paraíba	Cuité de Mamanguape	Marcação	Pedro Régis	Jacaraú	Curral de Cima	Baía da Traição	Itapororoca	Mataraca	Capim	Rio Tinto	Mamanguape	SIM	NÃO
1.Regulamentação LAI	1	1	1	1	1	1	1	1	1	1	1	100%	0%
2.Implementação do SIC	1	1	1	1	1	1	1	1	1	1	1	100%	0%
3.Pedido Eletrônico do SIC	1	1	1	1	1	1	1	1	1	1	1	100%	0%
4.Previsão e arrecadação de receitas	1	1	1	1	1	1	1	1	1	1	1	100%	0%
5.Empenho e pagamento da despesa	1	1	1	1	1	1	1	1	1	1	1	100%	0%
6.Unidade que financiou o gasto	1	1	1	1	1	1	1	1	1	1	1	100%	0%
7.PF ou PF beneficiária do pagamento	2	2	2	1	1	1	2	1	1	1	2	55%	45%
8.Indicação de Procedimento Licitatório	2	1	1	1	1	1	1	1	1	1	1	91%	9%
9.Prest. serviço ou inf.de entrega do bem	1	1	1	1	1	1	1	2	1	2	2	73%	27%

Fonte: Dados da pesquisa (2019)

Tabela 2 – Demonstração dos dados abertos nos portais dos municípios da microrregião do litoral norte do Estado da Paraíba que atendem ao requisito “tempo real”, conforme indicador 10.

Relação dos municípios da microrregião do litoral norte do Estado da Paraíba	Cuité de Mamanguape	Marcação	Pedro Régis	Jacaraú	Curral de Cima	Baía da Traição	Itapororoca	Mataraca	Capim	Rio Tinto	Mamanguape	SIM	NÃO
4.Previsão e arrecadação de receitas	1	1	1	1	1	1	1	1	1	1	1	100%	0%
5,Empenho e pagamento da despesa	1	1	1	1	1	1	1	1	1	1	1	100%	0%
6.Unidade que financiou o gasto	1	1	1	1	1	1	1	1	1	1	1	100%	0%
7.PF ou PF beneficiária do pagamento	2	2	2	2	1	1	2	2	1	1	2	36%	64%
8.Indicação de Procedimento Licitatório	2	2	1	1	1	1	1	1	1	2	2	64%	36%
9.Prest. serviço ou inf.de entrega do bem	1	2	1	2	1	1	2	2	1	2	2	45%	55%

Fonte: Dados da pesquisa (2019)

EXPLORANDO CONSULTAS SPARQL NA WIKIDATA COM PYTHON: TIPIFICAÇÃO DE METADADOS E RECONCILIAÇÃO DE DADOS

EXPLORING SPARQL QUERIES IN WIKIDATA WITH PYTHON: METADATA TYPING AND DATA RECONCILIATION

Luis Felipe Rosa de Oliveira [1], Dalton Lopes Martins [(2)]
(1) Universidade de Brasília, Brasília-DF, luisfelipeprf@gmail.com.
(2) Universidade de Brasília, Brasília-DF, dmartins@gmail.com.

Resumo:
Este estudo é desenvolvido sob a perspectiva da web semântica e dos dados abertos ligados, e é uma das vertentes de pesquisa do projeto Tainacan, a proposta é entender como desenvolver processos semiautomáticos de reconciliação de dados com a Wikidata. Como método, é apresentada a descrição do desenvolvimento de dois *scripts* de reconciliação em Python, um para tipificação de metadados, e outro para reconciliação de dados em texto. Como resultados são apresentados os scripts até o estado atual de desenvolvimento, apontando bibliotecas utilizadas, a estrutura da consulta SPARQL e como a busca e os resultados são processados e apresentados. Como conclusões entende-se que os *scripts* auxiliam no esforço de entender como dados de bases de dados podem ser ligados com objetos digitais da Wikidata, e espera-se que o desenvolvimento deles contemple uma das etapas de reconciliação de dados de acervos digitais disponibilizados na plataforma Tainacan, indicando o enriquecimento dos acervos e permitindo a aplicação de uma busca integrada semântica.
Palavras-chave: Wikidata; SPARQL; Reconciliação; Python

Abstract:
This study is developed from the perspective of the semantic web and open linked data, and is one of the research strands of the Tainacan project, the proposal is to understand how to develop semi-automatic data reconciliation processes with Wikidata. As a method, we present the description of the development of two Python reconciliation scripts, one for metadata typing, and another for text data reconciliation. As results are presented the scripts up to the current state of development, pointing out used libraries, the structure of the SPARQL query and how the search and results are processed and presented. As conclusions it is understood that scripts help in the effort to understand how database data can be linked with Wikidata digital objects, and it is expected that their development contemplate one of the steps of reconciling data from digital collections made available on the Tainacan platform, indicating the enrichment of the collections and allowing the application of an integrated semantic search.
Keywords: Wikidata; SPARQL; Reconciliation; Python

1. Introdução

Este estudo é coberto pela temática da web semântica e dos dados abertos ligados, mais especificamente das técnicas de reconciliação de dados, que são compreendias como formas de ligar dados com bases de conhecimento digitais (SANDERSON, 2016). Esse processo permite reconhecer entidades, seja no formato textual ou no formato de objeto digital em uma ou mais bases de conhecimento on-line, permitindo assim formar conexões entre bases por meio de identificadores únicos.

O caso de aplicação descrito neste estudo propõe entender como desenvolver processos semiautomáticos de reconciliação de dados com a Wikidata, que "é uma base de conhecimento livre e aberta que pode ser lida e editada por humanos e máquinas"[1].

O estudo aborda ainda a técnica de reconciliação de dados textuais para objetos digitais presentes na Wikidata, sendo descrito o desenvolvimento de dois *scripts*: o primeiro para tipificação de metadados, com o objetivo de reconhecer a qual instância da Wikidata determinado metadado pertence a partir de seu conjunto de valores. Já o segundo *script* busca a reconciliação de dados, em que determinados valores textuais são buscados na Wikidata em busca de identificar seus possíveis objetos representativos.

[1] Wikidata - https://www.wikidata.org/wiki/Wikidata:Main_Page

Esses scripts são desenvolvidos sob a necessidade de conectar diferentes fontes de conhecimento disponíveis on-line (por ex. GeoNames, Viaf, Wikidata) sem possuir um identificador em comum, através de valores textuais. Buscar esses valores nessas bases de conhecimento requer um conjunto de processos que vão desde um formato específico de busca (SPARQL[2]), até a validação dos objetos obtidos das bases de dados. Esse processo é denominado correspondência de entidades (DELPEUCH, 2019).

Um exemplo básico de aplicação desse processo seria identificar aspectos da vida de autores de obras de um acervo de obras de arte, em que só se tem o nome dos autores na base de dados. Utilizando a Wikidata é possível pesquisar o nome dos autores, e ter acesso a dados como, ano de nascimento e falecimento, nome de familiares, profissões e etc. permitindo contextualizar melhor o autor da obra, enriquecendo o conhecimento do acervo.

Porém, a consulta de nomes na Wikidata não é exata, como é um valor textual, podem aparecer vários resultados diferentes, por isso é necessária uma de validação do objeto retornado pela Wikidata, garantido que ele seja realmente representativo do valor textual buscado.

Além disso, o reconhecimento de valores textuais de uma base de dados em entidades de bases de conhecimento não é algo simples, exige conhecimentos técnicos como manipulação de consultas à dados no formato de triplas (sujeito, predicado e objeto), e formas de reduzir o ruído dos resultados obtidos.

Desse modo, espera-se com o desenvolvimento desses *scripts* de reconciliação, compreender melhor como dados de diferentes acervos podem ser ligados de forma semiautomática, complementando propostas de formas mais diretas de se reconhecer entidades em bases de conhecimento, e enriquecer bases de dados locais. Entendendo o modelo conceitual do acervo, não somente pelos documentos, mas pelos agentes relacionados a eles (ZENG, 2019).

O contexto de aplicação deste estudo se dá no âmbito do projeto Tainacan, uma plataforma livre de publicação de acervos digitais, desenvolvido a partir da necessidade de uma "solução tecnológica livre de fácil utilização e capaz de desmistificar o exercício da interoperabilidade entre modelos de dados dos diferentes domínios de acervos culturais" (MARTINS; CARVALHO JÚNIOR; GERMANI, 2019, p.60). Uma das vertentes de pesquisa do projeto é a aplicação de técnicas de web semântica e dados abertos ligados nos acervos de museus disponíveis na plataforma.

2. Objetivos

Este estudo tem como objetivo explorar o processo de reconciliação de dados com a Wikidata através de consultas SPARQL executadas por *script* em *Python*. Como objetivos específicos, busca-se descrever os processos de tipificação de metadados e reconciliação de dados, apontar as bibliotecas e funções utilizadas na obtenção de resultados da Wikidata e descrever como os resultados da busca são obtidos e podem ser apropriados pelo usuário.

3. Procedimentos Metodológicos

Para o desenvolvimento dos scripts de tipificação de metadados e reconciliação de dados com a Wikidata via consultas SPARQL, a primeira etapa foi identificar como as consultas podem ser estruturadas de forma a permitir busca de objetos através de texto livre.

Para isso, foi utilizada a ferramenta de consultas *on-line Wikidata Query Service*[3], o manual do utilizador, e as recomendações do W3C (World Wide Web Consortium) para a linguagem de consulta SPARQL[4]. Tanto o manual, quanto as recomendações permitiram realizar consultas de teste na ferramenta *on-line*. Dessa forma foi possível entender como pesquisar pelas *labels* dos objetos e como recuperar as instâncias que

[2] SPARQL é uma linguagem de consulta semântica, usada para recuperar dados de bases no formato RDF por exemplo. https://www.w3.org/TR/rdf-sparql-query/

[3] Wikidata Query Service - https://query.wikidata.org/

[4] W3C Recommentations SPARQL - https://www.w3.org/TR/sparql11-query/

os objetos pertencem. Assim, é possível buscar por um conjunto de valores de um metadado, e a partir das instâncias resultantes, identificar qual é mais representativa do metadado. Bem como, uma vez tipificado o metadado, buscar diretamente os seus valores na instância definida, e através das *labels* encontrar qual objeto potencialmente representa o valor na *Wikidata*.

A segunda etapa foi identificar quais as formas de acessar a da API (*Application Programming Interface*) SPARQL da *Wikidata* através da linguagem de programação Python, utilizando a consulta SPARQL como forma de requisitar resultados a partir de valores provenientes de um arquivo em formato tabular.

A linguagem Python foi escolhida pela habilidade de programação do pesquisador, e a forma encontrada para acessar a API SPARQL da *Wikidata* foi a biblioteca *requests*, que dentre outras funções, permite enviar parâmetros para um link (no caso o link de acesso da API), e recuperar os resultados em um formato manipulável. Desse modo, o script acessa a tabela de dados em CSV (*Comma Separeted Values*) a transformando em um *dataframe* através da biblioteca *pandas*, e para cada valor de cada coluna, uma consulta em SPARQL é efetuada. Utilizando o método *get*, da biblioteca *requests*, foram inseridos, a URL de acesso da API[5], os parâmetros de formato de dados em JSON (*JavaScript Object Notation*), e a query para cada valor da base dados.

A terceira etapa envolveu interpretar os resultados retornados pela API, e utilizá-los para indicar quais as melhores opções para representar cada um dos objetivos dos dois scripts, tipificar metadados como instâncias da *Wikidata* e reconciliar dados com seus respectivos objetos. Para isso, os resultados da busca retornados em JSON tiveram sua estrutura identificada, de forma que cada item do resultado aparece em uma lista obtida das chaves do JSON *'results'* e *'bindings'*, e cada item tem seu valor acessado pelo nome das variáveis da *query* SPARQL e pela chave 'value' (ex. 'sujeito' e 'value'). De posse dos valores, no caso do script de tipificação de metadados, para cada coluna da base (metadados) foram armazenados os objetos resultantes da Wikidata para cada valor, bem como a quais instâncias pertencem, e através da frequência em que uma instância é retornada para determinado conjunto de valores é possível indicar qual tem maior possibilidade de representar o metadado em questão.

No caso do script de reconciliação de dados, os resultados são obtidos do JSON com a mesma estrutura, e armazenados do mesmo modo, porém para cada valor da base de dados. E para avaliar qual o objeto da *Wikidata* corresponde melhor com o valor procurado, é utilizada a biblioteca *fuzzywuzzy*, que parte do preceito do *fuzzy- matching*, mesma abordagem utilizada pelo software OpenRefine[6], que compara conjuntos de caracteres e retorna um score de 0 a 100 para quantificar a proximidade entre os conjuntos, desse modo o valor em texto da base de dados é comparado com as *labels* do objeto na Wikidata e retornado com um score de 0 para muito diferente, e 100 para exatamente igual de comparação de texto.

Essas etapas metodológicas explicitam o desenvolvimento dos dois scripts de ligação de dados: a tipificação dos metadados a partir das colunas da base de dados (reconhecidas como metadados) e as instâncias dos objetos obtidos dos valores de cada coluna pesquisados na Wikidata; e a reconciliação de dados, realizada através da busca dos valores em texto por *labels* de objetos na Wikidata, podendo ou não essa busca ser reduzida ao escopo da instância representativa da coluna de valores (metadado).

Ambos os scripts preveem uma abordagem semiautomática onde o usuário, guiado pela aproximação textual com *score* oferecida pela técnica *fuzzy-match* ou pela frequência de aparição das instâncias, tem a possibilidade de escolher qual objeto/instância da Wikidata representa o valor/coluna da base de dados reconciliada.

[5] Url da API - https://query.wikidata.org/sparql

[6] Reconciliation OpenRefine - https://github.com/OpenRefine/OpenRefine/wiki/Reconciliation

4. Resultados

Os resultados deste estudo versam sobre o desenvolvimento dos scripts, portanto serão apresentadas e descritas partes dos scripts de tipificação de metadados e de reconciliação, ambos estão disponíveis no GitHub[7].

4.1 Script de Tipificação dos Metadados

O objetivo do script de tipificação de metadados (Apêndice A) é permitir identificar a quais instâncias da Wikidata os metadados da base de dados podem ser referenciados, e assim possibilitar refinar a reconciliação dos valores com objetos da Wikidata através de uma instância específica, reduzindo ruídos nos resultados.

As bibliotecas Python utilizadas foram, *requests*, para fazer a conexão com a API; *pandas* para processar a base de dados e armazenar os resultados; e a biblioteca *time*, para fazer com que o script aguarde um tempo entre as consultas, respeitando o limite de consultas da API.

Tanto a *query* SPARQL, quanto o processo de busca dos valores na API através da biblioteca *requests*, e a transformação dos resultados em um *dataframe* estão estruturados em uma função denominada *reconcilie_database*, essa função recebe um *dataframe* proveniente da base de dados, processa a pesquisa dos valores em SPARQL, e retorna outro *dataframe* com os resultados obtidos do JSON.

A query SPARQL utiliza as variáveis "sujeito" que faz referência ao objeto da Wikidata, "instancia" que se refere as instâncias dos objetos e "instanciaLabel" que retorna o nome das instâncias. Além disso, busca por objetos na Wikidata que tenham o valor em texto da base dados, e apresenta também as instâncias de cada possível objeto resultante. A sentença de filtro remove dos resultados objetos que são categoria da Wikidata e páginas de desambiguação.

Já os resultados, obtidos em JSON, foram iterados por item, e os valores das variáveis pesquisadas foram armazenados em um *dataframe*.

A etapa de apresentação dos resultados ainda está em implementação, mas espera-se calcular a frequência das instâncias obtidas para cada metadado, ordenando-as por número de aparição em uma lista de validação disposta ao usuário.

4.2 Script de Reconciliação de Dados

O script de reconciliação de dados (Apêndice B) utiliza as mesmas bibliotecas do script de tipificação de metadados, com adição da biblioteca *collections* que possui a classe *defaultdict* e permite criar dicionário de listas através de métodos, e também é utilizada a classe *fuzz* da biblioteca *fuzzywuzzy*, que permite aplicar métodos de comparação entre strings.

O que muda na query SPARQL desse script é que não se busca mais a instância do objeto, e as *labels* alternativas (*AltLabels*) do objeto são consideradas para os resultados. Desse modo, as variáveis utilizadas são, "sujeito" para recuperar o id do objeto na Wikidata, "sujeitoLabel" para recuperar sua *label* principal e "SujeitoAltLabel" para recuperar suas *labels* alternativas. A sentença de instância ainda permanece para que os resultados somente objetos que sejam de alguma instância da Wikidata, evitando assim páginas da Wikipedia.

A forma de requisição dos resultados também utiliza a biblioteca *requests*, que retorna o resultado em JSON. Os dados obtidos das *labels* alternativas são retornados no formato de lista, dessa forma, tanto a *label* principal, quanto as *labels* alternativas são unidas em uma lista para cada item pesquisado da base original em um dicionário de listas.

Após transformar os resultados em um dicionário, os valores são tratados para remover possíveis repetições.

Já a apresentação dos resultados, ainda em implementação, utilizará a classe *fuzz* da biblioteca *fuzzywuzzy* para comparar o valor da base de dados com os valores das *labels* retornadas na busca na Wikidata.

No *script* desenvolvido utiliza-se o método "token_set_ratio" da biblioteca *fuzzywuzzy*, que compara a ordenação dos caracteres de determinado texto e retorna um

[7] Scripts no GitHub - https://github.com/luisfeliperd/wikidata_sparql

score de 0 a 100 (0 para nenhuma semelhança entre os caracteres, e 100 para sequência de caracteres igual) para cada comparação. No caso a comparação é realizada entre o valor procurado na base de dados e o conjunto de *label* principal e *labels* alternativas resultantes da busca na Wikidata.

5. Considerações Finais

Os scripts ainda estão em desenvolvimento e fazem parte de um esforço de entender as possiblidades de reconciliação de dados com a Wikidata. Espera-se como resultados finais entender melhor a dinâmica de requisições a Wikidata, além da possiblidade de aplicar outras estratégias de aproximação de resultados que permitam ao usuário uma maneira semiautomática de ligar seus dados com a base de conhecimento da Wikidata.

As aplicações dessa iniciativa podem se estender para o aumento da semântica e enriquecimento de bases de dados, uma vez que os dados estarão reconciliados com a base de conhecimento da Wikidata, que possui conjuntos de informação específicos para cara objeto, e que podem ser recuperadas para aumentar o contexto dos dados.

Por fim, no contexto do projeto Tainacan, a reconciliação compõe uma etapa de ligação e enriquecimento de dados, que permitirá reconhecer os valores dos acervos com objetos digitais já existentes, permitindo coloca-los na rede de conhecimento disponível *on-line*, o que permite também potencializar uma busca integrada semântica entre os acervos.

Referências

DELPEUCH, A. A survey of OpenRefine reconciliation services. **arXiv preprint arXiv:1906.08092**, 2019.

MARTINS, D. L.; CARVALHO, J. M. C; GERMANI, L. Projeto Tainacan: Experimentos, Aprendizados e Descobertas da Cultura Digital no Universo dos Acervos das Instituições Memoriais. In: **TIC CULTURA Pesquisa Sobre o Uso das Tecnologias de Informação e Comunicação nos Equipamentos Culturais Brasileiros — 2018.** Comitê Gestor da Internet no Brasil, São Paulo, 2019.

SANDERSON, R. **"The Linked Data Snowball and Why We Need Reconciliation"**, 2016, Disponível em: https://www.slideshare.net/azaroth42/linked-data-snowball-or-why-we-need-reconciliation. Acesso em: 15/08/2019.

ZENG, M. L. "Semantic enrichment for enhancing LAM data and supporting digital humanities. Review article". **El profesional de la información**, v. 28, n. 1, 2019.

EXTRAÇÃO DE TÓPICOS APOIADA POR TÉCNICAS DE APRENDIZADO DE MÁQUINA EM REPOSITÓRIOS DIGITAIS: UM PRINCÍPIO PARA CONSTRUÇÃO SEMIAUTOMÁTICA DE ONTOLOGIAS

TOPICS EXTRACTION SUPPORTED BY MACHINE LEARNING TECHNIQUES IN DIGITAL REPOSITORIES: a principle for semi-automatic construction of ontologies

Fabio Piola Navarro[1],
José Eduardo Santarem Segundo [(2)]
(1) Universidade Estadual Paulista – Unesp, Marília/SP, navarro.fabio@gmail.com.
(2) Universidade de São Paulo – USP, Ribeirão Preto/SP, santarem@usp.br.

Resumo:

Ontologias provêm capacidade de leitura às máquinas, possibilidade de inferência melhorando consumo e geração de conteúdo em domínios específicos do conhecimento. A geração de documentos científicos tem crescido significativamente, a utilização de repositórios digitais para armazenamento e recuperação da informação acompanha esta demanda. Uma recuperação da informação (RI) assertiva nestes ambientes é importante para refletir a totalidade dos conteúdos, o que é inviável de ser realizado com estrita interação humana devido ao volume e a falta de ontologias para estes repositórios, necessitando de aportes da área de aprendizado de máquina. Topic modeling é uma técnica de machine learning, permite extrair tópicos dos documentos e para cada tópico palavras associadas. O objetivo deste trabalho é discutir a construção semiautomática de ontologias (*ontology learning*) pela utilização do modelo de tópicos LDA (Latent Dirichlet Allocation) para a geração tópicos. Para isso, utilizou-se de uma metodologia com característica tanto bibliográfica, quanto aplicada. Enquanto resultados, o topic modeling teve ótimo desempenho para geração de tópicos usados na construção de ontologias. Por fim, a aplicação das técnicas de machine learning contribui para fornecer bases a fim de obter amplo conhecimento sobre repositórios digitais.

Palavras-chave: Aprendizado de Ontologia; Modelo de Tópicos; Repositórios Digitais; Latent Dirichlet Allocation; Aprendizado de Máquina.

Abstract:
Ontologies provide machine readability, inference possibility improving consumption and content generation in specific domains of knowledge. The generation of scientific documents has grown significantly, the use of digital repositories for storing and retrieving information follows this demand. Assertive information retrieval (IR) in these environments is important to reflect the entire contents, which is not feasible to be performed with strict human interaction due to the volume and lack of ontologies for these repositories, requiring input from the learning area. machine. Topic modeling is a machine learning technique that allows you to extract topics from documents and associated words for each topic. The aim of this paper is to discuss the semiautomatic construction of ontology learning by using the Latent Dirichlet Allocation (LDA) topic model for topic generation. For this, we used a methodology with both bibliographic and applied characteristics. As a result, topic modeling performed well for generating topics used in the construction of ontologies. Finally, the application of machine learning techniques contributes to providing the foundation for a broad knowledge of digital repositories.

Keywords: Ontology Learning (OL); Topic Modeling; Digital Repositories; Latent Dirichlet Allocation; Machine Learning.

1. Introdução

Na Ciência da Informação, a recuperação da informação - RI tem sido cada vez mais estudada e atualmente sistemas computacionais têm um bom nível de busca sintática. No entanto, quando se trata de buscas semânticas, segundo [SANTARÉM SEGUNDO, 2010], estes repositórios de dados não estão aptos a retornar informação semântica e contextualizada.

A Web 1.0 ou Web Tradicional (anos 90) compreende os primórdios da Web, com documentos estáticos, pouca interação com o usuário que era somente um agente passivo e receptor de informação. Já a Web 2.0 (2000 a 2010) também conhecida como Web dos Documentos, já permite uma maior interação com o usuário que deixou de ser somente receptor de informação, mas também, participa e interage com os ambientes através de comentários, compartilhamento de informação e reviews.

Um dos problemas da Web 2.0 é a falta de padrão para escrita dos documentos, ou seja, formatos heterogêneos como (HTML, XML, entre outros) o que dificulta a leitura dos documentos por parte das máquinas, o que leva a pouca ou nenhuma geração de inferências, integrações, análises e talvez algum nível de inteligência.

A Web 3.0 ou ainda Web dos Dados (WoD), possui maior estruturação das informações e o uso de ontologias para prover vocabulário comum em ambientes permite fazer buscas (Recuperação da Informação) de maneira semântica. (Berners-Lee et al., 2001).

Ontologias tem caráter interdisciplinar, pois seu estudo e aplicações podem ser encontrados na Ciência da Informação (Vickery, 1997) e na Ciência da Computação (Guarino, 1998). Segundo (Gruber, 1995) ontologia é o estudo da semântica que pode ser analisada por uma especificação formal (leitura de máquinas) sobre uma conceituação compartilhada.

Segundo (Maedche and Staab, 2004) ontology learning é o processo de converter texto em ontologia e a capacidade de se gerar, de forma automática (sem interferência humana) ou ainda semiautomática (com alguma interferência humana), ontologias baseadas em termos de um determinado domínio.

Outro campo de estudo presente neste trabalho é o aprendizado de máquina ou ainda pelo termo em inglês, *machine learning*, partindo do teste de Turing com os primeiros softwares que conseguiam aprender a partir de dados de entrada.

Aprendizado de máquina é um subconjunto da inteligência artificial no qual programas são usados para aprender através deles mesmos a partir de dados e informação.

Segundo Mitchell (1997) o aprendizado de máquina trata da questão da construção de programas que automaticamente melhorem sua atuação com base na sua própria experiência, ou seja, os programas de computadores conseguem melhorar sua assertividade quanto mais são utilizados dentro de um determinado domínio de aplicação.

O método de pesquisa de tópicos (*topic modeling)* foi desenvolvido em 2003 por Blei que nomeou seu primeiro estudo na área como modelo LDA ou *Latent Dirichlet Allocation*. O LDA permite análise de estruturas temáticas não estritamente expostas o que possibilita extração de informação semântica para grandes volumes de texto que será base para extração de tópicos nos documentos de repositórios digitais. (BLEI, 2012).

Modelos de tópicos fazem uma busca por padrões nas relações entre documentos e termos (palavras), tentando identificar padrões significativos e possíveis relações. Este modelo pode trazer como resultado um conjunto de termos que podem possuir relação com um ou mais temas, ou ainda, ranquear documentos com maior relevância (Rani et al., 2017).

Extração de tópicos é uma tarefa definida dentro da área de aprendizado de máquina como sendo uma tarefa não supervisionada, ou seja, não há uma avaliação, nem uma rotulação prévia dos documentos em análise.

Os tópicos são estruturas com valor semântico que, no contexto de *text mining*, formam grupos de palavras que frequentemente ocorrem juntas. Esses grupos de palavras quando analisados, dão indícios a um tema ou assunto que ocorre em um subconjunto de documentos. A expressão “tópico” é usada levando-se em conta que o assunto tratado em uma coleção de documentos é extraído automaticamente, ou seja, tópico é definido como um conjunto de palavras que frequentemente ocorrem em documentos semanticamente relacionados. (DE et al., 2016)

O modelo de tópicos é muito mais que uma organização dos tópicos (palavras)

descobertas, este modelo faz uma análise de do conjunto de palavras que possuem maior frequência semântica dentro de cada documento.

Segundo (Xie et al., 2016) Repositórios digitais são sistemas de informação que recebem, armazenam, gerenciam, preservam e fornecem acesso a conteúdos digitais. Neste sentido, repositórios digitais temáticos são aqueles que, por exemplo, mantém todo fluxo informacional acima descrito para conteúdos na área da saúde.

Como os documentos estão armazenados em um corpus dentro de repositórios digitais, o modelo (LDA) permite a extração de palavras de cada documento e armazena estas palavras em tópicos. Portanto, cada documento pode ser visto como uma coleção de tópicos que podem ser analisados por uma tratativa computacional.

Neste cenário, a seção seguinte apresenta o objetivo geral deste artigo, ou seja, discutir a construção de ontologias de forma semiautomática através de modelos de aprendizado de máquina.

2. Objetivos

O objetivo geral deste trabalho é discutir a construção semiautomática de ontologias (*ontology learning*) apoiado pela utilização do modelo LDA (*Latent Dirichlet Allocation*) para recuperação da informação semântica em repositórios digitais temáticos.

De forma específica visa investigar como desenvolver a obtenção de tópicos dos documentos em um repositório digital temático para formar uma base para a construção de ontologias que serão utilizadas, ainda, discutir se estes tópicos descobertos podem ser utilizados como princípio para geração de uma ontologia de domínio para um repositório digital temático.

3. Procedimentos Metodológicos

Durante o desenvolvimento deste trabalho, aplicou-se uma metodologia exploratória e aplicada. Em um primeiro momento, realizou-se uma pesquisa bibliográfica e documental sobre as temáticas de *ontology learning*, *text mining*, recuperação da informação, aprendizado de máquinas e *topic modeling* em bases como SCOPUS e Google Scholar com os termos "ontology learning" para o primeiro com mais de oito mil resultados e para o segundo com mais de 1 milhão de resultados, mostrando ser um tema de grande procura e discussão.

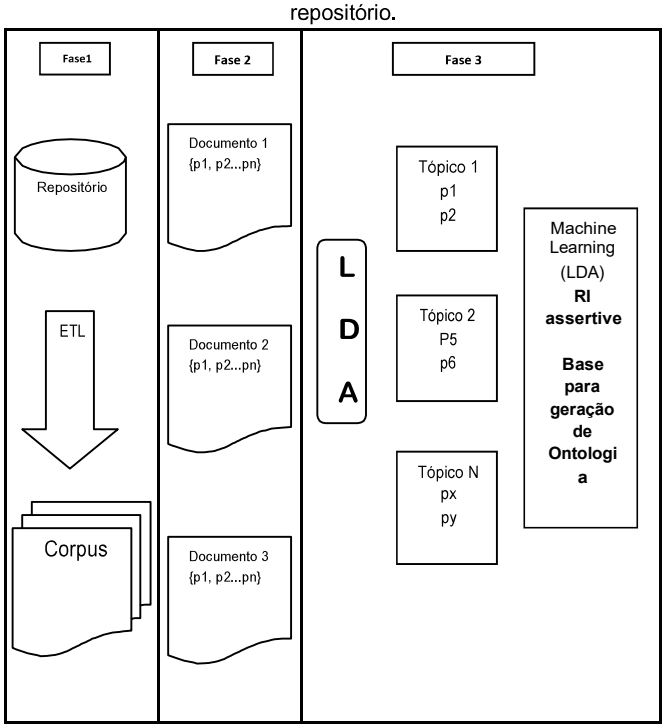
Posteriormente, realizou-se a discussão em que se vincula e reflete-se acerca das relações dessas temáticas, apresentando uma pesquisa aplicada, com uma análise documental para extração dos documentos com a técnica de *web scraping* e a construção de um ambiente computacional que possibilite carregar estes artigos e proporcione capacidade de análise, usando o método *topic modeling* a fim de comprovar a capacidade latente dos repositórios digitais em proporcionar informações para construção de uma ontologia para este repositório.

Em fase seguinte, o procedimento para a realização dessa prova de conceito, foi a transformação automática por meio de script dos arquivos de "PDF" dos documentos para "TXT". Essa mudança é necessária para que os algoritmos de análise possam avaliar cada documento.

Os procedimentos estão divididos em três fases, partindo da organização dos documentos (fase um), passando pela extração de palavras (fase dois) até a obtenção dos tópicos (fase três) que são base para a geração da ontologia. Essas fases estão apresentadas na figura 1.

Na primeira fase, aplica-se a técnica chamada de ETL (*Extract, Transform, Loading*), em que, primeiramente, é realizada uma extração dos arquivos do repositório, e posteriormente, após realizadas as devidas melhorias nos arquivos e transformações necessárias, faz-se o carregamento dos arquivos que gera como resultado o corpus ao qual será aplicado o algoritmo LDA.

Figura 1 – Fases do processo de aplicação do LDA no repositório.



Fonte: Elaborado pelos autores.

A fase 2 representa o momento em que os documentos já estão processados, contendo em cada documento centenas ou milhares de palavras. Esses documentos servem de base para proporcionar ao ambiente de programação a utilização plena do algoritmo de LDA, que posteriormente alimenta a fase 3.

Por fim, a fase 3 é onde de fato ocorre o processo do algoritmo LDA. A partir dos documentos o algoritmo faz uma leitura de cada palavra em cada documento e começa a fazer, em um primeiro momento, uma alocação de cada grupo de palavras para um determinado tópico.

Em um próximo passo, o algoritmo faz uma nova leitura abordando os seguintes aspectos:

a) Qual a proporção de uma determinada palavra no documento analisado, palavra esta que está atualmente atribuída a um determinado tópico?

b) Qual a proporção de atribuições para o tópico analisado, sobre todos os documentos, que venham da palavra em questão?

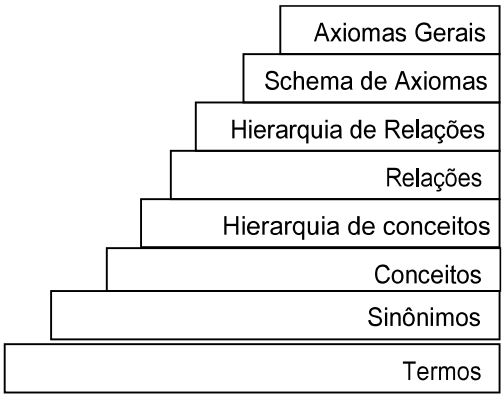
c) Analisando os passos a e b, reatribua a palavra analisada a um novo tópico baseado na probabilidade desta palavra pertencer a este novo tópico. Desta forma aloca palavras ou termos contextualizados que possuem aproximação semântica a um mesmo tópico por análise das proporções obtidas nos passos anteriores.

Como resultado deste processo, após a execução do algoritmo na fase 3, serão obtidos um conjunto de tópicos. Cada tópico gerado será composto de palavras que semanticamente possuem uma aproximação em relação a todo o corpus analisado.

Além disso, como cada documento é composto por uma mistura de um ou mais tópicos e cada tópico um conjunto de palavras, a busca não mais ficará baseada em palavras chave, mas sim a uma distribuição de tópicos com termos contextualizados que permitem a construção de sistemas de organização do conhecimento mais assertivos.

Segundo (Asim et al., 2018) as fases para obtenção de uma ontologia a partir de textos não estruturados, como no caso deste trabalho, são as seguintes:

Figura 2 – Bolo de camadas do aprendizado de ontologias



Fonte: Elaborado pelos autores, baseado em (Asim *et al.*, 2018)

Como pode ser analisado na Figura 2, o processo de construção de uma ontologia de domínio e de forma semiautomática começa extraindo termos e seus sinônimos do texto analisado. Em seguida, termos e sinônimos

correspondentes são combinados para formar conceitos. Em um passo posterior, são analisadas as possíveis relações taxonômicas e não taxonômicas entre esses conceitos. Finalmente, os esquemas dos axiomas são instanciados e os axiomas gerais são extraídos do texto não estruturado, todo esse processo é chamado de bolo de camada do aprendizado de ontologia.

3.1. Discussão sobre Técnicas

Como parte da discussão sobre as técnicas para aprendizado de ontologias ou ainda geração de ontologias de forma semiautomática, esta seção traz alguns trabalhos de destaque na área.

Um importante trabalho (Ding, 2002) sumarizou algumas características presentes em pesquisas sobre aprendizado semiautomático de ontologias, destas os seguintes destaques são trazidos: (i) a maioria dos sistemas de aprendizado de ontologia aprendiam a partir de outra ontologia e não a partir do zero (ii) técnicas de extração de tópicos se revelaram promissoras para construção da base de tópicos e (iii) identificar a relação entre os tópicos se consistia no maior desafio para geração das ontologias.

As técnicas para geração de ontologias se apoiam em campos multidisciplinares como a ciência da informação com representação do conhecimento e recuperação da informação, o aprendizado de máquinas da ciência da computação e o processamento de linguagem natural, entre outros. Para (Asim et al., 2018) estas técnicas estão classificadas em 3 categorias: linguísticas, estatísticas e programação lógica indutiva.

As técnicas linguísticas são usadas em todas as camadas do bolo de geração de ontologias, desempenham papel crucial, pois atuam propriamente na linguagem textual obtida em cada documento de um repositório digital e estão relacionadas desde pré processamento dos dados até a análise de termos, conceitos e extração de relações.

Técnicas estatísticas têm como característica principal os conceitos da própria estatística como o uso de probabilidades e não de análises semânticas. Mesmo assim, são usadas em fases de extração de termos, conceitos e relação taxonômica.

Já a Programação Lógica Indutiva (PLI) é uma disciplina da área de aprendizado de máquina, que se baseia em uma miríade de premissas sobre um conhecimento prévio e infere uma representação uniforme sobre determinado este conhecimento. Dentro das fases de geração da ontologia a PLI é utilizada no estágio mais alto do bolo de camadas para gerar os axiomas gerais.

4. Resultados

A utilização do LDA como método de *topic modeling* para extração dos tópicos de cada documento dentro de um repositório digital teve grande desempenho e acurácia, pois não houve nenhum documento que não pode ser extraído os tópicos.

Neste sentido, como a geração dos tópicos é estrutura básica para construção de ontologias esta etapa revelou-se suficiente para o aprendizado de ontologia satisfazer as primeiras camadas do bolo de aprendizado de geração de ontologias.

Sobre a discussão das técnicas tanto a técnica linguística quanto a programação lógico indutiva e também a estatística revelaram-se estritamente necessárias para se atingir todas as camadas do bolo de aprendizado de ontologias, pois como o LDA consegue atingir as 3 primeiras (termos, sinônimos e conceitos) as relações podem ser obtidas pela estatística e as últimas usando a técnica de programação lógico indutiva.

5. Conclusão ou Considerações Finais

A criação de uma ontologia é uma tarefa complexa que necessita de evoluções nas quais a interação humana é decisiva. No entanto, este trabalho tem a intenção de promover uma discussão sobre meios automatizados que possam auxiliar a geração de ontologias de domínio e para repositórios digitais, de maneira a diminuir a interação humana a fim de promover uma maior facilidade para criação de ontologias de domínio.

A Ciência da Informação traz todo o arcabouço necessário de teorias e técnicas para a construção de instrumentos que por

meio de análises de dados, possam favorecer os usuários para que possam recuperar informação que necessitam com maior precisão e semântica. O cenário apresentado apresenta múltiplas oportunidades de aprimorar a pesquisa daqueles que interagem com os ambientes informacionais digitais, no tocante à construção de ferramentais que possam melhorar as formas de como as informações são representadas e recuperadas.

Neste contexto, este trabalho discorre sobre como técnicas de machine learning, quando aplicado a cenários de grandes conjuntos de dados acadêmicos (repositórios digitais), podem auxiliar a recuperação da informação nestes ambientes tornando-os mais precisos e abrangentes quanto ao contexto das informações armazenadas.

O modelo construído demonstra um possível meio de relacionar diferentes campos de estudos, apresentando uma forma de utilizar o conteúdo dos repositórios, com o uso de algoritmos de topic modeling e LDA para a extração e sinônimos dos termos. A partir dessa aplicação dos algoritmos nos repositórios digitais, apresenta-se um possível modo de aprimorar a geração de ontologias de forma semiautomática.

O trabalho apresenta um caso em que o modelo proposto é aplicado, no qual se utiliza *machine learning* e *topic modeling* em um repositório digital real. O que visou demonstrar a viabilidade dessa proposta, mostrando o relacionamento entre *topic modeling*, *machine learning* e geração semiautomática de ontologias em repositórios digitais para viabilizar uma estrutura que possibilite recuperação da informação semântica e desta maneira mais abrangente e assertiva.

Finalmente, este trabalho demonstra um caminho inicial para a área em discussão e serve como base de pesquisa para um trabalho futuro (em andamento) no qual se pretende de fato, gerar ontologias sobre o repositório digital analisado, e discutir as técnicas envolvidas, os algoritmos utilizados e a construção de um modelo para reutilização.

Referências

BERNERS-LEE, T., HENDLER, J. AND LASSILA, O., 2001. The semantic web. **Scientific American**. 284(5), 28-37.

VICKERY, B. C. Ontologies. **Journal of Information Science**, London, v. 23, n. 4, p. 227-286, 1997.

GUARINO, N. Formal ontology and information systems. In N. Guarino (Ed.). **Formal ontology in Information Systems** (1998, pp. 3–15). Amsterdam: IOS Press.

GRUBER, Methods for Ontology Development. In: **Semantic Web: Concepts, Technologies and Applications**.

MAEDCHE, A., & STAAB, S., 2004. Ontology learning. **In Handbook on ontologies**, Springer Berlin Heidelberg, pp. 173-190.

MITCHELL, T. M. Machine Learning. Disponível em: <https://www.cs.ubbcluj.ro/~gabis/ml/ml-books/McGrawHill - **Machine Learning** - Tom Mitchell.pdf>. Acesso em: 12 set. 2019.

BLEI, D. M. Introduction to Probabilistic Topic Modeling. **Communications of the ACM**, 2012.

DE, T. et al. universidade de são paulo modelos probabilísticos de tópicos: desvendando o latent dirichlet allocation. **relatórios técnicos**. 2016.

RANI, M.; DHAR, A. K.; VYAS, O. P. Semi-automatic terminology ontology learning based on topic modeling. **Engineering Applications of Artificial Intelligence**, v. 63, n. August, p. 108–125, 2017

ASIM, M. N. et al. A survey of ontology learning techniques and applications. **Database**, v. 2018, n. 2018, p. 1–24, 2018.

Ding,Y. and Foo,S. (2002) Ontology research and development. **part 2-a review of ontology mapping and evolving. J. Inf. Sci.**, 28, 375–38

XIE, I. et al. Digital preservation. **Discover Digital Libraries**, p. 255–279, 1 jan. 2016.

SANTAREM SEGUNDO, José Eduardo **Representação Iterativa: um modelo para Repositórios Digitais** Tese de Doutorado em Ciência da Informação. – Universidade Estadual Paulista UNESP Marília, 2010

FUSÃO DE DADOS PARA COMPREENSÃO DE FENÔMENOS AMBIENTAIS POR MEIO DE FOTOGRAFIAS

DATA FUSION FOR UNDERSTANDING ENVIRONMENTAL PHENOMENA BY PHOTOGRAPHS

Danilo Camargo Dias[1], Danilo Dolci [2], Isaque Katahira [3], José Eduardo Santarém Segundo [4], Leonardo Castro Botega [5], Mariângela Spotti Lopes Fujita [6]

(1, 2, 3, 4, 5, 6) Programa de Pós-Graduação em Ciência da Informação da UNESP – Campus de Marília, Av. Hygino Muzzi Filho, 737, Bairro Mirante, Marília-SP, 17525-900, E.mail. danilo.dias@etec.sp.gov.br, danilo.dolci@fatec.sp.gov.br, isaque.katahira@fatec.sp.gov.br, santarem@marilia.unesp.br, leonardo.botega@unesp.br, mariangela.fujita@unesp.br

Resumo:
Com o crescimento dos repositórios digitais imagéticos, as técnicas de fusão de dados têm conquistado uma importância cada vez maior, devido à extensa quantidade e à heterogeneidade entre as fontes de dados. Esta pesquisa apresenta as potencialidades da fusão de dados em *datasets* de imagens de satélites da NASA que monitoram a qualidade do ar, temperatura dos oceanos, derretimento das calotas polares e emissão de CO2, com objetivo de estabelecer correlação entre a temperatura média do planeta e esses fenômenos, gerando uma visão global da qualidade ambiental do planeta. É realizada extração de metadados com imagens obtidas com ferramentas da NASA: *National Snow & Ice Data Center e a GES DISC Data Archive*. Após a seleção e identificação das imagens da amostra, a organização da fusão de dados foi realizada para verificação dos metadados que uma base e outra utilizam, a exploração da estrutura dos metadados, quantidade, descrição, tipo, entre outros que geraram 19 metadados potenciais convergentes. Posteriormente, esse estudo subsidiará a fusão de dados, especialmente, por meio de técnicas de redes neurais devido a sua ampla eficiência em análise imagéticas.
Palavras-chave: Fusão de dados; Metadados de imagens; Redes neurais; Fotografias.

Abstract:
With the growth of digital imagery repositories, data fusion techniques have become increasingly important due to the large amount and heterogeneity between data sources. This research presents the potential of data fusion in NASA satellite image datasets that monitor air quality, ocean temperature, polar ice caps and CO2 emissions, with the aim of establishing a correlation between the average temperature of the planet and these phenomena, generating a global view of the environmental quality of the planet. Metadata extraction is performed with NASA tools: National Snow & Ice Data Center and GES DISC Data Archive. After selecting and identifying the sample images, the data fusion was organized to verify the metadata that one base and another use, the exploration of the metadata structure, quantity, description, type, among others that generated 19 converging potential metadata. Subsequently, this study will subsidize data fusion, especially through neural network techniques due to its broad efficiency in image analysis.
Keywords: Data fusion; Image metadata; Neural networks; Photographs.

1. Introdução

Com o crescimento dos repositórios digitais imagéticos, as técnicas de fusão de dados têm conquistado uma importância cada vez maior. A intensidade e a heterogeneidade com que arquivos, especialmente os imagéticos, são produzidos, traz novos e grandes desafios, pois à medida que os dados gerados se multiplicam, há também a necessidade crescente de criar estratégias para sua organização e recuperação. Uma forma de aprimorar a descrição e a identificação de documentos imagéticos é a utilização de metadados.

O indexador deve considerar tanto a rapidez quanto a precisão proporcionada pelos metadados catalogados. No entanto, observando diferentes fontes, percebe-se que nem sempre os metadados são convergentes, fato que pode trazer dúvidas aos usuários ou explicitar lacunas de observação. Nesse sentido, a fusão de dados se configura como uma estratégia importante para extração dos parâmetros mais representativos. (HALL; JORDAN, 2010; BOTEGA, 2016).

É notório que as últimas décadas presenciaram o surgimento de novas aplicações para a tecnologia da informação, principalmente no que tange à organização de

dados dos repositórios de imagens digitais. Somado a isso, o compartilhamento de documentos aumentou, principalmente, quando se trata de fotografias. No entanto, segundo Oliveira e Vital (2015), não há na literatura científica da área da Ciência da Informação um acordo sobre como as imagens devem ser tratadas, pois diversos fatores implicam nessa análise, incluindo o tipo de imagem (pinturas, fotografias etc.) e o suporte no qual é divulgada.

Observando as potencialidades dos novos recursos disponibilizados pelas ferramentas computacionais, a hipótese que orientará a investigação realizada é a de que o registro de um fenômeno feito a partir de diversos critérios colabora para a sua compreensão mais global. Assim, a fusão de dados se mostraria relevante para obtenção de um resultado superior de recuperação, pois permitiria a comparação, a extração e o correlacionamento de dados sob diferentes perspectivas.

2. Objetivos

O objetivo desta pesquisa é demonstrar as potencialidades da fusão de dados obtidos a partir das imagens registradas por diferentes satélites. As imagens fotografadas selecionadas são resultadas do monitoramento da qualidade do ar, da temperatura dos oceanos e do derretimento das calotas polares.

3. Procedimentos metodológicos

Para alcançar os objetivos pretendidos, inicialmente, realiza-se revisão bibliográfica exploratória sobre padrões de metadados e fusão de dados imagéticos. Sequencialmente, consulta-se dois sistemas de monitoramento da NASA: *National Snow & Ice Data Center* e o *GES DISC Data Archive*, especificamente com relação aos metadados catalogados sobre as imagens de interesse, para, por fim, propor a fusão dos dados encontrados com base em redes neurais, a fim de melhor compreender os efeitos dos altos índices de emissão de CO2 para a qualidade de vida no planeta.

4. Resultados

4.1 Indexação de fotografias e fusão de dados

O uso de metadados e descritores em imagens fotográficas se mostra fundamental, pois existem informações significativas nesse tipo de conteúdo que podem ser pedidas, caso os processos de catalogação e indexação não sejam realizados de acordo com critérios validados.

Na esfera científica, as fotografias obtidas pelo sensoriamento remoto se materializam como valiosas fontes de informação para pesquisas relacionadas ao clima do planeta (ZANOTTA; FERREIRA; ZORTEA, 2019).

Segundo Zanotta *et al*. (2019), a crescente facilidade em se conseguir imagens em alta resolução da superfície terrestre provenientes de satélites possibilitou uma aproximação inovadora entre a tecnologia e a sociedade.

Chino, Romani e Traina (2010) afirmam que, como os dados são gerados por diversos tipos de satélites, com imagens registradas em resolução e periodicidades diferentes, é de suma importância investir na integração das informações obtidas.

Segundo Botega (2016), fusão de dados ou informações é a rotina de transformação de dados ou informações para produzir estimativas e predições de estados de entidades, visando maximizar o valor da informação e estimular a consciência situacional de analistas sobre um ambiente de interesse.

A título de exemplo, o trabalho de Srivastava *et al*. (2019) busca detectar áreas com alta probabilidade do incêndio florestal em uma região da Índia, com base na diferença da *Normalized Burn Ratio* (NBR), entre as condições pré e pós-incêndio. O estudo compara *datasets* gerados pelo satélite LANDSAT TM 5 por dois modelos de captura, o *Geographical Information Systems* (GIS) e o *Earth Observation* (EO). O resultado é a avaliação multicritério incorporando atributos como fontes antropogênicas e naturais, de modo a fundir os modelos e realizar a predição de zonas de alta probabilidade de incêndio.

As bases da modelagem computacional das Redes Neurais Artificiais (RNA) foram concebidas inicialmente por McCulloch e Pitts (1943) que desenvolveram estudos detalhados sobre a lógica das redes neurais, além de Von Neumann (1956) e Winograd e

Cowan (1963), que também trouxeram importantes contribuições para criar redes confiáveis com o uso de redundância. (HERTZ, 2018).

Meio século depois de McCulloch e Pitts, já com o uso de computadores que possibilitam aplicações práticas para as redes neurais, alguns autores elaboraram definições. Nigrin (1993) definiu uma rede neural artificial como um circuito composto por uma grande quantidade de unidades simples de processamento, inspiradas no sistema neural. Haykin (1994), analisando a estrutura não linear do cérebro, conceituou as RNA como sistemas massivamente paralelos e distribuídos, compostos por unidades de processamento simples que possuem uma capacidade natural de armazenar e utilizar conhecimento. Mais tarde, Haykin (2007) definiu uma RNA como uma rede implementada usando componentes eletrônicos ou simulada em software, que é projetada para modelar o caminho em que o cérebro executa uma tarefa ou função específica de interesse.

4.2 Sistemas de monitoramento e metadados catalogados

Com o intuito de confirmar a fusão de dados como ferramenta capaz de reunir, selecionar e evidenciar dados relevantes, é realizada a consulta e seleção de imagens e metadados de dois sistemas de monitoramento da NASA (*National Snow & Ice Data Center* e o *GES DISC Data Archive*). A partir dos dados selecionados, evidencia-se as potencialidades da fusão dos dados encontrados, a fim de melhor compreender como os altos índices de emissão de CO2 reverberam a qualidade de vida no planeta.

As ferramentas *National Snow & Ice Data Center* e *GES DISC Data Archive* foram acessadas pela área de dados abertos da *National Aeronautics and Space Administration* (NASA) a fim de extrair metadados de imagens obtidas por satélite que monitoram o derretimento das calotas polares, a qualidade do ar e a temperatura dos oceanos. Diversas informações são geradas e compactadas num arquivo que fica disponível na área do usuário, conforme Figura 1 (Apêndice A). Os campos que compõem a catalogação são: identificador do *datacenter*, data de registro, data da última alteração; nome do arquivo compartilhado; tamanho (MB); data e hora da captura; espaço de tempo inicial da captura; espaço de tempo final da captura; coordenadas da área observada (área retangular - norte, sul, leste e oeste); nome da plataforma; instrumento de sensoriamento utilizado; nome da campanha da NASA e identificador da aeronave utilizada. O *GES DISC Data Archive* é o serviço que fornece dados, informações e serviços de ciências da Terra para pesquisadores, cientistas de dados, usuários de aplicativos e estudantes, incluindo: composição atmosférica, ciclos de água e energia e variabilidade climática. É possível observar que também são arquivados conjuntos de dados aplicáveis ao Ciclo de Carbono e ao Ecossistema. Assim, há diversos *datasets* sobre a situação climática do planeta. O sistema gera e renderiza imagens e metadados sobre regiões em diversos padrões quanto à emissão de CO2, a temperatura dos oceanos e a temperatura atmosférica (Vide Figura 2 no Apêndice B).

Sobre a Figura 2 os metadados relacionados na catalogação são: data de registro; data da última alteração; nome da coleção a que pertence; indicador da coleção; tamanho (MB); período (diurno/ noturno); data e hora da captura; espaço de tempo inicial da captura; espaço de tempo final da captura; coordenadas da área observada (área retangular - norte, sul, leste e oeste); parâmetros medidos; parâmetro de qualidade automatizada e URL de acesso online ao recurso. Nesta pesquisa optou-se pelo padrão Nativo, que permitiu a identificação dos metadados.

No que tange à temperatura dos oceanos, selecionou-se, a título de exemplo, a Figura 3 (Apêndice C), cujos metadados relacionados na catalogação são: data de registro; data da última alteração; nome da coleção a que pertence; indicador da coleção; tamanho (MB); tipo de camada atmosférica observada; período (diurno/ noturno); data e hora da captura; espaço de tempo inicial da captura; espaço de tempo final da captura; domínio espacial vertical (máximo e mínimo); coordenadas da área observada (área retangular - norte, sul, leste e oeste); URL de acesso online ao recurso, descrição; tipo de

dado (aberto ou não); *mime type* e navegação em imagens associadas (URL, tamanho e descrição).

A última imagem selecionada se relaciona com a temperatura atmosférica (Vide Apêndice D). Para descrição da Figura 4, são fixados os metadados: data de registro; data da última alteração; nome da coleção a que pertence; indicador da coleção; tamanho (MB); tipo de camada atmosférica observada; período (diurno/ noturno); data e hora da captura; espaço de tempo inicial da captura; espaço de tempo final da captura; domínio espacial vertical (máximo e mínimo); coordenadas da área observada (área retangular - norte, sul, leste e oeste); URL de acesso online ao recurso, descrição; tipo de dado (aberto ou não) e *mime type.*

Após a seleção e identificação das imagens da amostra, a organização da fusão de dados foi realizada para verificação dos metadados que uma base e outra utilizam, ou seja, exploração da estrutura dos metadados, quantidade, descrição, tipo, entre outros; foram selecionados 19 metadados potenciais para a fusão de dados, conforme disposto no Quadro 1:

Quadro 1: Proposta: Visão global da qualidade do planeta Terra

National Snow & Ice Data Center e a GES DISC Data Archive
Data de registro
Data da última alteração
Nome da coleção a que pertence
Identificador da coleção
Tamanho (MB)
Período (Diurno/Noturno)
Data hora da captura
Espaço de tempo inicial da captura
Espaço de tempo final da captura
Coordenadas da área observada (Área retangular – Extremo Norte Global Fixo, Extremo Sul Global Fixo, Leste Variável e Oeste Variável)
Quantidade de CO2 registrado
Camada atmosférica onde foi registrada a medição da temperatura dos oceanos
Registro da temperatura média atmosférica continental
Dimensão das calotas polares registradas na faixa capturada
Correlação entre as médias das temperaturas aferidas
Correlação da quantidade de CO2 e a temperatura média atmosférica
Correlação da dimensão das calotas polares e a temperatura média atmosférica
URL de acesso aos recursos imagéticos

Fonte: Elaborado pelo Autor, com base nos dados obtidos nas ferramentas *National Snow & Ice Data Center e a GES DISC Data Archive, 2019*.

Com base no Quadro 1, é possível inserir os dados em um determinado algoritmo de Inteligência Artificial (IA) (OSÓRIO; BITTENCOURT, 2000) que faça a leitura e a interpretação dos dados, em um determinado período de tempo, e posteriormente, estabeleça a correlação existente entre eles. Para tanto, uma técnica potencial é a de redes neurais, que são amplamente utilizadas em softwares de análise de imagens (GONÇALVES, et al., 2008; MOREIRA, et al., 2002.). Destaca-se o Modelo de Hopfield de RNA para analisar as imagens obtidas por satélites, por apresentar os conceitos de redes com realimentação e comportamento dinâmico. (HAYKIN, 2007).

Uma outra proposta é a utilização do algoritmo de MAXVER, ou Máxima Verossimilhança, que, pela classificação supervisionada, "considera a ponderação das distâncias entre médias dos níveis digitais das classes e o pixel, utilizando parâmetros estatísticos, isto é, considerando a distribuição de probabilidade normal para cada classe." (RIBEIRO, 2001).

Nesse sentido, o algoritmo de IA especialista analisa a poluição do ar e o derretimento das calotas polares, estabelecendo, dessa forma, uma correlação com a elevação da temperatura global; tal procedimento acrescenta novas informações aos metadados das imagens, uma série de análises automatizadas da situação do

planeta com base na fusão dos dados (que estão separados e armazenados nos metadados) das imagens.

5. Considerações Finais

Diante das imagens e dos metadados apresentados com foco no monitoramento de grandes áreas, pela proposta de uso de Redes Neurais Artificiais, foi possível ampliar a compreensão e a avaliação de alterações ambientais (tanto positivas quanto negativas), ao correlacionar, por exemplo, os índices de emissão de CO2 à temperatura média do planeta e ao derretimento das calotas polares. Considerando as potencialidades dos novos recursos disponibilizados pelas ferramentas computacionais, foi possível comprovar que o registro de um fenômeno feito a partir de diversos critérios colabora para a sua compreensão mais global, como proposto pelas RNA.

Por fim, convém destacar que a fusão de dados imagéticos e os novos métodos de seleção se destacam no cenário atual como estratégias eficientes de representação e recuperação da informação. Para trabalhos futuros, o método poderia ser testado em um *corpus* maior de imagens, bem como em outras áreas do conhecimento, de modo a ampliar a validação do comparativo. Desse modo, é notória a necessidade de novas pesquisas voltadas para a organização e seleção de dados representativos em meio a avalanche de imagens produzidas atualmente.

Referências

BOTEGA, Leonardo Castro *et al*. **Modelo de fusão dirigido por humanos e ciente de qualidade de informação.** 2016. Disponível em: https://aberto.univem.edu.br/handle/11077/1483. Acesso em: 17 set. 2019.

CHINO, Daniel YT; ROMANI, Luciana AS; TRAINA, Agma JM. Construindo séries temporais de imagens de satélite para sumarização de dados climáticos e monitoramento de safras agrícolas. **Revista Eletrônica de Iniciação Científica**, v. 10, n. 3, p. 1-20, 2010. Disponível em:. Acesso em: 17 set. 2019.

GONÇALVES, Márcio Leandro et al. Classificação não-supervisionada de imagens de sensores remotos utilizando redes neurais auto-organizáveis e métodos de agrupamentos hierárquicos. **Revista brasileira de cartografia**, v. 1, n. 60, 2008.

HALL, D.; JORDAN, J. **Human-centered information fusion**. [S.l.]: Artech House, 2010.

HAYKIN, Simon. **Redes neurais:** princípios e prática. Porto Alegre: Bookman, 2007.

MCCULLOCH, Warren S.; PITTS, Walter. A logical calculus of the ideas immanent in nervous activity. The bulletin of mathematical biophysics, v. 5, n. 4, p. 115-133, 1943.

MOREIRA, Fabiano Cordeiro *et al*. Reconhecimento e classificação de padrões de imagens de núcleos de linfócitos do sangue periférico humano com a utilização de redes neurais artificiais. 2002. Disponível em: https://repositorio.ufsc.br/handle/123456789/8230**6.** Acesso em: 17 de set. 2019.

NIGRIN, Albert. Neural networks for pattern recognition. **MIT press**, 1993.

OLIVEIRA, R. A.; VITAL, L. P. Análise e indexação de imagens na rede Flickr. **Em Questão**, Porto Alegre, v. 21, n. 2, p. 7-30, mai/ago. 2015. Disponível em: <https://seer.ufrgs.br/EmQuestao/article/view/50968/33977>. Acesso em 1 de ago. 2019.

OSÓRIO, Fernando S.; BITTENCOURT, João R. Sistemas Inteligentes baseados em redes neurais artificiais aplicados ao processamento de imagens. In: I WORKSHOP DE INTELIGÊNCIA ARTIFICIAL UNISC–Universidade de Santa Cruz do Sul Departamento de Informática-Junho. 2000.

RIBEIRO, Selma Regina Aranha; CENTENO, Jorge Silva. Classificação do uso do solo utilizando redes neurais e o algoritmo MAXVER. **Simpósio Brasileiro de Sensoriamento Remoto**, v. 20, 2001.

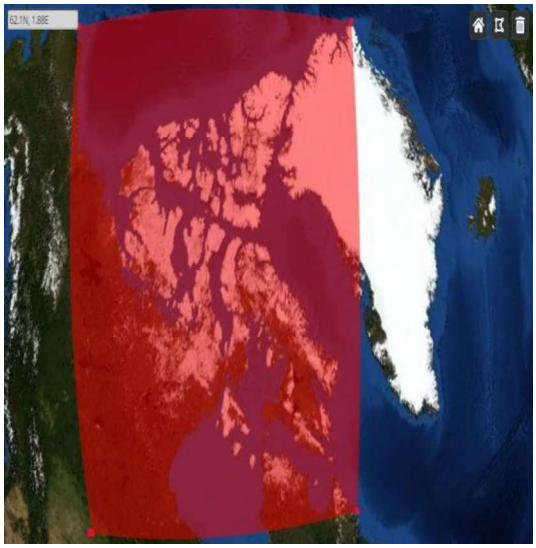
SRIVASTAVA, Prashant K. *et al.* Deriving forest fire probability maps from the fusion of visible/infrared satellite data and geospatial data mining. **Modeling Earth Systems and Environment**, v. 5, n. 2, p. 627-643, 2019.

VON NEUMANN, John. Probabilistic logics and the synthesis of reliable organisms from unreliable components. **Automata studies**, v. 34, p. 43-98, 1956.

WINOGRAD, Shmuel; COWAN, Jack D. **Reliable computation in the presence of noise**. Cambridge, Mass.: Mit Press, 1963.

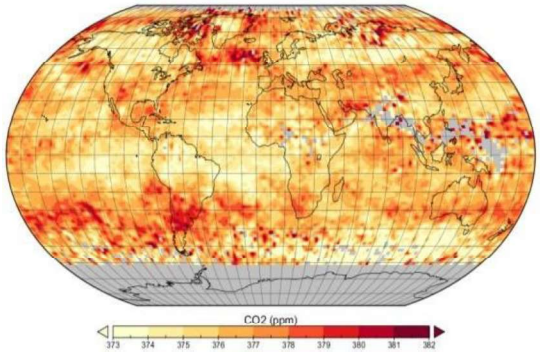
ZANOTTA, Daniel Capella; FERREIRA, Matheus Pinheiro; ZORTEA, Maciel. **Processamento de imagens de satélite**. Oficina de Textos, 2019.

Apêndice A – Figura 1: Arquivo gerado a partir da ferramenta da *National Snow & Ice Data Center*



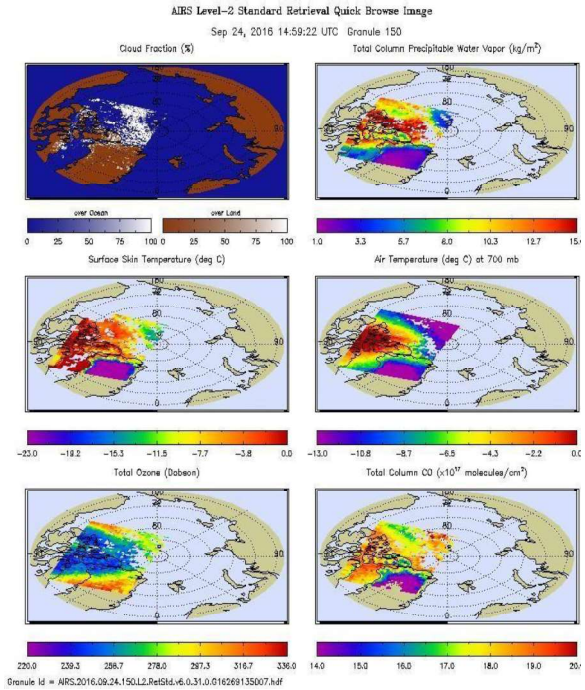
A área destacada em vermelho é a localização da qual foram extraídos os metadados.
Fonte: *National Snow & Ice Data Center*, 2019.

Apêndice B – Figura 2 - Dióxido de carbono na troposfera livre



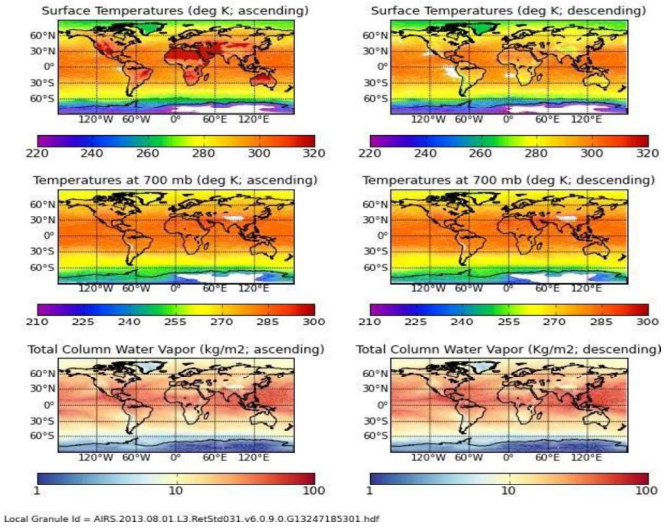
Fonte:https://docserver.gesdisc.eosdisc.nasa.gov/public/project/Images/AIRS3C28_005.png, 2019

Apêndice C – Figura 3 – Temperatura dos oceanos



Fonte:
https://airsl2.gesdisc.eosdis.nasa.gov/data/Aqua_AIRS_Level2/AIRX2RET.006/2016/268/AIRS.2016.09.24.150.L2.RetStd.v6.0.31.0.G16269135007.hdf.jpg, 2019.

Apêndice D - Figura 4 - Temperatura atmosférica



Fonte:
https://docserver.gesdisc.eosdisc.nasa.gov/public/project/Images/AIRH3SPM_006.png, 2019.

GOOGLE DATASET SEARCH (BETA): VISÃO GERAL E PERSPECTIVAS PARA INDEXAÇÃO E DISPONIBILIZAÇÃO DE CONJUNTOS DE DADOS CIENTÍFICOS ABERTOS

GOOGLE DATASET SEARCH (BETA): OVERVIEW AND PERSPECTIVES FOR INDEXING AND AVAILABILITY OF OPEN SCIENTIFIC DATASETS

Eduardo Diniz Amaral[1], Adilson Luiz Pinto[(2)]
(1) Universidade Estadual de Montes Claros - UNIMONTES, Av. Prof. Rui Braga, S/N - Vila Mauriceia, Montes Claros - MG, eduardo.diniz@unimontes.br.
(2) Universidade Federal de Santa Catarina – UFSC, R. Eng. Agronômico Andrei Cristian Ferreira, s/n - Trindade, Florianópolis - SC, adilson.pinto@ufsc.br.

Resumo: Este artigo tem por objetivo obter uma visão geral, funcionamento, padrões e perspectivas sobre a ferramenta *Google Dataset Search* – ferramenta lançada em 2018 que se propõe a identificar, indexar e disponibilizar datasets (conjuntos de dados) disponíveis na internet. A metodologia utilizada foi mapeamento de bibliografias sobre o tema, testes práticos, análise de buscas e impressões sobre as perspectivas. Foi realizado portanto levantamento bibliográfico sobre a plataforma, identificação do funcionamento interno, padrões e diretrizes utilizadas, além da explicação geral sobre os formatos e instituições de padronização que norteiam a plataforma. Foram feitos testes práticos e específicos na ferramenta. Os resultados obtidos mostraram uma plataforma que, apesar da necessidade de ajustes nos algoritmos de busca e indexação de metadados, parece ser bastante promissora para a comunidade científica, sobretudo na disponibilização das informações geradas na cauda longa de pesquisa, alinhada com padrões internacionais de interoperabilidade de dados, com poucos datasets em português indexados. Observou-se que as maiores iniciativas nacionais trabalham com os mesmos padrões de metadados desta ferramenta. Conclui-se que uma ferramenta com grande capacidade de armazenamento, catalogação, indexação, interoperação e disponibilização dos conteúdos disponíveis na internet, criada e mantida tecnológica e financeiramente pela empresa Google tem bastante potencial.

Palavras-chave: conjuntos de dados; datasets; acesso aberto; padrões de metadados; google dataset search

Abstract: This article aims to get an overview, operation, patterns and perspectives on the Google Dataset Search, launched in 2018, that aims to identify, index and make available datasets around the internet. The methodology used was mapping bibliographies about the subject, practical tests, search analysis impressions and perspectives about the future of this tool. Therefore, a bibliographic survey about the platform, identification of the internal functioning, standards and guidelines used, as well as a general explanation about the formats and standardization institutions that guide the platform were performed. Practical and specific tests were made on the tool. The results obtained showed a platform that, despite the need to adjust the metadata search and indexing algorithms, seems to be very promising for the scientific community, especially in the availability of the information generated in the long tail of the search, aligned with international standards of interoperability of data, with few Portuguese datasets indexed. It was observed that the largest national initiatives work with the same metadata standards of this tool, indicating future integrations. Good prospects for a tool with large storage capacity, cataloging, indexing, interoperation and availability of content on the Internet, created and maintained technologically and financially by the company Google.

Keywords: datasets; open access; scientific information; metadata patterns; google dataset search

1. Introdução

De acordo com Gavron (2017), o acesso aberto à informação científica exerce importante influência no desenvolvimento da ciência, pois, por meio desta torna-se possível conhecer o que está sendo realizado pelos pesquisadores em todas as partes do globo. Quanto mais atualizada, mais relevante será para os pesquisadores, promovendo um melhor diálogo e intercâmbio informacional entre eles. Neste contexto, os *datasets* (ou conjunto de dados) representam alta relevância. Trata-se de coleções brutas de dados organizados sobre um tema ou contexto específico, geralmente dispostos em colunas (como atributos) e linhas como dados individuais, elementos ou unidades, nos mais diversos formatos (planilhas, arquivos texto, listas, tabelas etc). No que tange o universo científico, os conjuntos de dados coletados, organizados e armazenados em

experimentos, por exemplo, são fundamentais para replicação, comprovação e novas análises destes. São informações valiosas que, se bem trabalhadas, podem gerar novos caminhos para pesquisas científicas.

Existem disponíveis na internet milhares de repositório de dados, provendo acesso a milhões de *datasets* (NOY, 2019). Assim, dada a importância destes repositórios, os esforços de instituições nacionais e internacionais para indexar e organizar *conjuntos de dados* abertos ao público é cada vez maior. No Brasil, por exemplo, temos, dentre estas iniciativas, o Portal Brasileiro de Dados Abertos: plataforma disponibilizada pelo governo brasileiro para que todos possam encontrar e utilizar dados e informações públicas (BRASIL, 2019). No cenário internacional a empresa GOOGLE – uma das gigantes da indústria informacional eletrônica contemporânea – lançou, em setembro de 2018, a ferramenta *Google Dataset Search (*referenciado pela empresa com a sigla GOODS*)*, que se propõe a se tornar um localizador e indexador de *datasets,* promovendo não só a descoberta destes repositórios (através de inteligência artificial, big data e outras tecnologias de dados) mas também estruturando um alinhamento a padrões de interoperabilidade de dados para que qualquer proprietário de *datasets* possa ali disponibilizá-los.

Dada a relevância das ferramentas disponibilizadas pela referida empresa, bem como a sua notável predominância neste mercado, o escopo e objeto de estudo deste trabalho orbita sobre o recém-lançado buscador *Google Dataset Search*.

2. Objetivos

O objetivo geral deste trabalho é obter uma visão geral do GOODS**,** identificando aspectos funcionais e técnicos, bem como dos padrões de dados utilizados e por fim perspectivas acerca da ferramenta. Como objetivos específicos, elencamos: a) Mapear estrutura de funcionamento e funções da ferramenta; b) Identificar procedimentos e padrões de interoperabilidade entre *datasets* e a plataforma; c) Realizar buscas de testes em sites de domínios **.br,** e; d) Conjecturar sobre as perspectivas da ferramenta e seus impactos informacionais para a comunidade científica.

3. Procedimentos Metodológicos

De acordo com Gerhardt e Silveira (org., 2009), esta pesquisa caracterizou-se como natureza aplicada, descritiva e de caráter exploratório. É também bibliográfica, pois fez uso de artigos, manuais e outros documentos a respeito ferramenta, disponibilizados pela própria empresa, disponíveis no site oficial: https://developers.google.com, e também por outros autores especialistas nesta temática.

Inicialmente, foi realizado um levantamento bibliográfico sobre a ferramenta, com o intuito de descrever seu funcionamento, estrutura tecnológica e padrões de interoperabilidade. O estudo bibliográfico acerca do *Google Dataset Search* foi realizado em meados do mês de agosto de 2019 através do buscador google.com, aplicando o termo "google dataset". Em seguida, foram observados e documentados aspectos funcionais através da realização de buscas específicas nobuscador da ferramenta com vistas a detectar e identificar resultados de buscas típicos de sites em domínios .br. Por fim, com base nos levantamentos e análises práticas, foram realizadas conjecturas sobre as perspectivas do GOODS para a organização dos conjuntos de dados para a comunidade científica.

4. Resultados

4.1. Sobre o Google Dataset Search

De acordo com os levantamentos bibliográficos realizados, foram encontrados 40.200 resultados, incluindo sobretudo notícias e postagens em fóruns técnicos especialistas. Poucos artigos e documentos científicos foram localizados, o que aparenta ser justificado pelo pouco tempo de lançamento da ferramenta. Grande parte destes foram escritos por desenvolvedores, funcionários e cientistas da própria Google.

Conforme a Google (2019), exemplos do que pode se qualificar como um conjunto de dados representáveis por metadados são: uma tabela ou um arquivo CSV com alguns dados; um conjunto organizado de tabelas;

um arquivo em formato proprietário que contenha dados; uma coleção de arquivos que unidos formam um conjunto de dados significativo; um objeto estruturado com dados em algum outro formato para processamento; imagens que capturam dados; arquivos relacionados ao aprendizado de máquina, como parâmetros treinados ou definições de estrutura de rede neural ou qualquer outro conjunto de dados representável e quantificável, em seus mais diversos formatos (Halevy, 2016): texto puro, planilhas, tabelas gigantes, sistemas de arquivos em nuvem, bases de dados relacionais etc. naturalmente ocasionando em uma ampla diversidade de metadados.

Colocadas estas considerações, o GOODS, segundo Noy (2019), surgiu partindo da dificuldade em encontrar repositórios de dados na internet. A autora afirma que, nos últimos anos, houve uma aumento significativo em quantidade, volume e tamanho, além da proliferação e expansão de dados desestruturados na web, que afetou também o mundo científico e as ferramentas de busca, que não conseguem localizar dados no espectro chamado "cauda longa da pesquisa" disponíveis na internet. A proposta portanto foi a criação de um buscador, segundo Halevy (2016), que fosse capaz de coletar, organizar e indexar de acordo com os padrões vigentes metadados de *datasets* acessíveis pela internet. Com base nestas premissas, o GOODS foi concebido em meados de 2015, lançado em 2018 e vem sendo aperfeiçoado. A figura 1 (apêndice A) mostra uma visão geral do *Google Dataset Search.* Robôs da Google coletam metadados das páginas web. Estes metadados são organizados, normalizados, indexados e organizados por prioridades para serem localizados por usuários através da interface de consulta (GOOGLE AI BLOG, 2019). Graças aos padrões de metadados, a plataforma consegue identificar os *datasets*, conectar com outras ferramentas (como o Google *Scholar* e o Google *Knowledge Graph*) e assim extrair de forma otimizada a informação desejada. A indexação dos metadados permitem ainda eliminar *datasets* duplicados ou disponibilizados em lugares diferentes.

4.2. Padrões, diretrizes e formatos aceitos para interoperabilidade entre *datasets*

As orientações gerais e formatos para desenvolvedores, segundo Google (2019) dizem que é possível processar dados estruturados em páginas da Web sobre conjuntos de dados das seguintes formas: ou usando a marcação de conjunto de dados do schema.org ou estruturas equivalentes representadas no formato de vocabulário do catálogo de dados (DCAT, na sigla em inglês) do W3C (páginas em inglês). Também estão sendo testadas suportes experimentais para dados estruturados com base no CSVW do W3C. Para auxiliar a compreensão, a tabela 1 traz definições sobre as siglas utilizadas neste trabalho e encontradas nas diretrizes de interoperabilidade do GOODS.

Tabela 1: informações básicas sobre as siglas de formatos de dados utilizadas neste trabalho.

SIGLA	DESCRIÇÃO
RDF	*Resource Description Framework*: representa meta dados no formato de sentenças sobre propriedades e relacionamentos entre itens na web.
JSON	Formato de interoperabilidade de dados entre sistemas, independente da linguagem de programação.
JSON-LD	JSON *Linked data*: é a maneira na qual a internet usa para conectar dados relacionados.
DCAT	*Data catalog vocabular*. Esquema de dados para facilitar a interoperabilidade entre dados de catálogos publicados na web.
URI	*Uniform resource identifier:* permite obter um identificador único para qualquer recurso na internet através de uma URL inteligente
Micro dados	Conjunto de etiquetas de organização de conteúdos que são legíveis por computadores e pessoas.
CSVW	CSV on the web working group: utiliza o padrão RDF para dados tabulares

Dados da pesquisa, 2019.

Criada pelo Google, Microsoft, Yahoo e Yandex, a schema.org é uma comunidade colaborativa com a missão de criar, manter e promover esquemas de dados estruturados na internet: em páginas web, mensagens de e-mail, conjuntos de dados e afins. O

vocabulário schema.org pode ser utilizado em diversas codificações (como RDFa, Microdados e JSON-LD), e seu vocabulário compreende entidades, relacionamentos entre entidades e ações entre elas para estruturar e organizar dados. Segundo o site oficial, mais de 10 milhões de páginas utilizam schema.org, dentre elas aplicações da empresa Google. As orientações e vocabulários para estruturar conjuntos de dados do schema.org estão disponíveis em https://schema.org/Dataset e incluem variáveis como autor, data de publicação, palavras chave, link de acesso, codificação, licença, versão, linguagem etc.

O DCAT (*data catalog vocabulary*) é um vocabulário RDF criado pela W3C para facilitar a interoperabilidade entre catálogos de dados publicados na internet. O W3C (*World Wide Web Consortium*) atualmente é uma organização composta por 450 membros, dentre eles órgãos governamentais, empresas, organizações independentes e comunidades científicas, com a finalidade de estabelecer padrões para criação e interpretação de conteúdo na internet (W3C, 2019).

Para serem inseridos no GOODS, os sites precisam seguir as diretrizes de dados estruturados (estar em JSON – padrão recomendado -, DCAT ou Microdados). Além dessas diretrizes, o site oficial indica as práticas recomendadas de *sitemap* (arquivos que auxiliam o Google a encontrar as URLs do site) e origem e procedência – sobretudo quando conjuntos de dados abertos são republicados, agregados e baseados em outros conjuntos de dados.

4.3. Testes exploratórios no GOODS

No tocante a utilização prática da ferramenta, o acesso pode ser feito pelo site https://toolbox.google.com/datasetsearch. Ao acessar a página, percebe-se o mesmo padrão visual do buscador da empresa. O usuário conta com três opções: 1) um campo de busca para realizar sua pesquisa por *datasets;* 2) uma caixa "sobre", com informações gerais sobre o GOODS e 3) um breve manual sobre como incluir conjuntos de dados na plataforma.

Os resultados de busca são dispostos da seguinte forma: à esquerda, com barra de rolagem, encontram-se os *datasets* encontrados, sendo os mais relevantes primeiro e os menos relevantes em seguida, de acordo com os termos utilizados na busca. Ao clicar em um deles, são exibidos à direita da tela, nesta ordem (de cima para baixo): o título do *dataset*, o link para acesso aos dados, informações gerais como data de publicação e atualização, fornecedor do conjunto de dados, licença, formato e descrição. A figura 2 (apêndice A) ilustra a tela inicial do *Google Dataset Search.*

Obedecendo a sugestão da própria ferramenta em sua página inicial, foi feita uma busca com o termo "*boston education data*". Foram obtidos mais de 100 resultados indicando *datasets* contendo estes termos. No entanto, ao realizar outras buscas, percebemos que a ferramenta limita a quantidade de resultados a "mais de 100". Observou-se que o buscador também permite alguns comandos de busca avançada do buscador tradicional da Google, como operadores "site:" para restringir a busca a um domínio específico, ou "inurl:" para identificar um termo dentro de uma URL específica, os caracteres asterisco (*) para substituir por qualquer conteúdo e aspas duplas para encontrar frase exata.

Com base nestes operadores e para fins de testes exploratórios, foram realizadas consultas de testes em sites e domínios governamentais do Brasil. Os termos utilizados, a quantidade de resultados encontrados e os itens de maior relevância (segundo ranking do GOODS) foram listados na tabela 2.

Tabela 2: Amostra de testes em sites de domínio.br realizados no GOODS.

	TERMO DE BUSCA	RESULTADOS	ITENS MAIS RELEVANTES
A	site:"*.gov.br"	91	ana.gov.br e cloud.csiss.gmu.edu
B	inurl:"*gov.br"	> 100	ana.gov.br e cloud.csiss.gmu.edu
C	inurl:"*edu.br"	0	-
D	inurl:"*com.br"	> 100	dados comerciais
E	site:"*.org.br"	3	-
F	universidade	> 100	GBIF
G	"são paulo"	> 100	Diversos

Dados da pesquisa, 2019.

Em uma análise superficial, os resultados de busca (A e B) apontam que grande parte dos dados indexados pelo GOODS são provenientes de dadosabertos.ana.gov.br (ANA - Agência Nacional de Águas), que oferece ferramentas para utilização de dados públicos sobre recursos hídricos no Brasil. Outro fato interessante é que muitos conjuntos de dados já estão indexados e disponíveis na plataforma csiss.gmu.edu, que pertence à GEOSS *Information Exchange DataHub*. A ferramenta informa que existem, na data deste trabalho, 13.187 conjuntos de dados etiquetados com "BRASIL" e 6.359 com a etiqueta "IBGE", mostrando forte relevância para a indexação de conjuntos de dados no Brasil. Em C nota-se que a ferramenta ainda não indexou *datasets* em domínios .edu.br. Observou-se que em D o GOODS indexou tabelas de preços e até especificações de produtos disponíveis em e-commerce. Em E nenhum dos resultados mostraram *datasets*, mostrando que a plataforma ainda precisa de ajustes. Em F, observou-se que existem diversas universidades brasileiras que disponibilizam seus dados pela plataforma internacional GBIF - Sistema Global de Informação sobre Biodiversidade. Em G tentou-se identificar instituições de ensino e pesquisa do Estado mais populoso do Brasil. Foram encontrados repositórios isolados ligados a ANA, Embrapa, e em plataformas como a *Zenodo, Kaggle* e *ResearchGate*, além de dados comerciais isolados.

Com relação ao formato dos datasets encontrados, os que mais se sobressaíram nos resultados de busca expostos pelos termos da tabela 2 foram *json, csv, pdf* e *xls*, além de API's próprias para consulta e exploração dos dados.

Outro fator relevante é que, naturalmente, os itens catalogados em português ainda são poucos se comparados ao idioma inglês.

É interessante notar também que, embora o portal brasileiro de dados abertos (dados.gov.br) siga os mesmos padrões e formatos seguidos pelo GOODS, não foram localizadas ocorrências que remetessem ao referido portal, embora o GOODS consiga indexar conteúdos em outras plataformas, como é o caso da GEOSS.

5. Considerações Finais

Certamente podemos afirmar que a empresa Google foi uma das principais agentes das transformações informacionais que a sociedade de hoje vivencia. Seja no tradicional buscador, ou nas ferramentas voltadas para a comunidade científica, a Google vem, entre prós e contras (listados por Canino, 2018), emplacando a sua contribuição para a comunidade científica.

Da mesma forma que o Google Acadêmico, que iniciou de maneira tímida seus trabalhos indo ao ar em 2004, e hoje é uma importante plataforma para a organização da informação científica, acredita-se que o Google *Dataset Search* tem grande potencial. A proposta de indexação e organização de datasets, que já é realidade em outras plataformas, mostra-se promissora quando vem sobre os ombros da gigante de tecnologia da informação, dotada de um gigantesco aparato computacional e tecnologias de ponta, como infraestrutura em nuvem, mineração de dados e inteligência artificial.

Fatores como robôs de busca e indexação de dados já existentes, padrões de dados internacionais e o poder financeiro e tecnológico credenciam esta perspectiva de esta ferramenta ser uma das mais importantes para a catalogação e disponibilização de *datasets* já nos próximos anos.

Partindo desta premissa, acredita-se numa não tardia integração com bases nacionais de conjuntos de dados, universidades e indústrias, sobretudo as que não possuem ferramentas para armazenar e distribuir seus datasets. Esta será uma grande contribuição para expor parte da cauda longa de pesquisa, tão importante para a comunidade científica.

Por fim, este breve estudo, ao oferecer uma visão geral e perspectivas para indexação e disponibilização de conjuntos de dados científicos abertos dentro da plataforma Google Dataset Search cumpre portanto os objetivos a que foi proposto.

Referências

BRASIL. Portal Brasileiro de Dados Abertos. 2019. Disponível em: <http://dados.gov.br>. Acesso em: 13/09/2019.

CANINO, Adrienne. **Deconstructing Google Dataset Search**. Public Services Quarterly, 15:3, 248-255, DOI: 10.1080 / 15228959.2019.1621793. Disponível em: <https://www.tandfonline.com/doi/full/10.1080/15228959.2019.1621793>. Acesso em: 13/09/2019,

FEBAB. 2017. Disponível em: <https://portal.febab.org.br/anais/article/view/1787>. Acesso em: 13/09/2019.

GAVRON, E. M.; CANTO, F. L. **Análise da utilização dos periódicos de acesso aberto de uma base de dados assinada pela Biblioteca Universitária da UFSC**. In: Anais do Congresso Brasileiro de Biblioteconomia, Documentação e Ciência da Informação.

GERHARDT E SILVEIRA (org.) **Métodos de pesquisa** / [organizado por] Tatiana Engel Gerhardt e Denise Tolfo Silveira. Porto Alegre: Editora da UFRGS, 2009. Disponível em: < http://www.ufrgs.br/cursopgdr/downloadsSerie/derad005.pdf>. Acesso em: 12/09/2019.

GOOGLE. **Conjuntos de diretrizes e orientações sobre o Google Dataset Search**. 2019. Disponível em: <https://developers.google.com/search/docs/data-types/dataset>. Acesso em: 13/09/2019.

HALEVY, A., Korn, F., Noy, N. F., Olston, C., Polyzotis, N., Roy, S., and Whang, S. E. **Goods: Organizing Google's datasets.** Google, 2016. Disponível em: < https://static.googleusercontent.com/media/research.google.com/pt-BR//pubs/archive/45390.pdf>. Acesso em: 11/09/2019.

NOY, Natasha. BURGESS, Matthew. BRICKLEY, Dan. **Google Dataset Search: Building a search engine for datasets in an open Web ecosystem.** WebConf'2019, May 2019, San Francisco, CA USA. Disponível em: <https://doi.org/10.1145/3308558.3313685>. Acesso em: 14/09/2019.

NOY, Natasha. Burgess, Matthew. **Building Google Dataset Search and Fostering an Open Data Ecosystem.** Google AI Blog. 2018. Disponível em: <https://ai.googleblog.com/2018/09/building-google-dataset-search-and.html>. Acesso em: 10/09/2019.

W3C - World Wide Web Consortium. **Data Catalog Vocabulary (DCAT).** 2014. Disponível em: < https://www.w3.org/TR/vocab-dcat/>. Acesso em: 09/09/2019.

W3C - World Wide Web Consortium. **Current Members**. Disponível em: <https://www.w3.org/Consortium/Member/List>. Acesso em: 09/09/2019.

Apêndice A – Relação de imagens referenciadas neste trabalho

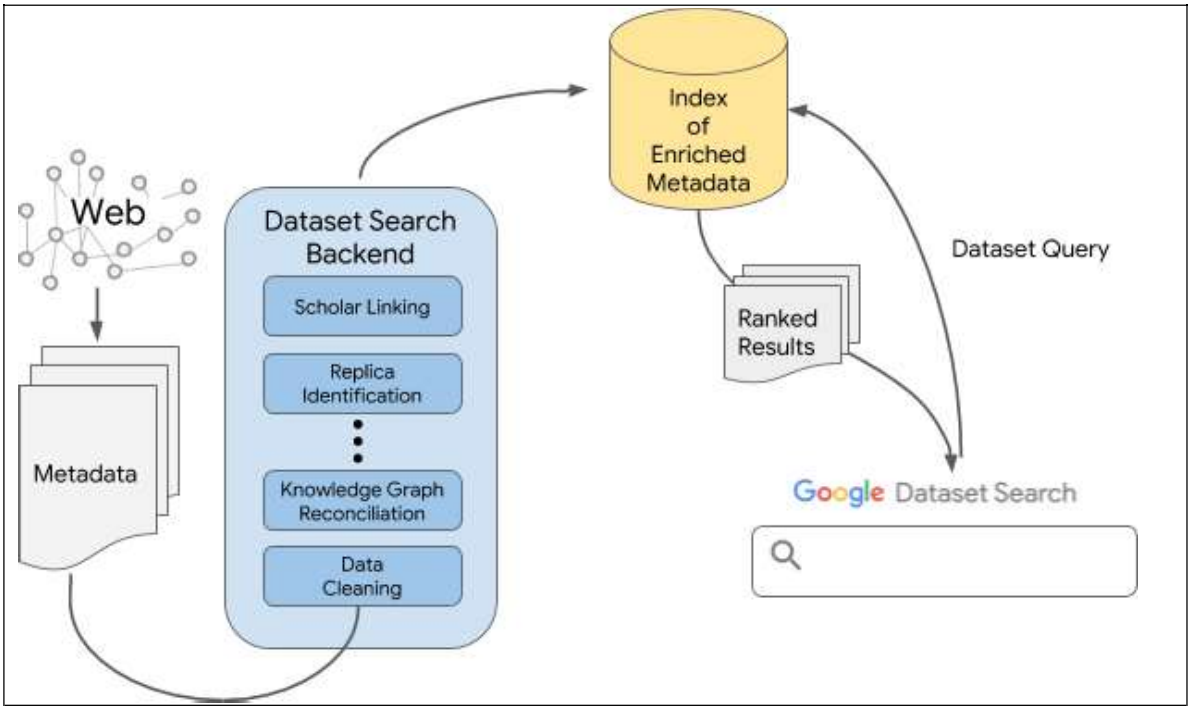


Figura 1: visão geral do Google Dataset Search (GOODS). Fonte: https://ai.googleblog.com/2018/09/building-google-dataset-search-and.html

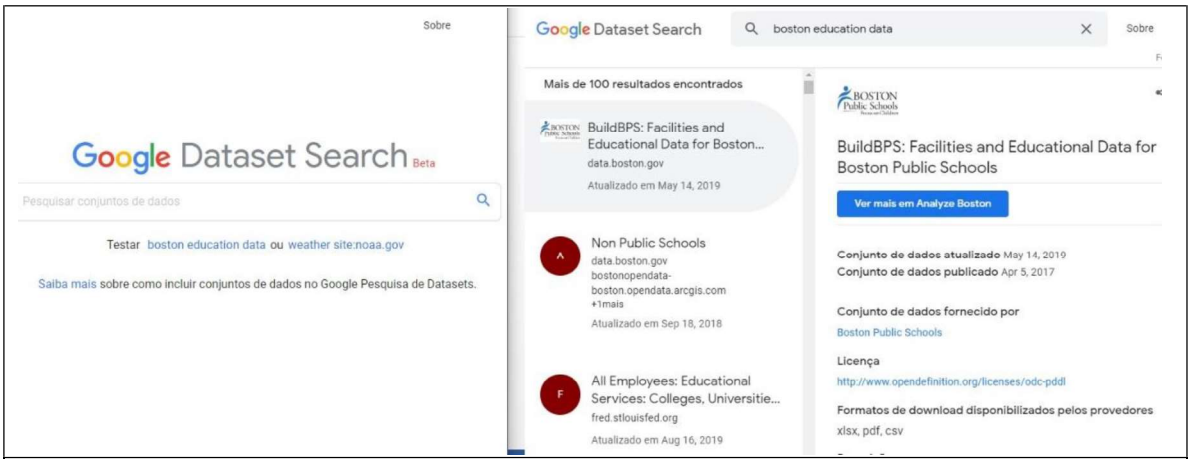


Figura 2: Visão geral da tela de busca do Google Dataset Search. Fonte: Própria, 2019

O DEBATE SOBRE PRIVACIDADE NO FÓRUM DE GOVERNANÇA DA INTERNET

The debate about Privacy in Internet Governance Forum

Adriana Veloso Meireles[1]
(1) Universidade de Brasília, dricaveloso@gmail.com

Resumo:
O artigo apresenta resultados quantitativos da análise dos discursos mobilizados em torno do tema da privacidade ao longo de onze anos do Fórum de Governança da Internet. O evento, realizado desde 2006, pela Organização das Nações Unidas é um ambiente qualificado que reflete os principais debates sobre internet e outras tecnologias da informação e comunicação. A metodologia de análise partiu de um recorte das atividades principais do evento, totalizando 137 transcrições. Para realizar a leitura dos arquivos padronizados foi escrito um script em Python para identificar o número de ocorrências das palavras chave; privacidade, direitos, vigilância e segurança. Os resultados indicam que este último tema teve um ápice na edição de 2013, ocorrida meses depois que Edward Snowden revelou as práticas de vigilância da Agência Nacional de Segurança dos Estados Unidos, seguido por um queda brusca do debate. Os dados enfatizam ainda que o termo direitos foi o mais mobilizado, com 3140 ocorrências, seguido pela palavra segurança com 2796 ocorrências. Por fim, os cálculos apontam que o termo privacidade foi mobilizado 1509 vezes e vigilância apenas 336 vezes.

Palavras-chave: privacidade, internet, segurança, vigilância, algoritmos.

Abstract:

The paper presents quantitative results of the discourse analysis regarding privacy along eleven years of the Internet Governance Forum. This event, held since 2006, by the United Nations, is a qualified environment that reflects the main debates about internet and other information and communication technologies. The methodology of the analysis selected the main activities held during the event, adding up to 137 transcripts. To read the standardized files, a Python script was written to identify the number of occurrences of the keywords; privacy, rights, surveillance, security. The results indicate that the latter theme peaked in the 2013 edition, months after Edward Snowden unveiled US National Security Agency surveillance practices, followed by a sharp drop in the debate. The data also emphasize that the term rights was the most mobilized, with 3140 occurrences, followed by the word security with 2796 occurrences. Finally, the calculations indicate that the term privacy has been mobilized 1509 times and surveillance only 336 times.

Keywords: privacy, internet, security, surveillance, algorithms.

1. Introdução

Nos últimos anos observou-se a passagem do acesso à internet dos computadores pessoais aos dispositivos móveis, primeiramente com os *smartphones* e mais recentemente com objetos inteligentes e a internet das coisas. Somado a isso, multiplicou-se a capacidade de armazenamento de informações. Graças sobretudo ao avanço das técnicas de compressão de dados, equipamentos relativamente pequenos e baratos são hoje capazes de guardar quantidades espantosas de informação, o que possibilitou o surgimento do fenômeno da *big data*, termo que se refere ao volume, processamento e cruzamento de dados gerados por dispositivos digitais (Baruh e Popescu, 2017; Tien, 2013). É neste sentido que ocorre uma "mudança na natureza da coleta de dados, realizada de forma automatizada e indiscriminada" (Nissembaum, 2009, p. 21).

Os dados pessoais são coletados, extraídos, analisados, processados e tratados por máquinas para finalidades pouco transparentes. Algoritmos simples, ou inteligentes, analisam, comparam e tratam informações dos mais variados formatos a todo momento. A principal diferença é que enquanto os algoritmos são códigos computacionais escritos para resolver problemas específicos, algoritmos inteligentes – sejam eles chamados de

inteligência artificial ou *machine learning* – são programados para solucionar problemas. O próprio programa assimila a resolução, mas como "eles não aprendem ou raciocinam como os humanos, isso pode fazer com que seus resultados sejam difíceis de prever e explicar" (Tutt, 2016, p. 87). Seus efeitos podem ser simples; um programa fechar sem salvar o trabalho feito, mas podem ser complexos quando suas consequências incidem diretamente na sociedade.

Estas transformações foram conceituadas por Zuboff (2019) como capitalismo de vigilância, fenômeno que ocorre há pelo menos duas décadas, ainda que o público em geral só tenha tido conhecimento sobre ele com as declarações de Edward Snowden, em 2013 sobre as práticas de vigilância da Agência Nacional de Segurança dos Estados Unidos. Suas revelações apontaram para uma convergência entre agências de defesa estadunidenses e empresas de tecnologia a partir dos atentados de 11 de setembro de 2001, em Nova Iorque.

Zuboff (2019) descreve como este acontecimento faz com que a prioridade do governo estadunidense se torne a segurança, portanto toda regulação que envolvia questões de privacidade e proteção de dados pessoais tornou-se irrelevante no contexto da "guerra ao terror". Essa conjuntura política, econômica e tecnológica, alicerçada em um modelo extremamente neoliberal de regulação, possibilitou a conformação do capitalismo de vigilância.

É neste contexto que o direito à privacidade deixa de ser suficiente para lidar com a quantidade de informações pessoais produzidas pela sociedade contemporânea. O direito à proteção de dados pessoais emerge, portanto, para proteger os indivíduos do uso indevido de suas informações, seja para fins de vigilância, como muitas vezes o tema é abordado, mas também para influenciar hábitos de consumo, preferências políticas, interferindo diretamente no exercício da cidadania e na autodeterminação informacional.

Diante do fenômeno do capitalismo de vigilância o artigo apresenta resultados quantitativos da análise dos discursos mobilizados em torno do tema da privacidade ao longo de onze anos do Fórum de Governança da Internet, evento organizado pelas Nações Unidas desde 2006, que reúne integrantes de diversos países e setores.

2. Objetivos

O objetivo do trabalho é mapear como o debate sobre privacidade e proteção de dados evoluiu a longo da última década. Para tanto a análise empírica identificou o Fórum de Governança da Internet como o principal espaço internacional em que temas como a internet e outras tecnologias são discutidas.

Para realizar a análise quantitativa foi elaborado um script em Python para filtrar palavras chave; direitos, privacidade, segurança e vigilância. O objetivo é detectar como os termos foram mobilizados pelos diversos interlocutores ao longo dos anos. Os dados servem também para orientar uma segunda etapa do estudo qualitativo, com o objetivo de construir uma cartografia de controvérsias em torno dos conceitos de privacidade, direitos, vigilância e democracia.

O recorte utilizado são as últimas doze edições do Fórum de Governança da Internet evento que agrega sociedade civil, empresas de tecnologia, academia, terceiro setor e representantes governamentais. Conhecido como modelo multissetorial, o evento reúne anualmente os principais atores que debatem governança da internet e seus impactos nas sociedades e democracias.

Com esta análise empírica busca-se mapear a evolução do discurso sobre privacidade, vigilância e direitos construindo parâmetros para uma cartografia sobre as principais controvérsias em torno destes temas (Babbie, 2015).

Entretanto, devido à limitação de espaço apresenta-se no artigo apenas a análise quantitativa dos dados e seus principais resultados, que apontam para algumas conclusões apresentadas a seguir.

3. Procedimentos Metodológicos

O IGF é estruturado com sessões principais, uma cerimônia de abertura e outra de encerramento e uma série de workshops que aumentam a cada ano. O primeiro IGF,

em Atenas, em 2006, teve onze sessões principais.

A partir do segundo encontro ocorreram atividades paralelas organizadas da seguinte forma; workshops, reuniões das coalizões dinâmicas, fóruns de melhores práticas, reuniões multilaterais, sessões temáticas do país sede, sessões instantâneas, fóruns abertos, diálogos inter-regionais, sessões de boas vindas.

A cada ano o número de atividades do evento aumenta, por isso, o recorte da análise focou nas sessões principais. A escolha teve como base o fato de que nem todas as atividades paralelas têm seu conteúdo transcrito e disponibilizado no site do IGF. Além disso, ao dar ênfase às sessões principais é possível realizar uma análise quantitativa sem viés, dado que como o número de atividades ao longo dos anos aumenta a ocorrência dos termos cresceria proporcionalmente.

O recorte proposto para a análise quantitativa engloba as sessões principais da primeira edição em Atenas à última edição, em Genebra, na Suíça, em 2017. O levantamento documental junto ao site do evento identificou 137 sessões principais.

Estes arquivos, que consistem na transcrição do que foi dito durante a atividade, foram baixados e quando necessário convertidos em arquivos de texto para que o programa pudesse realizar sua leitura (em um dos aos os arquivos estavam com codificação antiga e em outros em pdf, mas sempre padronizados em seu conteúdo).

Em seguida, com os arquivos padronizados e estruturados por ano aplicou-se o programa escrito em *Python* para identificar o número de ocorrências em que os principais termos de pesquisa eram mobilizados durante as sessões principais. O código simples e intuitivo contou com a colaboração do pesquisador João S. O. Bueno e se encontra no apêndice A do artigo.

O programa foi escrito para buscar palavras chave e exportá-las em um arquivo no formato de valores separados por vírgulas (o formato csv), que permite a importação das informações e resultados em planilhas. As palavras chave selecionadas nas sessões principais do evento foram; privacy, security, e surveillance, rights, em português; privacidade, segurança, vigilância e direitos. Os resultados da busca realizada pelo programa serão apresentados a seguir.

4. Resultados

Uma tabela com os resultados dos termos buscados pelo programa são apresentados no apêndice B.

Os dados indicam que o termo direitos foi o mais mobilizado nos discursos centrais do IGF, com mais de três mil ocorrências ao longo dos onze anos analisados. O ano de 2011 é aquele em que a palavra “direitos” é menos mencionada nas sessões principais do IGF.

O ano em que o termo foi mais mobilizado foi o de 2015, conforme observa- se no gráfico 1.

1. Gráfico: A palavra direitos nas sessões principais do IGF (Fonte: Dados da Pesquisa, 2019)

Das 137 sessões principais do IGF em apenas seis delas o termo “direitos” não é evocado pelos participantes. Dentre elas, duas ocorreram em 2007 e outras quatro em 2008.

Sendo assim, pode-se afirmar que de 2009 em diante o termo “direitos” foi mencionado pelos participantes em todas as sessões principais do IGF.

A atividade em que o termo foi mais mobilizado – com 305 ocorrências – foi a sessão de 2016 intitulada “Direitos Humanos; ampliando o debate”, ocorrida em 2016.

Observa-se uma queda no ano de 2014, entretanto não apenas com o termo “direitos”, como também os outros pesquisados, conforme será apresentado a seguir.

Dentre as palavras selecionadas para análise, o segundo termo mais utilizado durante as 137 sessões principais do evento foi “segurança”, com 2796 ocorrências.

Em 2006 e 2007 houveram sessões principais específicas sobre segurança. Em 2008 duas atividades principais trataram do tema. A primeira delas intitulada privacidade, segurança e abertura (security, openness and privacy), que se repete até o ano de 2012. A segunda atividade de 2008 que abordou o tema foi sobre “dimensões da cibersegurança e do cibercrime”.

Nos anos seguintes, apenas em 2015 a segurança voltou a ser tema dentre as sessões principais do evento, representando seu maior índice, conforme observa-se no gráfico 2;

2. Gráfico: Gráfico: A palavra segurança nas sessões principais do IGF (Fonte: Dados da Pesquisa, 2019)

Novamente observa-se uma queda de ocorrências no ano de 2014. O índice baixo é acompanhado do ano de 2016, quando também não houve nenhuma sessão principal tratando do tema da segurança.

Por sua vez, a palavra “privacidade” aparece nos discursos das sessões principais do IGF 1509 vezes.

O ano em que a palavra “privacidade” é mais mobilizada nas sessões principais do IGF é o de 2008, conforme observa-se no gráfico 3.

3. Gráfico: A palavra privacidade nas sessões principais do IGF (Fonte: Dados da Pesquisa, 2019)

Nos anos iniciais não houve uma sessão principal específica para privacidade, apenas para segurança. Conforme já mencionado entre 2008 e 2012 ocorreram trilhas principais sobre privacidade, segurança e abertura.

De 2013 em diante não houveram sessões principais com o tema da privacidade.

Por outro lado, neste ano de 2013, ocorreu uma atividade principal justamente sobre vigilância na internet. Na ocasião palavra “vigilância” foi mencionada em 87 vezes, representando a maior ocorrência do termo nas sessões principais do IGF.

No total, o termo “vigilância” foi mencionado 336 vezes ao longo dos anos nas atividades principais do IGF, sendo que apenas no ano de 2013 houveram 175 ocorrências, conforme observa-se no gráfico 4.

4. Gráfico: A palavra vigilância nas sessões principais do IGF (Fonte: Dados da Pesquisa, 2019)

Importante contextualizar que o Fórum ocorre poucos meses depois das revelações

de Edward Snowden sobre a vigilância em massa realizada pela agência nacional de segurança estadunidense.

É relevante pontuar ainda que neste ano de 2013, outra sessão principal, intitulada "Direitos humanos, liberdade de expressão e a livre circulação de informação na internet", o termo "vigilância" é mobilizado 64 vezes, o que contribui para o aumento de ocorrências do termo.

Por fim, deve-se sublinhar que nos primeiros Fóruns o termo "vigilância" praticamente não era abordado nas sessões principais. A palavra é mobilizada nestes primeiros anos nas trilhas sobre segurança. Por outro lado, ainda que em menor quantidade o termo segue presente nas próximas edições do evento.

Estes resultados parciais indicam que o tema dos direitos humanos é o mais frequente nos debates do Fórum de Governança da Internet, o que é compreensível dado que ele é organizado pela Organização das Nações Unidas.

A segurança – em suas variadas dimensões – é outro tema recorrente no IGF. Observa-se que a questão é acompanhada do debate sobre privacidade e proteção de dados pessoais em algumas ocasiões, mas em outras o tema é mobilizado a partir da perspectiva da prevenção e combate de crimes online.

O debate sobre privacidade é tímido nos anos iniciais 2006 e 2007. Nos anos seguintes a questão parece se consolidar até que se observa uma queda em 2012 e em 2014. Interessante sublinhar que no ano de 2013, quando o debate sobre vigilância ganha destaque, o termo privacidade não obtém ocorrências maiores. Por fim, os dados indicam uma queda de ocorrências da palavra privacidade nos últimos anos analisados, ou seja, 2016 e 2017.

O índice de maior relevância da análise quantitativa certamente é o número de ocorrências do termo vigilância em 2013. O tema que em anos anteriores aparecia de forma muito discreta ou quase nula, se torna central em 2013. Por outro lado, observa-se que a temática não se consolida nos anos subsequentes, estando presente nos debates das sessões principais do IGF de forma discreta.

5. Considerações Finais

O artigo apresentou resultados quantitativos da pesquisa que busca mapear as controvérsias em torno do debate sobre privacidade e proteção de dados pessoais durante a última década no âmbito do Fórum de Governança da Internet.

Portanto, ainda que estejam expostos resultados parciais, o objetivo central do artigo foi alcançado, uma vez que os quatro gráficos apresentados ilustram de forma eficiente a variação das palavras chave nas sessões principais do Fórum de Governança da Internet.

A análise empírica dos dados indica que o tema da vigilância foi central no ano de 2013, quando Edward Snowden divulgou as práticas da Agência de Segurança dos Estados Unidos.

Por outro lado, os dados indicam que tanto a privacidade como a vigilância perderam destaque nos anos mais recentes dentre os debates das sessões principais do IGF. Já o tema da segurança obteve seu ápice justamente no último ano analisado; 2017.

Pode-se concluir que o fenômeno da vigilância, que foi motivo de escândalos de grandes proporções em 2013, tornou-se secundário nos debates mais recentes. De forma semelhante a privacidade também perde destaque nas discussões centrais do IGF. Por outro lado, o tema da segurança segue presente.

É um diagnóstico preocupante principalmente diante da expansão da internet das coisas. Se a mudança da computação pessoal do *desktop* para os dispositivos móveis mostrou-se como um novo paradigma, a internet das coisas irá representar uma expansão ainda maior da coleta de dados pessoais. É neste sentido que se compreende que o "privado" nunca foi tão político como na atualidade.

Sendo assim, apenas regulações de proteção de dados pessoais não serão suficientes para lidar com a expansão da internet das coisas. Conforme mencionado na introdução do artigo, é necessário observar a evolução dos algoritmos, principalmente os que tomam decisões

"autônomas" que podem impactar diretamente na sociedade.

Diante deste contexto aplica-se a máxima *cypherpunk* "privacidade para os fracos e transparência para os fortes" (Assange, 2013), ou seja, proteção aos cidadãos e controle social para governos, empresas e seus algoritmos.

É cada vez mais relevante a regulação dos algoritmos, ou uma governança algorítmica, para que as plataformas se tornem mais transparentes com relação ao uso de dados pessoais.

Referências

ASSANGE, J. **Cypherpunks: liberdade e o futuro da internet**. São Paulo: Boitempo. 2013.

BABBIE, Earl R. **The practice of social research**. Nelson Education, 2015.

BARUH, L.; POPESCU M.. **Big data analytics and the limits of privacy self- management**. New Media & Society, vol. 19, nº 4, p. 579-96. 2017. NISSENBAUM, H. **Privacy in context: technology, policy, and the integrity of social life**. Stanford: Stanford University Press. 2009.

TIEN, J. M. **Big data: unleashing information**. Journal of Systems Science and Systems Engineering, vol. 22, nº 2, pp. 127-51. 2013.

TUTT, Andrew, **An FDA for Algorithms** (March 15, 2016). 69 Admin. L. Rev. 83 (2017).

ZUBOFF, Shoshana. **The age of surveillance capitalism: The fight for a human future at the new frontier of power**. Profile Books, 2019.

Apêndice A – Script Python

Este é o script do programa em Python utilizado para contar palavras.

```
import sys
import os
import csv
def procura_palavra(palavra):
    saida = open(f'dados_{palavra}.csv', 'wt')
    escritor = csv.writer(saida)
    escritor.writerow(['Palavra', 'Contagem', 'Arquivo'])
    for pasta, pastas, arquivos in os.walk("."):
        if "env" in pasta or  ".git" in pasta:
            continue
        if not "20" in pasta:
            continue
        for arquivo in arquivos:
            caminho = pasta + "/" + arquivo
            conteudo = open(caminho).read()
            conteudo_minuscula = conteudo.lower()
            contagem = conteudo_minuscula.count(palavra)
            escritor.writerow([palavra, contagem, caminho])
            # print(f"{palavra},{contagem},{caminho}")
    print("Concluído")
def principal():
    palavras  = ["privacy", "democracy", "rights",
"surveillance", "freedom", "security"]

    for palavra in palavras:
        procura_palavra(palavra)

principal()
```

Apêndice B – tabela de resultados do script

Tabela 1: Resultados dos termos pesquisados pelo programa

ANO	privacy	security	surveillance	rights
2006	45	332	7	189
2007	72	234	5	194
2008	207	339	1	193
2009	206	177	2	195
2010	183	230	1	156
2011	138	193	7	141
2012	119	139	21	265
2013	134	216	175	406
2014	71	70	21	167
2015	150	367	45	456
2016	92	127	30	427
2017	92	372	21	351
TOTAL	1509	2796	336	3140

O USO DA BLOKCHAIN PARA REGISTROS DE IDENTIDADE DE PESSOAS

Análise dos dados coletados e riscos à proteção de dados pessoais e privacidade

USING BLOKCHAIN FOR PERSONAL IDENTITY RECORDS

Analysis of collected data and risks to personal data protection and privacy

José Antonio Maurilio Milagre[1], José Eduardo Santarém Segundo[2]

(1) Universidade Estadual Paulista, Marília/SP, jose.milagre@unesp.br

(2) Universidade de São Paulo, Ribeirão Preto/SP, santarem@usp.br

Resumo:

As transformações trazidas pela rede descentralizada Blockchain são inúmeras. A estrutura, que nasceu para registrar as transações em criptomoedas hoje vem sendo utilizada para outras aplicações, como registro de obras autorais, autenticação de documentos e até mesmo para substituir a forma como as pessoas se identificam em empresas, serviços do governo e em ambientes, onde já se desenvolve aplicações de identidade digital baseadas na tecnologia, que substituiriam os documentos em papel. A presente pesquisa tem o objetivo de, a partir da criação de uma identidade digital Blockchain, valendo-se da aplicação Original.My, avaliar quais os dados pessoais coletados, como são armazenados e transacionados, e quais os riscos existentes à privacidade e uso indevido de dados pessoais. Como resultados, foi possível identificar que, apesar do titular não ter que ceder seus dados pessoais a todos os serviços, se identificando apenas com o ID digital, os dados pessoais associados ao ID podem ser identificados, o que pode gerar não apenas uma rastreio das atividades do titular, mas também existindo o risco de criação de identidades falsas com prejuízos aos titulares. Em conclusão, constata-se que embora a Identidade digital com base na Blockchain seja prática e mais segura, também oferecer riscos à privacidade e dados pessoais.

Palavras-chave: blockchain, privacidade, identidade, registro, dados pessoais.

Abstract:

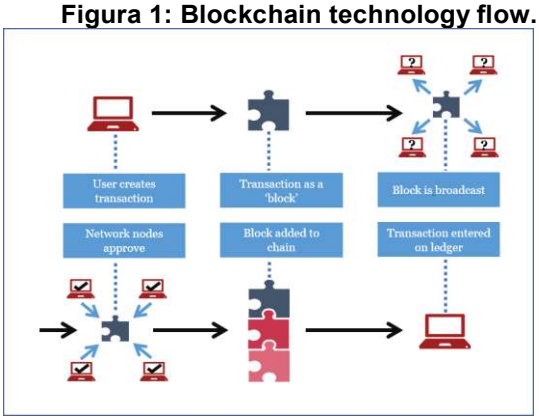
The transformations brought about by the decentralized Blockchain network are numerous. The framework, which was born to record cryptocurrency transactions today, is being used for other applications, such as copyright registration, document authentication and even to replace the way people identify themselves in business, government services and environments, where technology-based digital identity applications are already being developed that would replace paper documents. This research aims, from the creation of a Blockchain digital identity, using the Original.My application, to evaluate which personal data is collected, how it is stored and transacted, and what are the risks to privacy and misuse of personal data. As a result, it was possible to identify that, although the holder does not have to assign his personal data to all services, identifying himself only with the digital ID, the personal data associated with the ID can be identified, which can generate not only a tracking of the holder's activities, but there is also a risk that false identities may be created to the detriment of users. In conclusion, while Blockchain-based Digital Identity is practical and safer, it also poses risks to privacy and personal data.

1. Introdução

A Blockchain é uma tecnologia nova e emergente, desenvolvida para ser, a princípio, a base para as transações envolvendo ativos digitais, funcionando como um “livro razão”, registrando de forma indelével todas as transações na rede descentralizada.

Nakamoto (2008) confirma ser a Blockchain um meio descentralizado para transações de criptomoedas, de forma que cada transação é recebida por servidores e encaminhada aos pontos ou mineradores para que atestem a integridade da transação. Neste momento, a referida transação é gravada no bloco e permanece registrada, podendo ser consultada pelo código da transação, endereço do remetente e destinatário, dentre outras formas, exemplificado na Figura 1.

Figura 1: Blockchain technology flow.



Fonte: European Payments Council (2019)

A mesma estrutura que suporta as transações em criptomoedas vem sendo refletida para registrar e servir de base para inúmeras funções, como contratos eletrônicos, registros de obras, fatos e criações intelectuais, registros de prontuários médicos, dentre outras. Uma das funções que mais se destacam com o uso da tecnologia é a autenticação em ambientes, empresas e locais, por meio de um gerenciamento de identidades de pessoas registrados na Blockchain. Uma das vantagens da autenticação da identidade na Blockchain está relacionada a economia de cadastros, considerando que o titular, em toda sua vida se relacionando e fazendo usos de sites, portais, normalmente tem que fazer cadastros diversos, inclusive digitando senhas, que podem ser comprometidas. Além disso, justifica-se a criação de um padrão de identidade único e digital pela mora que o cidadão tem para realização de diversos cadastros e registros, como CNH, Passaporte, dentre outros. Para Rovoredo e Borges (2019), a tecnologia ainda permitiria o melhor controle de identidade e a aplicação ao direito ao esquecimento, em linha com os novos regulamentos de proteção de dados, do mesmo modo, considerando ser "descentralizado", o que aprimoraria o fator segurança.

O problema envolvendo identidades digitais geradas por diversas autoridades é um dos principais motivadores para criação de uma identidade digital única. Governos criam identidades diversas e inúmeros cadastros, muitos off-line e que não se comunicam entre si. A Blockchain pode oferecer uma tecnologia para contornar este problema, sem a necessidade de uma autoridade central para validar um documento, dando aos indivíduos o poder de não compartilhar dados com terceiros e o controle de quais pessoas podem acessar os dados. Uma identidade digital funciona como uma marca d'água, que pode ser assinada em qualquer transação online, ajudando organizações a identificarem a identidade em qualquer transação. Consumidores poderiam se cadastrar em serviços sem ter que digitar cadastros, username e senhas. A solução funciona com base em um par de chaves, sendo que a pública pode ser compartilhada com companhias e o próprio governo (JACOBOVITZ, 2016, p. 3)

Iniciativas estão em andamento, como a ID2020, uma organização não governamental que busca oferecer identidade a pessoas sem identificação formal, com base na Blockchain (ID2020, 2019)

Outras iniciativas de identificação com base na Blockchain são *Shocard, Uport, I/O Digital, BlockAuth, UniquID*, dentre outras (JOCOBOVITZ, 2016, p.9)

Por outro lado, sistemas de gestão da identidade baseados na Blockchain poderiam registrar dados pessoais e revelar um histórico indelével de ações tomadas pelo titular dos referidos dados, o que constituiria um risco à privacidade com identificação baseada na Blockchain.

Na presente pesquisa, analisou-se uma das plataformas utilizadas para criação e gestão de identidades com base na Blockchain (Blockchain ID da Original.my), avaliando-se os dados coletados, definindo sua classificação e como são tratados na Blockchain, investigando-se quais os riscos para a privacidade e proteção de dados dos titulares com o tratamento de tais dados.

Com efeito, pretende-se levantar oportunos resultados sobre os rumos da utilização da tecnologia na gestão da identidade digital e alertas para riscos existentes em relação à proteção de dados.

2. Objetivos

O objetivo da pesquisa é a avaliação da gestão de identidade da plataforma

Blockchain ID (Original.my, 2019) e a partir dela, identificar quais dados são coletados para criação de uma identidade Blockchain, como são custodiados, de que natureza são e, especificamente, avaliar a possibilidade de riscos à proteção de dados e privacidade diante das autenticações gravadas na Blockchain.

3. Análise da plataforma Original.My para gestão de identidades na Blockchain

A pesquisa considerou uma aplicação existente para gerenciamento online de identidade, com base no registro da BlockChain. Avaliamos na pesquisa, a plataforma Identidade Blockchain oferecida pela Original.My (2019), de modo a identificar os dados que foram coletados para uma identidade digital. A Identidade Blockchain é uma aplicação que registra dados do titular na plataforma, e então submete o resultado destes dados para a Blockchain.

Acessou-se o sistema para geração de uma identidade digital (https://originalmy.com/app) e a partir dele, criou-se um cadastro na plataforma de identidade digital

A realização de um cadastro apresentou a necessidade de coleta dos dados (Tabela 1), que foram categorizados pelos Autores.

Tabela 1: Dados coletados para geração de uma identidade digital.

Índice	Dado	Categoria	Detentor
1	Usuário	Pessoal indireto	BlockchainID
2	Senha	Pessoa indireto	BlockchainID
3	Validação aparelho celular	Pessoal indireto	BlockchainID
4	Validação de CPF	Pessoal direto	BlockchainID
5	Selfie do rosto	Pessoal direto	BlockchainID
6	Foto do documento	Pessoal direto	BlockchainID
7	Blockchain ID	Pessoal indireto	Titular
8	Blockchain ID chave pública	Pessoal Indireto	BlockchainID

Fonte: Os próprios Autores.

Identificados os dados, constatou-se, o sistema gerou 12 (doze) palavras para verificação de senha, que ficam custodiadas com o responsável pela identidade é a chave para acessar a identidade diante da troca do dispositivo móvel.

Como se verificou, embora os dados da identidade não fiquem armazenados diretamente nas plataformas privadas ou publicas da Blockchain, servindo apenas para checagem, a aplicação Original.my custodia uma série de dados e, como em uma transação em criptomoedas, a partir dos dados, gera uma "identidade blockchain" ou código *Hash* (fórmula matemática) que serve para autenticações. Um código criptográfico que pode ser aposto em contratos e que é negociado com entidades nas quais o titular transacionar, que podem consultar a aplicação e por sua vez a Blockchain para verificar a identidade atrelada ao indexador "Identidade Blockchain".

Os dados ficam registrados no aplicativo e a identidade é encorajada a ser utilizada, inclusive na assinatura de contratos. (ORIGINAL.MY, p1. 2019).

Deste modo, considerando que pode haver divulgação da identidade digital associada a documentos, os dados pessoais poderão ser com facilidade associados a uma identidade digital Blockchain, que é uma sequencia numérica. Com efeito, a privacidade e o rastreio de locais e plataformas online onde o titular realizou autenticações, como sites, governo, lojas virtuais, aplicativos e serviços pode ser comprometida. Uma alternativa seria a possibilidade de criar diversas "identidades" para uma mesma pessoa.

Como se constatou, a Blockchain ID (aplicação Original.My) é controladora (detentora) de 7 (sete) dos 8 (oito) campos que compõe a identidade digital. Assim, embora o titular não tenha em tese que contar com intermediários, como cartórios ou mesmo fornecer dados inúmeros repositórios para se autenticar, é fato que as aplicações que farão a gestão da identidade junto a Blockchain, como a Original.My, armazenam dados pessoais e até mesmo fotografias de documentos oficiais, sendo, portanto, custodiantes e controladoras dos dados que são usados para conectar o id na Blockchain à documentação pessoal fornecida.

4. Resultados

A custódia da identidade é feita hoje pelo próprio titular dos dados, que utiliza documentos em papéis ou fornece seus dados à diversas plataformas na Internet, que checam os dados em portais do governo e que passam a ser controladoras e responsáveis para com a segurança das referidas informações, categorizadas como dados pessoais.

Diante da criação de uma identidade digital associada à Blockchain, e que pretende facilitar o processo de identificação de pessoas em serviços, a partir da aplicação Original.My, identificou-se os riscos (Tabela 2. Dados de Pesquisa, 2019) envolvendo o uso de dados pessoais.

Tabela 2: Riscos envolvendo os dados pessoais atrelados à uma identidade digital BlockChain.

Id	Risco	Impacto	Severidade
1	Vazamento da Custódia da BlockChain ID chave Publica	Comprometida a integridade das transações	Alta
2	Comprometimento do Aplicativo intermediário de Identidade Digital	Rastreio das atividades com a identidade	Alta
3	CPF atrelado a identidade na Blockchain de forma indelével	Possibilidade de rastreio das autenticações e atividades do titular e locais onde ele apresentou sua identidade	Alta
4	Acesso indevido ao celular do titular e chave privada	Autenticações feitas como se fosse o titular dos dados. Uso de sua identidade na Internet	Alta
5	Perda das 12 (doze) palavras para recuperação da senha	Perda permanente da identidade.	Alta

Fonte: Os próprios Autores.

Como identificado com a pesquisa, os dados pessoais necessários e adequados para geração de uma identidade digital não são gravados na Blockchain, mas apenas um código (Id digital ou *hash*) gravado na rede é o código que fica associado aos dados cadastrais, que ficam custodiados ou armazenados nas aplicações que estão surgindo para identidade digital, como a Original.MY, a qual avaliamos nesta pesquisa. Assim, o comprometimento dos dados nestas aplicações pode permitir que um terceiro crie uma nova identidade Blockchain a partir dos documentos enviados as aplicações, o que permitira o denominado “furto de identidade”.

Identificado que a identidade digital continua lastreada em documentos expedidos pelo Estado, como RG e CNH, sendo que o acesso a estes documentos pode permitir a criminosos à criação de identidades digitais dos titulares, sem que estes saibam, o habilitando para fazer transações e se identificar de forma digital como se fossem as verdadeiras pessoas.

5. Considerações Finais

O fato da Blockchain ser descentralizada não significa que os dados estão gravados na rede de forma descentralizada. A rede não armazena os dados pessoais, mas apenas o código da identidade, que pode ser associada ao titular pela aplicação de identidade utilizada ou por meio de correlação. No caso desta pesquisa, utilizou- se a plataforma Original.my, simulando a criação de uma identidade digital.

Assim, não se pode afirmar que a identidade está mais segura do que armazenada em sistemas convencionais governamentais, considerando que em que pese não mais ser fornecida ou mantida por esses repositórios ou não necessariamente ser fornecida em cadastros de sites e portais, em caso de uso da identidade digital baseada na Blockchain, esta, necessariamente, mantém-se armazenada no sistema utilizado para interfacear a identidade do titular (código) gravado na Blockchain e seus dados pessoais coletados e armazenados na aplicação.

E em caso de fechamento ou encerramento das atividades desta empresa privada que faz a interface, não se pode garantir que a identidade funcionaria sem a previsão de mecanismos de portabilidade dos dados, o que pode sugerir que a evolução mais segura da identidade para a

baseada na Blockchain possa ser a realizada pelo próprio governo e autoridades públicas que hoje já gerenciam identidades dos cidadãos. Como visto, em diversas operações de uma identidade digital, faz-se possível comprometer a privacidade do titular dos dados, inclusive, permitindo-se que terceiros conheçam as atividades e comportamentos do titular com base em associações entre seu ID digital e seus dados.

Ainda, é possível a partir dos dados pessoais, solicitar identidades digitais em aplicações, sendo existente a possibilidade de furto de identidade, onde um terceiro registra na Blockchain uma identidade antes mesmo do seu real titular, com base em dados obtidos indevidamente.

Com efeito, foi possível identificar também os dados que são coletados em uma aplicação de gestão de identidade, sua criticidade, como são armazenados e se possível associar os referidos dados as titulares, sendo apresentados riscos consideráveis à privacidade e proteção de dados no uso de identidade digital, incluindo rastreio indelével de histórico de atividades de autenticações.

Como trabalhos futuros, pretende-se avaliar outras plataformas de Blockchain ID, realizando-se a criação dos cadastros e atividades em sites e portais, posteriormente buscando identificar pessoas por trás de transações, o que pode ser considerado um risco à privacidade. Igualmente propor um modelo para identidade digital que maximize a proteção de dados pessoais e evite as vulnerabilidades identificadas na presente pesquisa.

Referências

E-estonia.In: e-identity. Enterprise Estônia. Disponível em: < https://e-estonia.com/solutions/e-identity/id-card/ >.Acesso em: 03 set.2019.

European Payments Council. Disponível em: https://www.europeanpaymentscouncil.eu/news-insights/insight/blockchain-applications-payments >.Acesso em: 03 set.2019.

ID2020. Disponível em:<https://id2020wa.com/>. Acesso em: 31 ago. 2019

Jayachandran, Praveen. In: Blockchain Explained—The difference between public and private blockchain. Publicado por IBM Blockchain Blog em May 31, 2017. Disponível em: https://www.ibm.com/blogs/blockchain/2017/05/the-difference-between-public-and-private-blockchain/ Acesso em: 03 set.2019.

JACOBOVITZ, Ori. Blockchain for Identity Management. Disponível em:<

https://www.cs.bgu.ac.il/~frankel/TechnicalReports/2016/16-02.pdf>. Acesso em: 05 set.2019.

MILAGRE, José Antonio. Como fiz meu primeiro registro de livro na blockchain e protegi meus direitos autorais. Disponível em:< http://josemilagre.com.br/blog/2018/03/16/como-fiz-meu-primeiro-registro-de-livro-na-blockchain-e-protegi-meus-direitos-autorais/>. Acesso em: 01 ago. 2019.

MILAGRE, José Antonio; SANTARÉM SEGUDNO, José Eduardo. Possibilidade de identificação de violações a direitos autorais com base em metadados gerados na Blockchain: Avaliação da Plataforma Original.my. Disponível em:< http://enancib.marilia.unesp.br/index.php/XIXENANCIB/xixenancib/paper/viewFile/1327/1734 >Acesso em: 23. Fev. 2019

ORIGINAL.MY. Recomendações. Contratos. Disponível em:<http://docs.originalmy.com/pt_BR/latest/80-recomendacoes.html#contratos> Acesso em: 01 ago. 2019

PADOVAN, Thiago. E se no futuro, os cidadãos controlassem sua própria identidade. Disponível em:< https://blockchainacademy.com.br/e-se-no-futuro-os-cidadaos-controlassem-sua-propria-identidade/> Acesso em: 15 ago. 2019

ONTOLOGIAS MULTIMÍDIA: um estudo comparativo para reúso

MULTIMEDIA ONTOLOGIES: A comparative study for reuse

Daniela Lucas da Silva Lemos[1]
(1) Universidade Federal do Espírito Santo, Vitória, Brasil, danielalucas@hotmail.com

Resumo:

O artigo trata de um estudo sistemático realizado em iniciativas de padrões de metadados, vocabulários, modelos e ontologias voltados ao domínio da descrição multimídia, que culminou na obtenção de um *ranking* de ontologias a partir de uma análise comparativa e uma avaliação criteriosa sobre dimensões concernentes a reúso de recursos de conhecimento disponíveis na Web de dados. Metodologicamente, a pesquisa foi classificada como sendo de natureza qualitativa e quantitativa, de caráter exploratório e descritivo à luz de literatura científica já publicada e material empírico específico, o que a torna bibliográfica e documental. Para a seleção e a análise das ontologias multimídia foi utilizado o guia *NeOn Methodology* que orientou na definição de critérios e categorias de análise fundamentais ao procedimento de coleta, organização e análise dos dados. Como resultado, evidenciou-se as ontologias mais proeminentes para o domínio da descrição multimídia em que se tornaria possível selecionar os recursos de conhecimento provenientes de suas estruturas visando a proposição de um modelo conceitual de referência para a organização e representação desse tipo de domínio.
Palavras-chave: Ontologias Multimídia; Padrões de Metadados; Reúso de Ontologias; Descrição Multimídia.

Abstract:
The article deals with a systematic study carried out in initiatives of metadata standards, vocabularies, models and ontologies focused on the domain of multimedia description, which culminated in obtaining a ranking of ontologies from a comparative analysis and a careful evaluation of dimensions concerning reuse of knowledge resources available on the data web. Methodologically, the research was classified as qualitative and quantitative, exploratory and descriptive in the light of published scientific literature and specific empirical material, which makes it bibliographic and documentary. For the selection and analysis of multimedia ontologies, the NeOn Methodology guide was used to guide the definition of criteria and categories of analysis fundamental to the procedure of data collection, organization and analysis. As a result, it was evidenced the most prominent ontologies for the domain of multimedia description in which it would be possible to select the knowledge resources from its structures aiming at proposing a conceptual reference model for the organization and representation of this type of domain.
Keywords: Multimedia Ontologies; Metadata Standards; Reuse of Ontologies; Multimedia Description.

1. Introdução

Nos últimos anos, observou-se um crescimento significativo de dados semanticamente relacionados e distribuídos na Web. Nesse contexto, padrões de metadados recomendados pelo *World Wide Web Consortium* (W3C) vêm sendo utilizados para descrever e representar recursos multimídia, possibilitando ampliar os pontos de acesso e melhorar a gestão, a organização e a recuperação de acervos digitais. Entretanto, o relacionamento entre multimídia e a Web de dados ainda é um ramo de pesquisa que carece de estudos avançados voltados a tecnologias eficientes para geração, exposição, descobrimento e consumo de recursos multimídia semanticamente vinculados na Web (OSSENBRUGGEN; NACK; HARDMAN, 2004; NACK; OSSENBRUGGEN; HARDMAN, 2005; SCHANDL *et al*. 2011, SILVA, SOUZA, 2014).

Pesquisas têm sido desenvolvidas progressivamente nos campos das Ciências da Informação e da Computação, visando a estudos sobre a problemática do excesso de informações e sua organização, com o objetivo de melhorar a eficácia dos sistemas de recuperação de informação. Citam-se algumas pesquisas nessa perspectiva voltadas à exploração semântica da informação, tais como: a) Web Semântica e sua proposta emergente de dados interligados ou *Linked Data*, que intencionam criar metodologias, tecnologias e padrões de metadados para aumentar o escopo da interoperabilidade e da integração plena de informações heterogêneas entre sistemas de informação (BERNERS-LEE; HENDLER; LASSILA, 2001; BIZER; HEATH; BERNERS-LEE, 2009); b) instrumentos de

representação de relacionamentos semânticos e conceituais como ontologias (GUARINO 1998; ALMEIDA, 2013; SOERGEL, 2017) e vocabulários controlados (ANSI, 2005) objetivando endereçar problemas relacionados à interoperabilidade de sistemas e bases de dados, além das dificuldades intrínsecas à manipulação da linguagem natural como, por exemplo, as questões de polissemia e sinonímia; e c) modelos conceituais, de referência e ontológicos que orientam a modelagem da realidade documental e o processo de busca e recuperação da informação em contextos digitais como os *Functional Requirements for Bibliographic Records* - FRBR (IFLA, 2009); o *International Committee for Documentation/Conceptual Reference Model* - CIDOC CRM (LE BOEUF, 2018); a *Multimedia Metadata Ontology* - M3O (SAATHOFF; SCHERP, 2010); e o *Europeana Data Model* – EDM (EUROPEANA, 2017).

2. Objetivos

Durante as últimas décadas, surgiram várias iniciativas na produção de ontologias baseadas em RDF/OWL voltadas a descrever dados multimídia (SILVA; SOUZA 2014; LEMOS; SOUZA, no prelo) cujos esforços objetivaram transformar padrões de metadados multimídia, como o MPEG-7 (SALEMBIER, 2001; SALEMBIER; SMITH, 2001; MARTÍNEZ, 2004), em formatos semelhantes a ontologias. Nessa perspectiva, este artigo tem como objetivo apresentar o resultado de um estudo sistemático sobre iniciativas de padrões de metadados, vocabulários, modelos e ontologias voltados ao domínio da descrição multimídia, que culminou na obtenção de um *ranking* de ontologias a partir de uma análise comparativa e uma avaliação criteriosa sobre dimensões concernentes a reúso de recursos de conhecimento disponíveis na Web de dados.

3. Procedimentos Metodológicos

A pesquisa foi classificada como sendo de natureza qualitativa e quantitativa, de caráter exploratório e descritivo à luz de literatura científica já publicada e material empírico específico, o que a torna bibliográfica e documental.

O primeiro passo foi procedido por um estudo de domínio envolvendo fontes documentais, incluindo normas, artigos e bibliotecas de esquemas *Extensible Markup Language* - XML relacionados a padrões para descrição de documentos multimídia. Um conjunto de requisitos funcionais para o domínio multimídia foi organizado após o procedimento de análise do domínio e serviu como base para identificar, analisar e comparar ontologias para descrição multimídia no aspecto de características concernentes a padrões de metadados consolidados nas comunidades de Biblioteca Digital, Web Semântica e Multimídia.

O segundo passo foi o de adotar um guia metodológico atual, testado e validado em diferentes domínios e áreas que seguisse diretrizes para a construção de ontologias em rede *Linked Open Data* – LOD. Para tal, realizou-se uma revisão na literatura da área de Engenharia de Ontologias, e, dentre um conjunto de propostas aventadas (SILVA; SOUZA; ALMEIDA, 2008; SUÁREZ-FIGUEROA; GÓMEZ-PÉREZ; FERNÁNDEZ-LÓPEZ, 2012), selecionou-se o guia *NeOn Methodology* por dispor de práticas em iniciativas LOD e ser oriundo de *frameworks* metodológicos amplamente aceitos em áreas maduras como Engenharia de Software e Engenharia do Conhecimento. Buscou-se, então, a partir das orientações do guia, identificar ontologias multimídia fazendo um levantamento na literatura e buscas em repositórios da Web Semântica. Após um processo de refinamento frente às ontologias previamente selecionadas para análise, nove ontologias foram selecionadas, a saber: Media Ontology (STEGMAIER *et al*. 2009); M3 Multimedia (ATEMEZING, 2011); *Multimedia Metadata Ontology* - M3O (SAATHOFF; SCHERP, 2010); *Bootstrapping Ontology Evolution with Multimedia Information* – Boemie (DASIOPOULOU *et al*., 2008); *Core Ontology for Multimedia* – COMM (ARNDT *et al*., 2009); Polysema MPEG-7 MDS (VALKANAS; TSETSOS; HADJIEFTHYMIADES, 2007); MPEG-7 de Hunter (HUNTER, 2001); SmartWeb (OBERLE *et al*., 2007); e Rhizomik (GARCÍA; CELMA, 2005).

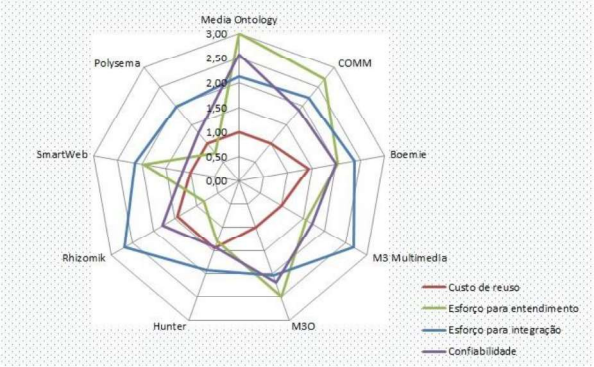
O terceiro e último passo foi conduzido por uma análise das ontologias selecionadas em que seus conteúdos (códigos) e documentações subjacentes foram inspecionados e analisados. Os critérios determinados para analisar e avaliar as ontologias foram em sua maioria oriundos do guia, os quais foram originados de casos de uso em diversas experiências de projeto envolvendo desenvolvimento e reúso de ontologias. A organização desses critérios se deu em quatro dimensões elucidadas como se segue: i) custo de reúso: estimativa de custos relacionados a tempo e a economia necessários ao reúso da ontologia avaliada;
ii) esforço para entendimento: estimativa de esforços necessários para entendimento do conteúdo da ontologia avaliada; iii) esforço para integração: estimativa de esforços empreendidos para integrar a ontologia avaliada à ontologia que está sendo construída; e iv) confiabilidade: análise da confiança em relação à ontologia avaliada frente a aspectos de tratamento semântico nas declarações (ex. axiomas presentes; recursos de conhecimento utilizados), avaliação (ex. testes disponíveis) e projetos renomados que fazem uso.

O método para a obtenção das pontuações para cada ontologia se deu por média ponderada envolvendo pesos determinados e valores mensurados para os critérios. Para este último, a escala de valores foi determinada de 0 a 3 correspondendo, em sequência, aos qualificadores (D)esconhecido, (B)aixo, (M)édio e (A)lto. Nesse sentido, a pontuação resultante permaneceu sempre numa escala de 0 a 3.

4. Resultados

A partir da análise comparativa realizada tornou-se possível delinear considerações relevantes sobre as dimensões para reúso frente às ontologias estudadas. O Gráfico 1 a seguir apresenta uma visão panorâmica e comparativa envolvendo as quatro dimensões.

Gráfico 1 - Comparação das dimensões para reúso



Fonte: elaborado pela autora.

A dimensão *custo de reúso* como dimensão que influencia negativamente o *ranking* para reúso se manteve estável para a maioria das ontologias e, portanto, sem considerável influência na pontuação final destas. O aspecto custo econômico de uma forma geral foi avaliado como baixo pelo fato de o acesso as nove ontologias ter ocorrido de maneira gratuita por meio de repositórios indicados na literatura ou de *links* apontados por máquinas de busca da Web Semântica. Já o aspecto tempo requerido variou entre baixo e médio. As ontologias avaliadas com tempo baixo para acesso e abertura no Protégé (editor de ontologias utilizado) foram prontamente analisadas. As ontologias do projeto Boemie, a MPEG-7 Hunter e a MPEG-7 Rhizomik foram avaliadas com valor médio em função de alguns impasses no acesso às suas bases de conhecimento.

A dimensão *esforço para entendimento* foi a dimensão que se apresentou com pontuações mais baixas para as ontologias analisadas. Contribuíram para essa realidade a Polysema MPEG-7, a MPEG-7 Rhizomik, a MPEG-7 Hunter e a M3 Multimedia, em geral por motivos de escassez de fontes documentais e/ou ausência de anotações, ou mesmo sem contribuição semântica, nos elementos de suas estruturas. Tal constatação as desfavorecem no aspecto consumo de tempo para se conseguir entender seus propósitos, escopos e conceituações visando alinhamentos consistentes. Por outro lado, a Media Ontology se destacou com pontuação alta em todos os critérios (qualidade da documentação, disponibilidade de conhecimento externo, claridade no código e

anotações na terminologia compatibilizada). Acredita-se que por ser uma proposta oriunda de um grupo de pesquisa do W3C (*Media Annotation Working Group*) especializado em questões de anotação semântica de mídias na Web, a equipe envolvida buscou empreender esforços na produção e na disponibilização de documentos concernentes ao conhecimento da ontologia. O mesmo procedimento ocorreu para a clarificação do código da ontologia, incluindo organização taxonômica favorável com conceitos delimitados e declarações conceituais adequadas para a maioria de seus elementos ontológicos, facilitando, assim, a interpretação semântica por parte do ontologista envolvido na análise. Em decorrência à qualidade da proposta, comprovada nos resultados aqui presentes, informações relevantes a respeito da Media Ontology são encontradas também em outros projetos que praticaram reúso com a sua estrutura, como foram os casos da M3 Multimedia e da M3O.

A dimensão *esforço para integração* foi a mais bem qualificada em comparação com as outras duas com influência positiva no *ranking* e também a que se manteve geometricamente mais estável. Isso reforça que o método aplicado na seleção de ontologias para descrição multimídia a compor o corpus da pesquisa foi bem sucedido para uma dimensão que contempla um aspecto importante relacionado à cobertura de requisitos funcionais.

A dimensão *confiabilidade* pode ser considerada uma característica presente na maioria das ontologias analisadas pelas seguintes constatações: i) todas possuem uma equipe de desenvolvimento com boa reputação; ii) todas são assistidas por entidades importantes no cenário mundial, tais como W3C, *European Commission*, *German Federal Ministry of Education and Research*, conceituadas universidades europeias e renomados centros de pesquisa; e iii) grande parte (M3O, COMM, Boemie, M3 Multimedia, Rhizomik) se propôs a disponibilizar ricas axiomatizações em suas conceituações, as quais são fundamentadas, na maioria dos casos, em ontologias de alto nível, em padrões de projeto multimídia, e no padrão de metadados multimídia MPEG-7.

Finalmente, o *ranking* de ontologias, produto da análise comparativa, evidenciou as ontologias mais proeminentes para o domínio em estudo, a saber, e nesta ordem: Media Ontology (**1,56**); M3O (**1,23**); COMM (**1,19**); e M3 Multimedia (**0,95**). A partir das constatações de características multimídia concernentes a cada uma delas, tornar-se-ia possível selecionar os recursos de conhecimento provenientes de suas estruturas e propor um modelo conceitual de referência para a organização e representação desse tipo de domínio. O sentido de "referência" é o que caracteriza o modelo como um artefato subjacente a esforços multidisciplinares de pesquisas voltados a modelos e tecnologias para processamento de metadados multimídia.

5. Considerações Finais

Um problema comumente verificado nas instituições que fazem uso de acervos em rede das mais variadas naturezas está no tratamento integrado das bases de dados heterogêneas e na ausência de padronização nos formatos de descrição. Tal prática culmina em situações problemáticas para os sistemas de recuperação da informação como, por exemplo: i) busca feita por palavras isoladas e descontextualizadas, o que dificulta maior visibilidade do acervo sob a ótica dos usuários e, consequentemente, dos mecanismos de busca; ii) falta de contexto nos itens midiáticos descritos (por exemplo, como fotos e vídeos se relacionam com o texto?); iii) ambiguidade conceitual (de qual conceito precisamente está se falando?); e iv) pouca relevância para o recurso recuperado.

A comparação de várias propostas de ontologias no domínio da descrição multimídia frente a padrões de metadados ISO evidenciou características relevantes que podem e devem ser descritas para melhor recuperação de recursos multimídia, principalmente no contexto da Web. A necessidade de integração semântica e disponibilização global de recursos multimídia na rede é um propósito comum entre as propostas de ontologias pesquisadas.

Referências

ALMEIDA, M. B. Revisiting ontologies: a necessary clarification. **Journal of the American Society of Information Science and Technology**, [S.l.], v. 64, n. 8., p. 1682-1693, 2013.

ANSI/NISO Z39.19-2005 (R2010). **Guidelines for the construction, format, and management of monolingual controlled vocabularies**. Baltimore: NISO Press, 2005. 184 p.

ARNDT, R. *et al*. **COMM:** a core ontology for multimedia annotation. 2009. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.154.5510>. Acesso em: 05 set. 2019.

ATEMEZING, Ghislain Auguste. **Analyzing and ranking multimedia ontologies for their reuse**. 2011. Tesis (Master) - Facultad de Informática,Universidad Politécnica de Madrid, Madrid, 2011.

BERNERS-LEE, T; HENDLER, J.; LASSILA, O. The semantic web. **Scientific American**, [S.l.], v. 284, n. 5, p. 34-43, May 2001.

BIZER, C.; HEATH, T.; BERNERS-LEE, T. Linked Data - the story so far. **International Journal on Semantic Web and Information Systems**, [S.l.], v. 5, n. 3, p. 1-22, 2009.

DASIOPOULOU, S. **Multimedia content and descriptor ontologies**: final version. 2008. Disponível em: <https://www.academia.edu/2721370/Multimedia_content_and_descriptor_ontologies-final_version>. Acesso em: 05 set. 2019.

EUROPEANA. **Definition of the Europeana Data Model v5.2.8**. 2017. Disponível em: < https://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation//EDM_Definition_v5.2.8_102017.pdf > Acesso em: 05 set. 2019.

GARCÍA, R.; CELMA, O. semantic integration and retrieval of multimedia metadata. In: INTERNATIONAL WORKSHOP ON KNOWLEDGE MARKUP AND SEMANTIC ANNOTATION, 5th, 2005, Galway. **Proceedings...** Galway, 2005, p. 69–80.

GUARINO, N. **Formal ontology in information systems**. 1998. Disponível em: <http://citeseer.ist.psu.edu/viewdoc/download;jsessionid=E88DA9B5B5A9797C83C1F2E3C907991F?doi=10.1.1.29.1776&rep=rep1&type=pdf>. Acesso em: 05 set. 2019.

HUNTER, J. Adding multimedia to the semantic web – building an MPEG-7 ontology. In: INTERNATIONAL SEMANTIC WEB WORKING SYMPOSIUM, 1st, 2001, Stanford. **Proceedings...** Disponível em: <https://files.ifi.uzh.ch/ddis/iswc_archive/iswc/ih/SWWS-2001/program/full/paper59a.pdf>. Acesso em: 05 set. 2019.

INTERNATIONAL FEDERATION OF LIBRARY ASSOCIATIONS AND INSTITUTIONS (IFLA). **Functional requirements for bibliographic records**. 2009. 142 p. Disponível em: < https://www.ifla.org/files/assets/cataloguing/frbr/frbr_2008.pdf >. Acesso em: 05 set. 2019.

LE BOEUF, Patrick *et al*. (Ed.). **Definition of the CIDOC Conceptual Reference Model**: version 6.2.3: International Council of Museums (ICOM); International Committee for Documentation(CIDOC), 2018. Disponível em: <http://www.cidoc-crm.org/Version/version-6.2.3-0>. Acesso em: 05 set. 2019.

LEMOS, D.L. da S.; SOUZA, R. R.. Ontologias na representação de documentos: um panorama atual para descrição de conteúdo multimídia em rede. **Informacao &**

Sociedade-Estudos. No prelo 2019.

MARTÍNEZ, J. M. **MPEG-7 overview (version 10)**. 2004. Disponível em: < https://mpeg.chiariglione.org/standards/mpeg-7 >. Acesso em: 05 set. 2019.

NACK, F.; OSSENBRUGGEN, J.V.; HARDMAN, L.H. That obscure object of desire: multimedia metadata on the web -part 2. **IEEE MultiMedia**, [S.l.], v.. 12, n. 1 , p. 54-63, 2005.

OBERLE, D. *et al*. On foundational and domain models in the smartweb integrated ontology (SWIntO). **Journal of Web Semantics**., [S.l.], v. 5, n. 3, p. 156-174, Sept. 2007.

OSSENBRUGGEN, J. V.; NACK, F.; HARDMAN, L. H. That obscure object of desire: multimedia metadata on the web - part 1. **IEEE MultiMedia**, [S.l.], v. 11, n. 4, p. 38-48, Oct./Dec. 2004.

SAATHOFF, C.; SCHERP, A. Unlocking the semantics of multimedia presentations in the web with the multimedia metadata ontology. In: INTERNATIONAL CONFERENCE ON WORLD WIDE WEB, 19th, 2010, Raleigh. **Proceedings...**New York: ACM, 2010. p. 831-840.

SALEMBIER, P. Overview of the MPEG-7 standard and of future challenges for visual information analysis. **EURASIP Journal on Advances in Signal Processing**, New York, v. 2002, n. 2, p. 343-353, Apr. 2002.

SALEMBIER, P.; SMITH, J. MPEG-7 multimedia description scheme**. IEEE Transactions on Circuits and Systems for Video Tecnology**, [S.l.], v. 11, n. 6, June 2001.

SCHANDL, B. *et al*. Linked Data and multimedia: the state of affairs. **Multimedia Tools and Applications**, [S.l.], online first, p. 1-34, 2011.

SILVA, D.L. da; SOUZA, R. R.; ALMEIDA, M. B. Ontologias e vocabulários controlados: comparação de metodologias para construção. **Ciência da Informação**, Brasília, v. 37, n.3, p. 60-75, set./dez. 2008.

SILVA, D.L. da ; SOUZA, R. R . Representação de documentos multimídia: dos metadados às anotações semânticas. **Tendências da Pesquisa Brasileira em Ciência da Informação**, v. 9, n.2, p. 1-22, 2014.

SOERGEL, Dagoberto (Org.). Ontologias na ciência da informação: estado da arte no Brasil. **Ciência da Informação**, Brasília, DF, v.46, n.1, p.1-227, jan./abr. 2017.

STEGMAIER, F. et al. How to align media metadata schemas? design and implementation of the media ontology. In: INTERNATIONAL CONFERENCE ON SEMANTIC AND DIGITAL MEDIA TECHNOLOGIES, 4th, 2009, Graz. **Proceedings…** [S.l.]: CEUR-WS.org, 2009. Workshop on semantic multimedia database technologies (SeMuDaTe 2009).

SUÁREZ-FIGUEROA, M. C.; GÓMEZ-PÉREZ, A.; FERNÁNDEZ-LÓPEZ, M. The NeOn methodology for ontology engineering. In: SUÁREZ-FIGUEROA, M. C. et al. (Ed.). **Ontology Engineering in a Networked World**. Berlin: Springer, 2012. p. 9-34.

VALKANAS, G.; TSETSOS, V.; HADJIEFTHYMIADES, S. The polysema MPEG-7 video annotator. In: INTERNATIONAL CONFERENCE ON SEMANTICS AND DIGITAL MEDIA TECHNOLOGIES, 2nd, 2007, Genova. **Proceedings...** Berlin: Springer, 2007.

OS ACERVOS CULTURAIS BRASILEIROS NO REPOSITÓRIO WIKIMEDIA COMMONS: A VISUALIZAÇÃO DAS COLEÇÕES DE MUSEUS DO INSTITUTO BRASILEIRO DE MUSEUS

The Brazilian Cultural Collections in Wikimedia Commons Repository: *The visualization of the museum collections of the Brazilian Museum Institute*

Danielle do Carmo[(1)], Dalton Lopes Martins[(2)]
(1) Universidade Federal de Goiás, Avenida Esperança s/n, Câmpus Samambaia, Goiânia- GO, docarmo.danielle@gmail.com.
(2) Universidade de Brasília, Campus Universitário Darcy Ribeiro s/n, Asa Norte, Brasília – DF, dmartins@gmail.com.

Resumo:
O presente trabalho apresenta resultados preliminares de uma pesquisa exploratória e descritiva que tem como objetivo compreender o alcance das mídias relativas à coleções de acervos de instituições culturais no repositório Wikimedia Commons. A ferramenta GLAMorgan foi utilizada para a coleta de dados acerca da quantidade de visualização de arquivos de mídias, os dados são relativos aos 12 meses do ano de 2018. Desse modo foi possível obter dados quantitativos sobre a visualização de arquivos de mídia que estavam em categorias referentes às coleções de sete museus que estão sob administração direta do Instituto Brasileiro de Museus (Ibram), sendo que a coleção do Museu de Belas Artes gerou no ano de 2018, mais de 33 milhões de visualizações e uma média mensal de quase 3 milhões de visualizações. Os resultados obtidos demonstram que a plataforma Wikimedia Commons se revela como meio potencial de disseminação de acervos para as instituições culturais.

Palavras-chave: Wikimedia Commons; Instituto Brasileiro de Museus; Repositório de mídias; Acervos culturais; Instituições Culturais.

Abstract:
The present work presents preliminary results of an exploratory and descriptive research that aims to understand the reach of media related to collections from cultural institutions in the Wikimedia Commons repository. The GLAMorgan tool was used for gathering quantitative data regarding media archives visualizations in the 12 months of 2018. Thus it was possible to identify data on the visualization of media files that were in categories referring to the collections of seven museums that are under Instituto Brasileiro de Museus (Ibram) direct administration. This way was possible to identify that the collection of Museu de Belas Artes generating in 2018 more than 33 million views and a monthly average of almost 3 million views. The results show that the Wikimedia Commons platform reveals itself as a potential mean of dissemination of collections for cultural institutions.

Keywords: Wikimedia Commons; Instituto Brasileiro de Museus; Media Repository; Cultural Collections; Cutural Institutions.

1. Introdução

O Wikimedia Commons é um repositório online que armazena e disponibiliza, de forma gratuita, diversos tipos de arquivos de mídias. É uma plataforma gerida pela organi- zação sem fins lucrativos Fundação Wikime- dia e alimenta os vários projetos da fun- dação como a Wikipédia, Wikitionaty, Wiki- books, Wikisource, Wikinews, Wikiversite, o Wikiquote, o Wikidata e outros. Lançado no ínicio de setembro de 2004, o repositório de mídia Wikimedia Commons disponibiliza atualmente mais de 55 milhões de arquivos de mídias. Os arquivos de mídias encontram- se sob licenças individuais que permitem a cópia, o uso e a modificação, de acordo com os termos especificados. O banco de dados do Wikimedia Commons, e seus textos, são licenciados sob a licença *Creative Commons Attribution / Share-Alike* (WIKIMEDIA COMMONS, 2019).

Do mesmo modo que outros projetos da fundação Wikimedia, o repositório é alimentado e mantido de forma coletiva e colaborativa por meio da ação de usuários voluntários. Segundo dados fornecidos pelo projeto, em setembro de 2019, a plataforma registrou um total de 35,285 usuários ativos e 161 *bots* que auxiliam na edição do conteúdo (WIKIMEDIA COMMONS, 2019).

Entre as coleções de mídias de diversos tipos e temáticas encontradas no Wiki-

media Commons é possível identificar conteúdos referentes a itens de coleções de diversas instituições culturais do mundo. Esses arquivos de mídia podem ser disponibilizados por usuários que realizaram *uploads* desse material de modo espontâneo, ou em alguns casos por meio de parcerias estabelecidas com instituições culturais. Segundo Stinson, Fauconnier e Wyat (2018, p.17) há um longo histórico de colaboração entre a comunidade Wikimedia e o conjunto de diferentes tipos de instituições culturais que são reunidas sob o termo guarda-chuva GLAM (acrônimo em inglês - *Galleries, Library, Archives and Museums* - usado para se referir a instituição galerias, bibliotecas, arquivos, museus e outras instituições do patrimônio cultural).

Segundo Zeinstra (2013) o conteúdo GLAM reunia cerca mais de dois milhões de objetos digitais o que corresponderia um total de 13,14% do conteúdo da Wikimedia Commons no ano de 2013.

> Em outras palavras, um em cada oito arquivos na Wikimedia Commons é disponibilizado através de alguma colaboração com GLAM´s. Tanto a GLAM coloca suas coleções em domínio público ou abre a licença de suas coleções online e voluntários realizam *upload,* ou alguma colaboração ativa entre GLAMs e a comunidade Wikimedia como Wiki Loves Monuments cria conteúdo GLAM. Dessa forma voluntários colocam esses objetos de mídia em artigos da Wikipédia criando um incrível aumento em sua visibilidade. Instituições contribuintes enxergam o potencial da Wikimedia como um canal de distribuição. (ZEINSTRA, 2013)

Nesse sentido, as autoras Vilaespesa e Navarete (2019) observam que o Google e a Wikipédia são ambientes de intensa utilização, que dão fácil acesso a coleções independente de onde elas estejam. Pois ao realizar uma busca utilizando os mecanismos Google e Google Imagens, ou por meio da busca por voz de assistentes virtuais como a Siri e Alexa é possível identificar, ainda nas primeiras páginas, conteúdos oriundos da Wikimedia Commons, Wikidata e Wikipédia. Esssa privilegiada exposição de conteúdos se daria devido o "aos ambientes estruturados fornecidos pelas plataformas Wiki que influenciaram fortemente os resultados da pesquisas" (VILLAESPESA, NAVARETTE, 2019). As autoras ainda atentam que devido ao atual comportamento de busca de informações pelos usuários somado aos algoritmos dos mecanismos de busca torna-se possível que o conteúdo de uma instituição cultural apareça em uma posição predominante.

Dessa forma, o ecossistema informacional composto pelo repositório de mídias Wikimedia Commons, a enciclopédia *online* Wikipédia e o repositório de dados estruturados Wikidata se apresentam como importantes aliados na socialização das informações referentes aos acervos das instituições culturais

2. Objetivos

A presente pesquisa tem como objetivo principal verificar o alcance, em termos de visualizações, dos arquivos de mídias referentes a objetos digitais de coleções de acervos culturais disponibilizados por meio repositório Wikimedia Commons. Para o presente trabalho serão apresentados dados coletados sobre os museus vinculados ao Instituto Brasileiro de Museus (Ibram).

3. Procedimentos Metodológicos

Para realizar a coleta de dados, foi utilizada a ferramenta, desenvolvida pelo wikimedista Magnus Manske, chamada GLAMorgan. Com essa ferramenta é possível obter dados de visualização das páginas dos arquivos das mídias do Wikimedia Commons com base nas suas categorias (META-WIKI, 2019). Para efeito do presente trabalho, utilizamos categorias que determinavam a origem dos itens, no caso as coleções de determinadas instituições. Nosso recorte busca verificados dados de visualização relativo à coleções museus que são vinculados ao Instituto Brasileiro de Museus (Ibram). Com a lista dos museus em mãos, realizamos a busca de categorias de suas coleções na lista na plataforma Wikimedia Commons, dessa forma foi possível obter as categorias e posteriormente utiliza-las para realizar as buscas na ferramenta GLAMorgan.

Os dados coletados refletem números sobre a visualização dos arquivos de mídias das coleções dos 12 meses do ano de 2018, de janeiro a dezembro. Os dados coletados foram planilhados, analisados e posteriormente descritos no presente artigo.

4. Resultados

Para verificar o alcance das coleções brasileiras disponibilizadas na Wikimedia Commons foram coletados dados sobre o número de visualizações de arquivos de mídia de coleções de instituições culturais brasileira, mais especificamente os museus vinculados ao Ibram. Como descrito no tópico anterior, realizamos o processo de busca por categorias que faziam referência a coleções dos museus elencados. Dos 30 museus sob administração direta do Ibram identificarmos categorias de coleções de nove museus, como pode ser observado no quadro 1.

Quadro 1. Categorias de coleções dos museus do Ibram

#	Museu	Categoria referente à coleção do museu
1	Museu da Inconfidência	Collections of the Museu da Inconfidência
2	Museu da República	Collections of the Museu da República
3	Museu Histórico Nacional	Collections of the Museu histórico Nacional
4	Museu Imperial	Collections of the Museu Imperial
5	Museu Nacional de Belas Artes	Collections of the Museu Nacional de Belas Artes
6	Museu do Açude (equipamento dos Museus Castro Maya)	Collections of the Museus Castro Maya
7	Museu Casa de Benjamin Constant	Media contributed by the Museu Casa de Benjamin Constant
8	Museu Regional de São João del-Rei	Collections of the Museu Regional de São João del-Rei
9	Museu Victor Meirelles	Collections of the Museu Victor Meirelles

Fonte: Dados da pesquisa, 2019.

Com a categoria da coleção foi possível obter os dados de visualização de coleções de nove museus no Wikimedia Commons por meio da ferramenta GLAMorgan. Das nove categorias de coleções de museus identificadas foi possível obter dados de sete delas. Ao consultar as categorias referentes às coleções do Museu Histórico Nacional e Museu Imperial, a ferramenta apresenta a mensagem "*Loading file data* " e não retornou resultados. Para os demais museus foi possível obter o total de visualizaçoes do ano 2018 e média de visitação mensal, os dados podem ser observados no quadro do apêndice A.

Segundo os dados coletados, foi possível observar que o museu que possuí um maior número de arquivos da árvore de sua categoria é o Museu Nacional de Belas Artes, com 1.237 arquivos, que no ano de 2018 receberam um total de 33.018.251 visualizações, equivalente a uma média de quase 3 milhões de visualizações por mês. O segundo e terceiro museus com o maior número de visualizações totais no ano de 2018 foram respectivamente o Museu da República com 2.790.915 visualizações no ano de 2018 e média de aproximadamente 215 mil visualizações por mês, e o Museu Castro Maia com 3.057.720 visualizações no ano e uma média de aproximadamente 254 mil visualizações por mês. Apesar da categoria da coleção do Museu Regional de São João del Rei apresentar 6 registros em sua árvore da categoria, os resultados obtido na consulta de visualizações retornou o número 0 nos resultados, como pode ser observado no apêndice A.

5. Considerações Finais

O presente trabalho teve como objetivo apresentar os resultados de uma pesquisa que tem como objetivo principal verificar o alcance, em termos de visualizações, dos arquivos de mídias referentes de coleções de instituições culturais brasileiras no repositório de arquivos de mídia da plataforma Wikimedia Commons a objetos digitais de coleções de

acervos culturais disponibilizados por meio repositório Wikimedia Commons. Como amostra para o presente trabalho foram apresentados dados coletados sobre os museus sob administração direta do Ibram.

Dessa forma foi possível identificar que dos 30 museus do Ibram, nove possuem categorias no Wikimedia Commons. Das nove categorias identificadas, foi possível coletar dados sobre a visualização de arquivos de mídia de coleções de sete museus.

Os números obtidos por meio da ferramenta GLAMorgan, permitiu a obtenção dos números totais de visualização do ano de 2018 e a média mensal de visualizações. Os números obtidos são impressionantes em termos de alcance de público, e reforça a ideia de que essa plataforma seja um meio potente de disseminação de acervos culturais. Dessa forma o repositório Wikimedia Commons, assim como seus projetos irmãos Wikidata e Wikipédia, se revela para as instituições culturais como um meio estratégico para a disseminação de mídias sobre seus acervos na internet.

Referências

MANSKE, Magnus. **GLAMorgan**. 2019. Disponível em: https://tools.wmflabs.org/glamtools/glamorgan.html?&category=Collections_in_Brazil&depth=12&year=2018&month=5. Acesso em 18 out. 2019

META-WIKI. **GLAMorgan**, 2019. Disponível em: https://commons.wikimedia.org/wiki/Main_Page. Acesso 15 de set. de 2019.

WIKIMEDIA COMMONS. **Main page**, 2019. Disponível em: https://commons.wikimedia.org/wiki/Main_Page . Acesso 15 de set. de 2019.

STINSON, Alexander D.; FAUCONNIER, Sandra; WYATT, Liam. **Stepping Beyond Libraries**:The Changing Orientation in Global GLAM-Wiki. JLIS.it 9, 3 set. 2018. Disponível em: https://www.jlis.it/article/download/12480/11333. Acesso em: 18 jul. 2019.

VILLAESPESA, Elena; NAVARRETE, Trilce. **Museum Collections on Wikipedia**: Opening Up to Open Data Initiatives. Boston, 2019. Disponível em: https://mw19.mwconf.org/paper/museum-collections-on-wikipedia-opening-up-to-open-data-initiatives. Acesso em: 18 jul. 2019.

ZEINSTRA, Maarten. **Report on Requirements for Usage and Reuse Statistics for GLAM content**. Kennisland, 2013. Disponível em: https://www.kl.nl/wp-content/uploads/2014/09/report-on-requirements-for-usage-and-reuse-statistics-for-glam-content.pdf. Acesso em: 20 de setembro de 2019.

7. Apêndice A - Número de visualizações por coleção de museu.

#	Museu	Categoria referente à coleção	Arquivos na árvore da categoria	Total de visualizações dos arquivos no ano de 2018	Média mensal de visualização de arquivos de mídia no ano de 2018
1	Museu da Inconfidência	Collections of the Museu da Inconfidência	31	2.018.081	168.173,4
2	Museu da República	Collections of the Museu da República	31	2.790.915	232.576,2
3	Museu Nacional de Belas Artes	Collections of the Museu Nacional de Belas Artes	1.237	33.018.251	2.751.520,9
4	Museu do Açude (equipamento dos Museus Castro Maya)	Collections of the Museus Castro Maya	111	3.057.720	254.810
5	Museu Casa de Benjamin Constant	Media contributed by the Museu Casa de Benjamin Constant	31	3.632	302,6
6	Museu Regional de São João del-Rei	Collections of the Museu Regional de São João del-Rei	6	0	0
7	Museu Victor Meirelles	Collections of the Museu Victor Meirelles	84	551.482	45.956,8

Fonte: Dados da pesquisa, 2019.

PROPOSTA DE APLICAÇÃO DA FUSÃO DE DADOS E INFORMAÇÕES NO APOIO À PREVENÇÃO DE ACIDENTES DE TRÂNSITO NAS RODOVIAS FEDERAIS BRASILEIRAS

PROPOSAL OF APPLICATION OF THE DATA AND INFORMATION FUSION TO SUPPORT THE PREVENTION OF TRAFFIC ACCIDENTS ON BRAZILIAN FEDERAL HIGHWAYS

Jordan Ferreira Saran[1], Ronnie Shida Marinho[1], Clayton Martins Pereira[1], Leonardo Castro Botega[1], José Eduardo Santarem Segundo[2]

(1) Universidade Estadual Paulista (UNESP), Faculdade de Filosofia e Ciências, Marília – SP, {jordan.saran, ronnie.shida, clayton.martins, leonardo.botega}@unesp.br

(2) Universidade de São Paulo (USP), Faculdade de Filosofia, Ciências e Letras, Ribeirão Preto – SP, santarem@usp.br

Resumo:
A responsabilidade pela coleta e disponibilização de dados oficiais sobre acidentes, infrações e condições nas rodovias federais brasileiras está descentralizada em diferentes órgãos públicos do Poder Executivo Federal. Como efeito desta descentralização, não há uma unificação das bases de dados para o planejamento de ações de fiscalização e de prevenção de acidentes nessas vias. Por outro lado, os modelos de fusão de dados e informações surgem como forma de orientar os processos de desenvolvimento de sistemas para a aquisição, inferência, avaliação e representação de informações situacionais de alto nível. Nesse contexto surge o presente trabalho de pesquisa, que tem como objetivo auxiliar o planejamento de ações de prevenção de acidentes de trânsito em rodovias federais, ao propor a aplicação dos métodos da fusão de dados e informações, a partir de dados coletados de forma semiautomática de órgãos governamentais, e de informações inseridas manualmente pelos agentes envolvidos. Assim, ao utilizar a fusão de dados na aplicação proposta neste artigo, espera-se obter representações mais completas, precisas e enriquecidas com dados de múltiplas fontes de acidentes de trânsito. Almeja-se com isso minimizar problemas de acidente de trânsito nas rodovias federais brasileiras.
Palavras-chave: Prevenção de Acidentes de Trânsito; Fusão de Dados e Informações; Rodovias Federais Brasileiras.

Abstract:
The responsibility for the collection and availability of official data on accidents, infractions and conditions on the Brazilian federal highways is decentralized in different public agencies of the Federal executive branch. As a result of this decentralization, there is no unification of the databases for the planning of surveillance and accident prevention actions in these ways. On the other hand, data and information fusion models emerge as a way to guide the processes of systems development for the acquisition, inference, evaluation and representation of high-level situational information. In this context arises the present research work, which aims to assist the planning of actions to prevent traffic accidents on federal highways, through the application of methods of merging data and information, from data collected from Government agencies, and information entered manually by the agents involved. Thus, by using data fusion in the application proposed in this article, we expect to obtain more complete, accurate and enriched representations with data from multiple sources of traffic accidents. It aims to minimize traffic accident problems on Brazilian federal highways.
Keywords: Prevention of Traffic Accidents; Data and Information Fusion; Brazilian Federal Highways.

1. Introdução

De acordo com a Organização Pan-Americana de Saúde (OPAS, 2009), cerca de 1,35 milhão de pessoas no mundo morrem, a cada ano, em decorrência de acidentes de trânsito. Preocupados com estes números alarmantes, a Organização das Nações Unidas (ONU) em sua "Agenda 2030 para o Desenvolvimento Sustentável" fixou uma meta ambiciosa quanto à segurança no trânsito, que consiste em reduzir pela metade, até o ano de 2020, o número de mortos e feridos por acidentes de trânsito em todo o planeta.

Dentre os principais fatores relacionados à ocorrência dos acidentes de trânsito estão: os erros humanos; o excesso de velocidade; a condução de veículos sob influência de álcool e outras substâncias; a distração ao dirigir; a falta de segurança da infraestrutura viária e dos veículos; a falta de obediência às normas e leis de trânsito; entre outros (OPAS, 2009).

Consoante aos fatores e, considerando especificamente as rodovias federais brasileiras, a responsabilidade pela coleta e disponibilização de dados oficiais sobre acidentes, infrações e condições nestas vias está descentralizada em diferentes órgãos públicos do Poder Executivo Federal.

Como efeito desta descentralização, não há uma unificação das bases de dados para o planejamento de ações de fiscalização e de prevenção de acidentes nas rodovias federais brasileiras. Neste sentido, diante deste complexo cenário que envolve as ocorrências de acidentes de trânsito, a formulação de políticas públicas e de campanhas educativas, e a adoção de melhorias na segurança das vias, necessitam do emprego de elementos de suporte ao conhecimento (DNIT, 2014).

Por outro lado, os modelos de fusão de dados e informações surgem como forma de orientar os processos de desenvolvimento de sistemas para a aquisição, inferência, avaliação e representação de informações situacionais de alto nível. Tais sistemas são alimentados por várias fontes de dados heterogêneas, na maioria dos casos de forma dinâmica, tal como seria no caso das rodovias federais brasileiras.

Nesse contexto surge o presente trabalho de pesquisa, que tem como objetivo auxiliar o planejamento de ações de prevenção de acidentes de trânsito em rodovias federais, por meio da proposta de aplicação dos métodos da fusão de dados e informações, a partir de dados coletados de forma semiautomática de órgãos governamentais, e de informações inseridas manualmente pelos agentes envolvidos.

A metodologia empregada neste trabalho, de natureza qualiquantitativa e de tipo descritivo, consiste primeiramente na realização de uma pesquisa bibliográfica, de forma a permitir uma breve revisão sobre os modelos de fusão de dados e informações, bem como estudar os trabalhos que seguem a mesma linha de pesquisa deste projeto. Por fim, foi realizada uma análise documental, onde foram obtidos fontes, dados e estatísticas sobre acidentes de trânsito no Brasil, além dos elementos necessários para a proposta de aplicação da fusão de dados e informações na prevenção de acidentes de trânsito nas rodovias federais brasileiras.

2. Acidentes de trânsito nas rodovias federais brasileiras

No Brasil, de acordo com dados do Ministério da Infraestrutura, a malha rodoviária federal possui atualmente 75.800 Km de extensão total, sendo que 65.400 Km são de rodovias pavimentadas e 10.400 Km de rodovias ou trechos não pavimentados. A maior parte desta malha está concentrada nas regiões sul e sudeste do país, sendo o estado de Minas Gerais o que possui a maior extensão de rodovias federais em sua área territorial.

De acordo com dados do IPEA (2015), nesta extensa malha rodoviária federal ocorrem aproximadamente 20% das mortes em acidentes de trânsito registradas anualmente no país, além de deixar cerca de 26 mil feridos graves por ano, com fortes impactos para o governo e para as famílias dos acidentados. Quanto à apuração das causas dos acidentes de trânsito em rodovias federais, estas são analisadas sob três fatores: 1) Fator Humano: subavaliação da probabilidade de acidente; desatenção; cansaço; deficiências (visual, auditiva, motora); consumo de álcool; consumo de drogas; excesso de velocidade; desrespeito à distância mínima entre veículos; ultrapassagem indevida; não-uso de cinto ou de capacete; imprudência de pedestres, ciclistas e motociclistas; 2) Fator Veículo: violência do choque; defeitos de manutenção; utilização incorreta; 3) Fator Infraestrutura: condições de conservação da via; mudanças do contexto da rodovia; evolução do tráfego; condições meteorológicas.

Atualmente, os dados utilizados para esta análise de acidentes de trânsito, sob os fatores humano e veículo, são obtidos exclusivamente da Polícia Rodoviária Federal (PRF) a partir das seguintes bases: 1) Acidentes de Trânsito em Rodovias Federais; e 2) Infrações de Trânsito em Rodovias Federais.

Já os dados que possibilitam a análise de acidentes de trânsito sob o fator infraestrutura pode ser obtidos a partir do Departamento Nacional de Infraestrutura de Transportes (DNIT), que fornece informações como dados de contagem de tráfego e estatísticas de Acidentes.

Ainda sob o fator infraestrutura, os dados geográficos, meteorológicos e pluviométricos podem ser obtidos a partir do Centro Nacional de Monitoramento e Alertas de Desastres Naturais (CEMADEN).

Quanto às ações de fiscalização e às campanhas de prevenção de acidentes, dado a enorme extensão da malha rodoviária federal, que é distribuída em um país de dimensões continentais como o Brasil, estas acabam sendo pontuais (locais com maior incidência de acidentes) e restritas aos centros urbanos das vias mais importantes. Muitas vezes estas ações são planejadas sem contar com um sistema de apoio à tomada de decisão que unifique as informações provenientes desta diversidade de bases de dados descentralizadas.

Assim, este artigo pretende apresentar uma representação que minimize tais lacunas apresentadas. Para tal, propõe a aplicação de um processo de fusão de dados e informações, pelo qual seja possível integrar, por meio de técnicas apropriadas, diferentes dados oriundos de órgãos como PRF, DNIT e CEMADEN, com o objetivo de disponibilizar uma ferramenta para suporte às ações de fiscalização e às campanhas de prevenção de acidentes nas rodovias federais brasileiras.

3. Fusão de dados e informações

Em um primeiro instante, a definição de Fusão de Dados era dada como um conjunto de métodos para lidar com dados e informações de uma ou mais fontes, realizando associações, correlações e combinações múltiplas entres elas, com o intuito de obter estimativas mais precisas sobre um determinado objeto e possíveis relações entre estes, sendo o processo um ciclo contínuo de estimativas e validações, assim como a aplicabilidade de fontes externas adicionais ou alterações no processo (HALL e JORDAN, 2010).

Entretanto, para Blasch (2013) a Fusão de Dados e Informações é aplicada como sistema para apoiar a avaliação de situações e a tomada de decisão em sistemas complexos. Além disso, possibilita ainda a redução da dimensionalidade dos dados, a agregação de valor à informação, o aumento da representatividade e a produção de subsídios para a construção do conhecimento sobre situações.

Em razão do crescimento do fluxo de dados e informações, da computação e da tecnologia de sensores, a área de fusão vem se tornando um importante objeto de estudo interdisciplinar. Inúmeras técnicas para combinar dados têm sido exploradas em diferentes áreas tais como: inteligência artificial, processamento de sinais, teoria de controle, entre outras.

No entanto, como forma de auxiliar no avanço da área de fusão, ao realizar a aderência a outras áreas, o modelo *Joint Directors of Laboratories* (*JDL*) foi o primeiro e mais clássico a surgir com o intuito de normalizar e difundir o conhecimento sobre as bases dos sistemas de Fusão de Dados. Devido sua extrema facilidade de compreensão e padronização para diversos problemas, auxiliando também na avaliação da relevância da solução, o modelo *JDL* é o mais conhecido e amplamente aplicado para demonstrar a utilidade da Fusão de Dados como suporte à tomada de decisão (LLIANS, 2004; HALL e MACMULLEN, 2004).

Este modelo é composto por diferentes módulos e funções que dão suporte e possibilitam a fusão, dentre os quais: preparação e processamento das informações; avaliação de objetos e situações; refinamento das informações e análise de impacto e risco; importação de dados por meio de fontes externas.

O modelo *JDL* apresenta níveis de processos para acomodar e executar os módulos e funções citados acima, sendo eles: Pré-processamento (Nível 0), Identificação e Refinamento de Objetos (Nível 1), Elaboração e Refinamento da Situação (Nível 2), Definição e Refinamento de Risco (Nível 3) e Refinamento do Processo (Nível 4) e Interface Humano-Computador (Nível 5) (STEINBERG et al., 1999). Mais adiante, na seção 5, o modelo *JDL* será descrito de forma mais detalhada.

Assim, ao aplicar a fusão de dados proposta neste artigo, espera-se obter representações mais apuradas dos acidentes de trânsito em rodovias federais brasileiras, a partir do uso de dados disponibilizados pelos seguintes órgãos: PRF, DNIT e CEMADEN. Almeja-se com isso minimizar a ocorrência de

acidentes de trânsito nestas rodovias. A seguir são apresentados alguns trabalhos que, de alguma forma, relacionam-se com a proposta deste artigo.

4. Trabalhos relacionados à aplicação da fusão de dados e informações na prevenção de acidentes de trânsito

Diversos esforços têm sido realizados com o intuito de amenizar as causas dos acidentes de trânsitos e beneficiar a segurança rodoviária e veicular. Ryder et al. (2017) desenvolveram um sistema de apoio à decisão para veículos, que faz uso de alertas aos condutores informando o risco do trajeto a ser percorrido com base no histórico de acidentes do local. Para isso, no momento em que o condutor se aproxima de um local onde já tenha sido registrada uma ocorrência de trânsito, automaticamente esse condutor é notificado pelo sistema para que diminua a velocidade e tenha prudência naquele trecho. Diferentemente do trabalho descrito acima, o trabalho proposto por Abulatif (2018) apresenta um processo de integração de diversas fontes de dados relacionadas à acidentes de trânsito, de modo que os dados gerados por tal processo sejam utilizados para subsidiar a elaboração de ações de segurança viária em cinco capitais brasileiras: Belo Horizonte/MG, Campo Grande/MS, Curitiba/PR, Palmas/TO e Teresina/PI. Para tanto, foi realizada a integração de dados do Sistema de Informação sobre Mortalidade (SIM) e do Sistema de Informação Hospitalares (SIH), ambos mantidos pelo Ministério da Saúde.

Em uma abordagem diferente, Sohn et al. (2003) fizeram uso da análise de dados, por meio de algoritmos de agrupamento e fusão de dados, para verificar o nível de gravidade dos acidentes de trânsito em rodovias da Coreia do Sul. Para tal, utilizaram dados como a largura da via, a velocidade antes do acidente, o formato do carro, além da verificação da existência de dispositivos de proteção nos veículos.

Além da utilização de algoritmos de agrupamento e fusão de dados para a descrição de acidentes de trânsito, pode-se citar o trabalho de Sun et al. (2018), onde os autores tiveram a ideia de prever a ocorrência de acidentes de trânsito por meio da conscientização de incidentes e anomalias. Para tal, utilizam dados reais de uma estrada chamada Kunshi, onde são coletados registros como: taxa de fluxo, velocidade média, entre outras informações. Mediante esses dados é proposto um método com base nos vizinhos mais próximos, chamado *k-Nearest Neighbors* (*kNN*), muito utilizado para classificação de dados em aprendizado de máquina.

Costa *et al.* (2014) propuseram a aplicação de mineração de dados nos boletins de ocorrências registrados nas rodovias federais brasileiras pela Polícia Rodoviária Federal no ano de 2012, com o intuito de identificar associações entre variáveis relacionadas aos acidentes de trânsito neste tipo de rodovia. Para isso, utilizaram algoritmos de aprendizado de máquina visando a extração de modelos e padrões.

Com base no contexto e nos trabalhos supracitados, a seguir é descrita a proposta deste artigo, que visa apresentar um processo fusão de dados e informações para ser aplicado na prevenção de acidentes de trânsito nas rodovias federais brasileiras.

5. A Fusão de Dados e Informações no contexto da prevenção de acidentes de trânsito nas rodovias federais brasileiras

Para cumprir com o objetivo deste trabalho de pesquisa, é apresentada a seguir uma proposta de aplicação da fusão de dados e informações na prevenção de acidentes de trânsito no Brasil. Cabe destacar que essa proposta seguirá os princípios do modelo *JDL*, mencionado anteriormente, mas, no entanto, não seguirá todos os níveis que o modelo dispõe, uma vez que o modelo permite esse tipo de adaptação (STEINBERG et al., 1999).

O objetivo do processo de fusão aqui descrito é proporcionar uma melhor representação, organização e confiabilidade das informações, de forma a auxiliar no planejamento e tomadas de decisão voltadas para prevenção de acidentes de trânsito nas rodovias federais brasileiras, dado o alto nível de variação, inconsistência e inconfiabilidade das informações que são disponibilizadas pelos órgãos diretamente envolvidos nesta questão.

Nos parágrafos seguintes são descritos os passos de cada nível do processo, com

base no modelo *JDL*, que será empregado nesta proposta de aplicação da fusão de dados e informações. O processo de fusão aqui proposto utilizará apenas os níveis de pré-processamento, identificação de objetos, construção de situações e interação humano-computador.

Como ponto de partida, o processo de fusão de dados e informações necessita de algum tipo de insumo para que sejam processados os dados e informações. No contexto de acidentes de trânsito deste trabalho de pesquisa, são utilizados três tipos de fontes, sendo elas: 1) DNIT, fornecendo informações do trafego das rodovias federais; 2) PRF, fornecendo informações sobre a quantidade de acidente e infrações nas vias; 3) CEMADEN, fornecendo informações de dados geográficos e das condições meteorológicas das vias.

Com esses três tipos de fontes iniciais, é possível obter um volume de informações satisfatórios para aplicação do processo de fusão, e assim, seguir para o nível 0 (de acordo com o modelo *JDL*). No entanto, pode haver mais fontes além das mencionadas anteriormente caso seja necessário. Cabe destacar que tais dados são fornecidos de forma estática pelos órgãos e entidades responsáveis, não entrando assim este trabalho na questão dos dados fornecidos de forma dinâmica e em tempo real.

A etapa de pré-processamento, conhecida como nível 0 do modelo *JDL*, se caracteriza como nível de entrada dos dados e informações, e de preparação dos mesmos para os níveis de fusão seguintes (HALL e JORDAN, 2010).

No contexto deste trabalho de pesquisa, as informações advindas das fontes mencionadas anteriormente, recebem alguns tratamentos como limpeza, normalização e padronização das informações, com objetivo de organizar e facilitar na identificação, análise e recuperação das mesmas, e assim, facilitar a aplicação das técnicas do nível 1 (identificação de objetos) do modelo de fusão.

No nível 1, o processo de análise e associação de dados de múltiplas fontes tem como intuito definir, da maneira mais confiável possível, quais as entidades estão envolvidas no processo, como por exemplo, associar características de infrações de trânsito aos acidentes em uma determinada rodovia, e por sua vez, às informações climáticas.

Tal processo de associação oferece a possibilidade de se descobrir uma concentração grande de um dado tipo de situação, como por exemplo, o volume de acidentes de trânsito em um trecho específico de uma rodovia federal, ou até mesmo classificar uma região da rodovia com alto risco de acidentes provocados por excesso de velocidade. Todo os pontos descritos anteriormente se caracterizam com base em técnicas computacionais aliadas ao conhecimento de especialistas neste tipo de domínio.

Ao realizar a identificação dos objetos das múltiplas fontes de dados, o nível 2 elabora uma interpretação contextual dos dados, ou seja, analisa como os objetos encontrados no nível 1 estão relacionados com o ambiente em que estão empregados (HALL e JORDAN, 2010).

Um exemplo disso, é ao identificar em quais rodovias de pista simples o nível de acidentes por excesso de velocidade ocorre devido ao volume de tráfego de veículos de grande porte (caminhões ou ônibus), na qual uma possível medida de prevenção seria incluir mais faixas adicionais a esse tipo de via ou colocar radares em localizações estratégicas para averiguar, a médio ou longo prazo, se o índice de acidentes irá diminuir.

Já os níveis 3 e 4 do modelo *JDL* não são utilizados nesta proposta, devido ao alto nível de complexidade e falta de padronização das próprias fontes, onde suas informações em certos casos são divergentes e acarretam intepretações erradas durante a construção de uma situação.

Por fim, no nível 5, o processo de fusão busca apresentar as informações ao usuário de forma mais clara e compreensível, para auxiliar na tomada de decisão. As informações devem ser representadas, por meio de indicadores de situações, tais como: nível de acidentes de trânsito, características envolvendo situações nível excessivo de excesso de velocidade em quilômetros específicos, entre outras informações.

6. Resultados Esperados e Considerações Finais

Ao realizar o processo de fusão de dados e informações proposto neste trabalho, é esperada a obtenção de informações mais claras e coerentes, com alto nível de confiabilidade e consistência, para possibilitar o máximo de entendimento das situações identificadas, e quais ações de prevenção podem ser adotadas, bem como os fatores de relevância que podem ser integrados para melhoria de todo o processo de fusão.

Como trabalhos futuros, pretende-se a aplicação do processo de fusão proposto neste artigo em um caso prático, com o intuito de validar se os resultados obtidos a partir deste processo auxiliam na representação das informações referentes à prevenção de acidentes de trânsito em uma rodovia federal brasileira.

Referências

ABULATIF, L. I. Processo de integração de dados: um modelo de gestão da informação para múltiplas bases de dados de acidentes de trânsito no Brasil. **Epidemiologia e Serviços de Saúde.** Vol. 27, e2017160, 2018.

BLASCH, E. et al. Revisiting the JDL model for information Exploitation. In: Proceedings of the 16th **International Conference on Information Fusion**. IEEE, p. 129-136, 2013.

BOTEGA, L. C, et. al. Quality-aware human-driven information fusion model. In **International Conference on Information Fusion**. IEEE Computer Society. p. 1–10, Xian, 2017.

COSTA, J, J., BERNARDINI, F. C., VITERBO J. F. A mineração de dados e a qualidade de conhecimentos extraídos dos boletins de ocorrência das rodovias federais brasileiras. **AtoZ: novas práticas em informação e conhecimento**. Vol. 3, n. 2, p. 139-157, 2014.

Departamento Nacional de Infraestrutura de Transportes (DNIT). **Operações rodoviárias**. 2014. Disponível em: <https://www.dnit.gov.br/rodovias/operacoes-rodoviaria>.

HALL, D. L., MCMULLEN, S. A. H. **Mathematical Techniques in Multisensor Data Fusion**. [SI]: Artech House, 2004.

HALL, D. L., JORDAN, J., **Human-centered information fusion.** [SI]: Artech House, 2010.

INSTITUTO NACIONAL DE PESQUISAS ECONÔMICAS APLICADAS (IPEA). **Acidentes de trânsito nas rodovias federais brasileiras: caracterização, tendências e custos para a sociedade.** Relatório de pesquisa. Brasília. 2015. Disponível em: <http://www.ipea.gov.br/portal/images/stories/PDFs/relatoriopesquisa/150922_relatorio_acidentes_transito.pdf>.

LLINAS, J., BOWMAN, C., ROGOVA, G., STEINBERG, A., WALTZ, E., WHITE, F. Revisiting the JDL data fusion model II. In: **Space and Naval Warfare Systems Command.** San Diego, CA. 2004.

MINISTÉRIO DA INFRAESTRUTURA. **Programa Rodovida.** Disponível em: <http://infraestrutura.gov.br/rodovida.html>.

Organização Pan-Americana de Saúde (OPAS). **Folha Informativa - Acidentes de Trânsito.** Disponível em: <https://www.paho.org/bra/index.php?option=com_content&view=article&id=5147:acidentes-de-transito-folha-informativa&Itemid=779>.

RYDER, B., GAHR, B., EGOLF, P., DAHLINGER, A. Preventing Traffic Accidents with In-Vehicle Decision Support Systems - The Impact of Accident Hotspot Warnings on Driver Behavior. **Decision Support Systems.** vol. 99, pp. 64-74, 2017.

SOHN, S. Y., LEE, S. H. Data fusion, ensemble and clustering to improve the classification accuracy for the severity of road traffic accidents in Korea. **Safety Science**. Vol. 41, n. 1, p. 1-14, 2003.

STEINBERG, A, N., BOWMAN, C. L., WHITE, F. E., Revisions to the JDL data fusion model. In: **Proceedings of SPIE**. Vol. 3719, n.1, p. 430-441, 1999.

SUN, B. et al. Anomaly-Aware Traffic Prediction Based on Automated Conditional Information Fusion. In: **2018 21st International Conference on Information Fusion (FUSION)**. IEEE, p. 2283-2291, 2018.

UMA ESTRATÉGIA PARA RECOMENDAÇÃO DE COLABORADORES EM REPOSITÓRIOS DE DADOS CIENTÍFICOS

A Strategy for Contributors Recommendation in Scientific Data Repositories

Felipe Affonso[1], Thiago Magela Rodrigues Dias[1], Monique de Oliveira Santiago[1]
(1) Centro Federal de Educação Tecnológica de Minas Gerais,
Av. Amazonas, 7675 - Nova Gameleira, Belo Horizonte - MG, 30510-000
felipe-affonso@hotmail.com; thiagomagela@gmail.com; moniqueosantiago@gmail.com

Resumo:
Em uma rede de colaboração científica, uma conexão é formada quando dois ou mais cientistas publicam um trabalho em conjunto, nesse caso, os trabalhos representam as arestas, e os cientistas representam os nós da rede. Utilizando conceitos da analise de redes sociais, é possível compreender melhor o relacionamento entre os nós. O trabalho em questão tem como objetivo realizar a predição de ligações em redes de coautoria formadas pelos doutores com currículos cadastrados na Plataforma Lattes, na área de Ciências da Informação. Atualmente, a Plataforma Lattes conta com 6.4 milhões de currículos de indivíduos e representa um dos repositórios científicos mais relevantes e reconhecidos mundialmente. Com isso, é possível compreender o comportamento da rede e acompanhar a sua evolução ao longo do tempo. Para tanto, algumas etapas se fazem necessárias, são elas: extração dos dados, criação das redes de coautoria, definição dos atributos a serem utilizados, criação de um conjunto de dados, e por fim, utilizar os mesmos como entrada em um algoritmo de aprendizado de máquinas. Através dos resultados é possível estabelecer, com precisão, a evolução da rede de colaborações cientificas dos pesquisadores a nível nacional, auxiliando assim as agências de fomento na escolha de futuros pesquisadores de destaque.

Palavras-chave: redes; colaboração científica; predição de ligações.

Abstract:
In a scientific collaborative network, a connection is formed when two or more scientists publish a paper together, in which case the work represents the edges, and the scientists represent the nodes of the network. Using concepts from social network analysis, it is possible to better understand the relationship between nodes. The objective of this work is to predict the connections in co-authoring networks formed by PhDs with curricula registered in the Lattes Platform, in the area of Information Sciences. Lattes Platform currently has 6.4 million individual curricula and represents one of the most relevant and recognized scientific repositories worldwide. With this, it is possible to understand the behavior of the network and monitor its evolution over time. To do so, some steps are necessary, they are: data extraction, creation of co-authoring networks, definition of the attributes to be used, creation of a data set, and finally, use it as input to a machine learning algorithm. Through the results it is possible to establish, precisely, the evolution of the network of scientific collaborations of researchers at the national level, thus assisting the funding agencies in choosing future outstanding researchers.
Keywords: networks; scientific collaboration; link prediction.

1. Introdução

No final da década de 90, diversos pesquisadores dedicaram atenção aos estudos de redes. Foram realizados trabalhos sobre a área da biologia, a internet, roteadores, entre outros (Newman, M. E., 2001; Newman, M. E. e Park, J., 2003; Barabási, A.-L. e Albert, R. 1999). Tais estudos permitiram entender o relacionamento entre os nós. Ao se estudar essas ligações por algum tempo, surge a pergunta "como ocorre a evolução da rede ao longo do tempo?", porém Hasan e Zaki (2011) explica que compreender a evolução da rede como um todo é uma tarefa complexa.

Com esses conceitos em mente, Liben-Nowell e Kleinberg (2003) propuseram o problema da predição de liga-

ções. Inicialmente foram utilizados métodos que calculavam a similaridade entre dois nós da rede. Quanto mais parecidos os nós, maior a chance de possuírem uma ligação entre si. Portanto, diversos outros métodos foram propostos para melhor resolução do problema da predição de ligações (Acar, E., Dunlavy, D. M., Kolda, T. G., 2009; Zhou, T., Lü, L., Zhang, Y.-C., 2009; Liu et al., 2011).

Atualmente, utiliza-se métodos probabilísticos, métodos baseados em álgebra linear, e também, os métodos que transformam esse problema em um de classificação binária, dessa forma, diversos algoritmos podem ser utilizados pra sua resolução. Neste trabalho, trataremos o problema da predição de ligações como um problema de classificação, dessa forma, algoritmos da área de sistemas de recomendação são utilizados para realização dos objetivos propostos.

Aplicando tais conceitos a um domínio mais especifico, podemos dirigir as atenções às redes pertencentes à comunidade cientifica. Ao se publicar um trabalho com outro cientista, uma ligação é formada pela colaboração realizada. Nestas redes os autores representam os nós, e as colaborações científicas representam as arestas (Maruyama, W. T. e Digiampietri, L. A., 2019). Tais redes são chamadas de redes de coautoria, e serão o objeto de estudo deste trabalho.

Neste contexto, a Plataforma Lattes, mantida pelo CNPQ[1], tem sido fonte de dados de diversos trabalhos que visam analisar redes de colaboração cientificas, principalmente por englobar dados de grande parte da produção científica nacional. Atualmente, a Plataforma Lattes conta com 4 milhões de currículos de pesquisadores e representa uma das fontes de dados científicos mais relevantes e reconhecidos mundialmente (Lane, 2010). O conjunto de dados, registrados nos currículos cadastrados na Plataforma Lattes possui atributos como: nome, formação acadêmica, experiência profissional, projetos, publicações cientificas, entre outros. O grande volume de dados presente nos currículos pode fornecer informações valiosas e até então desconhecidas (Dias et al., 2013).

2. Objetivos

Será realizada a predição de ligações em redes de coautoria formada pelos dados de doutores com currículos cadastrados na Plataforma Lattes. Com isso, será possível compreender o comportamento dessa rede e acompanhar a sua evolução ao longo do tempo. Através deste estudo, também será possível identificar os pesquisadores que poderão colaborar em um futuro instante do tempo. Em um segundo momento, partindo da analise proposta, também se torna possível identificar os pesquisadores mais influentes na rede de coautorias.

3. Procedimentos Metodológicos

Para que seja possível atingir os objetivos propostos, alguns passos se fazem necessários. Nesta seção serão destacados os métodos utilizados para que seja possível realizar a predição de futuras ligações em uma área específica. Para tanto, foi escolhida a grande área de Ciências Sociais e Aplicadas e posteriormente a subárea Ciência da Informação. Este conjunto de dados possui 1.094 pesquisadores com título de doutor. Inicialmente será apresentado o *framework* utilizado para extração dos dados. Em um segundo momento, as redes de colaboração científicas criadas, e por último, os atributos selecionados para a predição serão caracterizados.

Para início do desenvolvimento do trabalho, foi necessário realizar a extração dos dados a serem utilizados. Para tanto, o *LattesDataExplorer (*Dias, T., 2016*),* um *framework* para extração e

[1] Conselho Nacional de Desenvolvimento Científico e Tecnológico

Quadro 1 – Atributos Utilizados

Atributo	Descrição
Vizinhos em Comum (VC)	De acordo com Liben-Nowell e Kleinberg (2003), a forma mais simples de realizar a predição de arestas, é através da métrica vizinhos em comum. que pode ser entendida como a quantidade de nós em comum que dois nós específicos possuem.
Coeficiente de Jaccard (JC)	mede a probabilidade de que ambos *x* e *y* possuam um vizinho *v*, escolhido aleatoriamente que *x* ou *y* possuam. Hasan e Zaki (2011) explicam que ao contrário do atributo Vizinhos em Comum, o coeficiente de Jaccard normaliza o número de vizinhos em comum.
Adamic/Adar (AA)	Essa formulação atribui às características mais raras um peso maior. Podemos entendê-la como o número de propriedades compartilhadas pelos nós, dividido pelo *log* da frequência das características.
Resource Allocation (RA)	Seguindo mesmo raciocínio, a métrica *Resource Allocation* atribui peso na relação de dois nós favorecendo as relações entre aqueles que possuem poucos relacionamentos.
Preferential Attachment (PA)	Considerando apenas o tamanho das vizinhanças dos nós, a métrica *Preferential* Attachment foi proposta. Em suma, estabelece que a probabilidade de um novo relacionamento com outros vértices é baseada no grau do nó em questão.
Menor Caminho (MC)	O fato de que amigos de amigos podem criar uma ligação sugere que a distância entre os nós de uma rede podem influenciar na formação de novas ligações. Podemos entendê-la como o caminho mínimo entre dois nós.
Colaborações em conjunto (Peso)	Dessa forma é possível identificar colaboradores que já trabalham juntos a mais tempo, e possivelmente possuem uma maior influência nos próximos instantes de tempo.

Fonte: Adamic, L. A. e Adar, E., 2003; Digiampietri, L. et al., 2015; Maruyama, W. T. e Digiampietri, L. A., 2019; Liben-Nowell, D. e Kleinberg, J., 2007; Potgieter, A. et al., 2009; Hasan, M. A. e Zaki, M. J., 2011; Chen, H., Li, X., Huang, Z., 2005; Lü, L. e Zhou, T., 2011.

tratamento dos dados foi utilizado. Já com os dados extraídos e organizados, é necessário realizar a criação das redes. Dias, T. M. R. e Moita, G. F. (2015) apresentam um método para identificação de colaborações científicas em grandes bases de dados, com a utilização de baixo poder computacional. Portanto, este método foi utilizado para geração das redes utilizadas neste trabalho.

Após a criação das redes de colaboração, se faz necessário identificar quais atributos serão utilizados para a predição. Para tanto, um conjunto básico de características, oriundos de outros trabalhos que abordaram esse tema foram selecionados e são apresentados no Quadro 1.

Após definir os atributos, alguns passos se fazem necessários. Em primeiro momento é necessário definir os períodos para treino e teste, portanto, 3 redes diferentes foram criadas. Para a rede 1, foram definidas as publicações realizadas no período entre 1960 e 2000, que será chamado de período inicial. Já a segunda rede foi criada para o período de 2001 a 2010. Por fim, foi estabelecido o período de 2011 a 2018 para a terceira e última rede. Tais períodos compreendem a data do primeiro trabalho registrado na plataforma até o último ano finalizado anteriormente a apresentação deste trabalho.

O conjunto de dados contendo os pesquisadores, as ligações entre si e os atributos selecionados foi então utilizado como entrada em um algoritmo de aprendizado de máquina. Cada linha do conjunto de dados é composta pelos seguintes itens: Identificação do primeiro pesquisador, identificação do segundo pesquisador, Vizinhos em Comum, Coeficiente de Jaccard, Adamic/Adar, *Resource Allocation*, *Preferential Attachment*, Menor Caminho, Peso, e por fim, a presença, ou não, de uma aresta.

Nessa etapa do trabalho, o problema do desbalanceamento de classes vem à tona. O número de ligações possíveis em um grafo é quadraticamente relacionado ao número

Tabela 1 - Métricas geradas a partir das predições

Algoritmo	Precisão	Revogação	F1	AUC
Regressão Logística	0.67	0.66	0.65	0.70
K-Vizinhos Mais Próximos	0.71	0.68	0.68	0.71
Baías Ingênuas	0.76	0.62	0.56	0.70
Florestas Aleatórias	0.70	0.68	0.67	0.71

Fonte: Dados da Pesquisa, 2019.

de nós, no entanto, o número de ligações reais representa apenas uma pequena fração deste número (Hasan, M. A. e Zaki, M. J., 2011). Uma técnica tradicional para superar o desbalanceamento das classes é chamada de sob amostragem. Ela consiste em reduzir o número de amostras da classe determinante, de forma randômica, igualando assim o número de componentes para ambos os casos. Essa técnica foi utilizada no trabalho aqui apresentado. Inicialmente o conjunto de dados apresentava uma proporção de 152 arestas ausentes, para cada aresta presente. Após a aplicação da sob amostragem, o número de arestas presentes e ausentes é o mesmo. Com os dados balanceados, o algoritmo para predição de ligações foi executado.

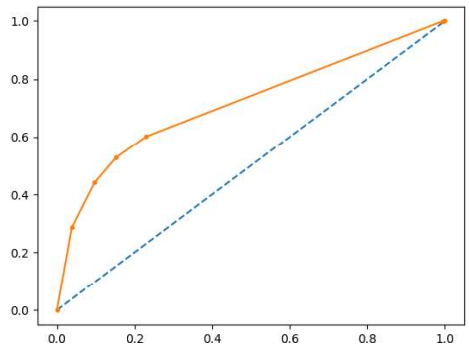
4. Resultados

Ao longo do processo descrito na seção anterior, o conjunto de dados sofreu algumas alterações. Os 1.094 pesquisadores podem possuir um total de 597.871 arestas. Destas, apenas 3.831 representavam arestas positivas na Rede 3, portanto, através do balanceamento das amostras, um conjunto randômico de outras 3.831 arestas ausentes foi escolhido. Sendo assim, o conjunto de dados utilizado para entrada no algoritmo de predição de dados é composto por 7.662 registros. Dessa forma, foram selecionadas 5.746 ligações (escolhidas aleatoriamente) para treino, representando 25% do conjunto total, e outros 1.916 ligações para teste.

Diversos algoritmos podem ser utilizados para resolução de problemas de classificação, dentre estes, alguns foram selecionados para execução do trabalho, são eles: Regressão Logística, K-Vizinhos Mais Próximos, Baías Ingênuas e Florestas Aleatórias. Cada uma dessas técnicas apresenta uma particularidade diferente, e consequentemente, diferentes consequências. Portanto, seus resultados serão apresentados na Tabela 1 utilizando as métricas: precisão, revogação, F1 e área sob a curva (AUC). Normalmente, em algoritmos utilizados para predição de ligações, a área sob a curva é utilizada pela maioria dos autores, portanto, a utilizaremos como base.

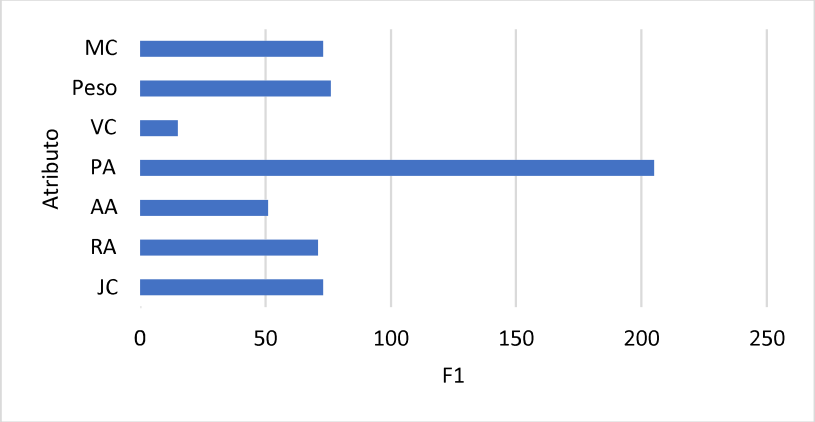
Cada uma das métricas utilizadas para validação dos resultados possui características próprias. A precisão tem como objetivo responder a seguinte pergunta: de todos os valores preditos positivos, quantos realmente estão corretos, uma alta precisão está relacionada a poucos falsos positivos. Já considerando todos os valores positivos, a revogação tem como objetivo saber quantos destes foram realmente preditos. A métrica F1 leva em consideração a precisão e a revogação, fazendo assim uma média ponderada dessas duas métricas. Por último, a área

Figura 1 - Área Sob a Curva (AUC) para o algoritmo K-Vizinhos Mais Próximos



Fonte: Dados da Pesquisa, 2019.

Figura 2 - Importância de cada atributo



Fonte: Dados da Pesquisa, 2019.

sob a curva, ou *area under the curve* (AUC), em inglês, é utilizada para exibir o desempenho de um modelo de classificação ao longo de todo o processo de aprendizagem.

Analisando a Tabela 2, é possível perceber que os algoritmos escolhidos obtiveram bons resultados. Ao observar a área sob a curva, percebemos que todos obtiveram um resultado acima do que um mero acaso. Essa situação é melhor explicada na Figura 1, onde a linha pontilhada em azul representa uma chance de 50% de acerto, ou seja, probabilidades iguais para a predição ser da classe correta ou incorreta, e a linha laranja representa os valores das predições realizadas. Dessa forma, fica claro que o algoritmo conseguiu utilizar o conjunto de dados e características apresentado para realizar predições corretas a respeito de futuras ligações.

Dentre os algoritmos utilizados, o que apresentou um melhor desempenho, levando em conta todas as métricas, foi o K-vizinhos mais próximos, seguido por Florestas Aleatórias, Baías Ingênuas, e, por último Regressão Logística. Porém, existe uma pequena diferente entre os resultados obtidos, deixando claro que, para o problema em questão, ainda não podemos estabelecer qual técnica deveria ser utilizada como padrão. Ao se analisar o processo de aprendizado levando em consideração os atributos utilizados, é possível identificar encontrar a ordem de influência de cada um deles no resultado final. Podemos observar, a partir da Figura 2, que a ordem de importância dos atributos para a predição aqui realizada é: *Preferential Attachment*, Peso das Colaborações, Caminho Mais Curto, Coeficiente de Jaccard, *Resource Allocation*, Adamic/Adar, e, por fim, Vizinhos em Comum. Tal fato apresenta um comportamento, até então, diferente da maioria dos referenciais teóricos aqui estudados, onde, na maior parte das vezes, o atributo mais relevante é o Vizinhos em Comum. Já nos estudos aqui realizados, a métrica *Preferential Attachment* é responsável por uma boa parte do resultado final.

5. Considerações Finais

Os resultados aqui apresentados demonstram que é possível realizar a predição de ligações utilizando informações da própria rede estudada. O objetivo proposto foi então alcançado, uma vez que, a partir da utilização destes dados é possível saber, por exemplo, se dois pesquisadores da área citada acima, irão colaborar em um futuro instante de tempo. O desempenho das métricas de avaliação ficou em torno de 70% representando um bom resultado, porém valores maiores podem ser alcançados a partir da utilização de mais atributos.

Como trabalhos futuros, destaca-se a importância de aumentar o conjunto de dados, ou até mesmo buscar outras formas de solucionar o problema do

desbalanceamento de classes, aumentando assim o número de amostras presentes para treino do algoritmo. A partir disso, espera-se que os classificadores apresentem um desempenho ainda melhor.

Referências

Acar, E., Dunlavy, D. M., Kolda, T. G. Link Prediction on Evolving Data Using Matrix and Tensor Factorizations. In: **Ieee. Data Mining Workshops, 2009. Icdmw'09. Ieee International Conference On,** 2009, p. 262–269.

Adamic, L. A., Adar, E. Friends And Neighbors On The Web. **Social Networks, Elsevier,** 2003, v.25, n. 3, p. 211–230.

Barabási, A.-L. e Albert, R. Emergence of scaling in random networks. science, **American Association for the Advancement of Science,** 1999, v.286, n.5439, p. 509–512.

Chen, H., Li, X., Huang, Z. Link Prediction Approach to Collaborative Filtering. In: **Ieee. Proceedings Of The 5th Acm/Ieee-Cs Joint Conference On Digital Libraries (Jcdl'05)**, 2005, p.141–142.

Dias, T. M. et al. Modelagem E Caracterização De Redes Científicas: Um Estudo Sobre A Plataforma Lattes. In: **Brasnam-Ii Brazilian Workshop On Social Network Analysis And Mining**, 2013. p. 10–20.
Dias, T. M. R. e Moita, G. F. A Method For The Identification Of Collabora- tion In Large Scientific Databases. **Em Questão**, 2015, Vol. 21, N. 2, p. 140–161.

Dias, T. Um Estudo Da Produção Científica Brasileira a Partir De Dados Da Plataforma Lattes. **Programa De Pós-Graduação Em Modelagem Matemática E Computacional, Centro Federal De Educação Tecnológica De Minas Gerais, Belo Horizonte (Doutorado)**, 2016, 181p.

Digiampietri, L. et al. Um Sistema De Predição De Relacionamentos Em Redes Sociais. In: **Brazilian Symposium on Information Systems**, 2015, V. 11.

Hasan, M. A. e Zaki, M. J. A Survey of Link Prediction In Social Networks. In: **Social Network Data Analytics. Springer**, 2011, p. 243–275.

Lane, J. Let's Make Science Metrics More Scientific. **Nature, Nature Publishing Group,** 2010, v. 464, n. 7288, p. 488.

Liben-Nowell, D. e Kleinberg, J. The Link-Prediction Problem for Social Networks. **Journal of The American Society For Information Science And Technology, Wiley Online Library,** 2007, v.58, n.7, p.1019–1031.

Liu, Z. et Al. Link Prediction in Complex Networks: A Local Naïve Bayes Model. **Epl (Europhysics Letters), Iop Publishing**, 2011, v.96, n.4, p.48007.

Lü, L. e Zhou, T. Link Prediction in Complex Networks: A Survey. **Physica A: Statistical Mechanics And Its Applications, Elsevier,** 2011, v.390, n.6, p.1150–1170.

Maruyama, W. T. e Digiampietri, L. A. Co-Authorship Prediction In Academic Social Network. In: **Sbc. Anais Do V Workshop Brasileiro De Análise De Redes Sociais E Mineração**, 2019. p.79–90.

Newman, M. E. The structure of scientific collaboration networks. **Proceedings of the national academy of sciences, National Acad Sciences,** 2001, v.98, n.2, p. 404–409.

Newman, M. E.; Park, J. Why social networks are different from other types of networks. **Physical Review E, APS,** 2003, v. 68, n. 3, p. 036122.

Potgieter, A. et al. Temporality in Link Prediction: Understanding Social Complexity. Emergence. **In Complexity & Organization (E: Co), Citeseer**, 2009, v.11, n.1, p.69–83.

Zhou, T., Lü, L., Zhang, Y.-C. Predicting Missing Links Via Local Information. **The European Physical Journal B, Springer,** 2009, v.71, n.4, p.623–630.

UMA SOLUÇÃO SEMI-AUTOMÁTICA PARA EXTRAÇÃO, TRANSFORMAÇÃO E CARGA DE DADOS ABERTOS CONECTADOS

A Semi-automatic Solution for Extraction, Transformation and Loading of Open Linked Data

Sérgio Souza Costa[1], Mateus Vitor Duarte Sousa[2], Micael Lopes da Silva[3], Eddye Cândido de Oliveira[4], José Victor Meireles Guimarães[5]

(1) Universidade Federal do Maranhão (UFMA) São Luís - MA - Brazil, prof.sergio.costa@gmail.com.
(2) Universidade Federal do Maranhão (UFMA) São Luís - MA - Brazil, mateusriograndense@gmail.com.
(3) Universidade Federal do Maranhão (UFMA) São Luís - MA - Brazil, micaelopes32@gmail.com.
(4) Universidade Federal do Maranhão (UFMA) São Luís - MA - Brazil, eddyeoliver@gmail.com.
(5) Universidade Federal do Maranhão (UFMA) São Luís - MA - Brazil, jvictormguimaraes@gmail.com.

Resumo:
Este artigo propõe uma metodologia semi-automática para a extração de dados públicos, sua transformação e carga para um servidor de dados abertos e conectados. Primeiro é realizada a extração dos dados dos portais públicos da instituição e então mapeados para o formato de dados abertos e conectados; em seguida, eles são carregados para um servidor de dados abertos e conectados. Para isso utilizou-se diversas tecnologias como servidor de banco de dados, *framework* web, container e Plataforma como Serviço. Como resultado, foram extraídas e mapeadas diversas entidades do sistema acadêmico da Universidade Federal do Maranhão; incluindo discentes, docentes, unidades acadêmicas, cursos e monografias. A solução mostrou-se simples e replicável em instituições que já possuem dados públicos disponíveis na web. As tecnologias utilizadas foram suficientes, e um simples teste de eficiência foi realizado. Nele foram necessários em média 3,4 segundos para retornar 169.228 triplas executando no ambiente Google Colab.
Palavras-chave: dados abertos, dados conectados, universidade, triple-store

Abstract:
This paper proposes a semi-automatic methodology for extraction of public data, it's transformation, and it's loading to an open and linked data server. First, data are extracted from the institution's public portals and then mapped to open and linked data format; then the data are loaded onto an open and linked data server. For this, we used several technologies such as database servers, web frameworks, containers and platforms as service (PaaS). As a result, several entities from the academic system of the Federal University of Maranhão were extracted and mapped; including students, professors, academic units, courses, and monographs. The solution proved simple, and replicable in institutions that already have public data available on the web. The technologies used were sufficient, and a simple efficiency test was performed. It took an average of 3.4 seconds to return 169,228 triples running it on the Google Colab environment.
Keywords: open data, linked data, university, triple-store

1. Introdução

No Brasil o acesso aos dados de instituições públicas já estava previsto pela Constituição de 1988, porém foi reforçado através da Lei de Acesso à Informação (Lei n.º 12.527/2011). Em resposta a esta demanda, instituições têm disponibilizado um grande volume de dados públicos através da web. Em geral, o acesso a estes dados requer a interação direta com um usuário, dificultando assim a sua análise e a sua utilização através de novas aplicações. Uma solução seria a disponibilização destes dados de modo aberto e conectado, como será apresentado neste artigo. Então, antes de prosseguir se faz necessário distinguir alguns conceitos destacados na Figura 1.

Figura 1: Dos dados, aos dados abertos e conectados.

Os dados estão na base da pirâmide, e são definidos como símbolos que representam propriedades de objetos, eventos e o seu ambiente (ACKOFF, 1989). Dados abertos são todos aqueles que podem ser usados, modificados e compartilhados livremente por qualquer pessoa para qualquer fim (OKF, 2009). Eles precisam ainda seguir alguns princípios, como, serem completos, primários, atuais, acessíveis, compreensíveis por máquina, não proprietários e livres de licença (OGWG, 2007). Por outro lado, o conceito de dados conectados, proposto por Berners-Lee, refere-se a um conjunto de boas práticas para publicar e interligar os dados estruturados na Web (HEATH e BIZER 2011). Por fim, dados abertos e conectados são aqueles que além de abertos, estão conectados.

É importante frisar que nem todos dados conectados são necessariamente abertos, inclusive existe uma agenda de pesquisa em dados fechados conectados (COBDEN et al., 2011). Além disso, nem todos os dados abertos são conectados. É possível disponibilizar dados abertos através de outras tecnologias e padrões, como o estilo arquitetural de software denominado REST (*Representational State Transfer*) proposto em (FIELDING, 2000).

Este artigo tem foco nos dados abertos e conectados. Abertos, pois é aplicado aos dados públicos de uma universidade federal. Já a escolha por dados conectados deve-se as vantagens desse paradigma, como a integração com outros dados e uma linguagem de consulta.

Alguns trabalhos recentes já têm concentrado esforços para disponibilizar bases de dados abertos e conectados de algumas universidades, como (PANTOJA, 2013; D'AQUIN et. al, 2014; KESSLER; KAUPPINEN, 2012; DAGA et al, 2015; OLIVEIRA; GUIMARÃES; COSTA, 2018; ALENCAR; XAVIER; SOUZA, 2018). Porém, muitos destes projetos tem enfrentado o desafio de manter e expandir essa base de dados abertos e conectados. Esse desafio pode ser agravado em instituições públicas, como é o caso das universidades. Elas podem não possuir políticas específicas e/ou recursos humanos para a realização e manutenção destas bases.

2. Objetivos

Este artigo tem como objetivo apresentar uma metodologia semi-automática para a extração de dados públicos, e a sua transformação e carga como dados abertos conectados. Essa metodologia poderá então ser replicada em qualquer instituição que já possua os seus dados públicos disponibilizados na web.

3. Procedimentos Metodológicos

A metodologia descrita nesta seção foi influenciada pelo conceito ETL (*Extract Transform Load*) que é usualmente utilizado no contexto de *Data Warehouse* (VASSILIADIS; SIMITSIS; SKIADOPOULOS, 2002). Uma visão geral da metodologia é apresentada na Figura 2.

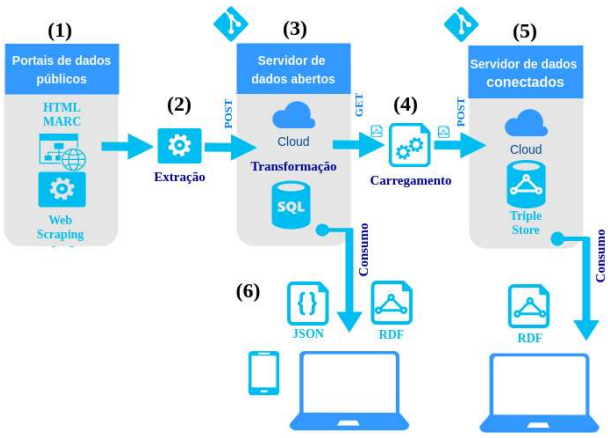


Figura 2: Visão geral da metodologia

Essa metodologia inclui a criação de dois servidores. O primeiro é um servidor de dados abertos (3) e o segundo é um servidor de dados conectados (5). O processo inicia com a execução da extração (2) dos dados dos portais públicos da instituição (1). Essa extração é uma rotina que pode ser

executada de acordo com um dado agendamento, rodando localmente ou como um *worker* dentro de um serviço de nuvem. Além de extrair, os dados serão convertidos para o formato JSON (*Java Script Object Notation*) e enviados ao servidor de dados abertos (3) através de requisições HTTP POST. O servidor de dados abertos (3) irá armazená-los num banco de dados e disponibilizá-los a um cliente (6) nos formatos JSON e RDF (*Resource Description Framework*). Por fim, durante o processo de implantação do servidor de dados conectados, será executado um conjunto de rotinas que receberá os dados do servidor de dados abertos (3) e os enviará para o servidor de dados conectados (5). Na próxima seção, essas etapas serão melhor detalhadas em um caso específico.

4. Resultados

Está seção apresenta como a metodologia foi aplicada aos dados públicos da Universidade Federal do Maranhão (UFMA).

4.1. Extração

A UFMA, como diversas outras instituições, já disponibiliza diversos dados públicos através da web. Então, o primeiro passo foi extrair estes dados usando *web scrapping*. Que é uma prática de recuperação de dados que usam algoritmos que analisam e extraem as informações entre *tags* de páginas HTML (MITCHEAL,2015). O web *scrapper* implementado utilizou a biblioteca BeautifulSoup[1]. Os dados extraídos foram persistidos em um banco de dados, e disponibilizados através de uma API (*Application Program Interface*) REST. Para o desenvolvimento da API foi utilizado o *framework* Flask[2]. Na versão atual, foram disponibilizados cinco *endpoints* com o total de 45.141 objetos, como apresentados na Tabela 1. Essa API pode ser acessada pelo endereço: https://dados-ufma.herokuapp.com/.

Tabela 1: *Endpoints* da API REST

Recurso	Atributos	Total
/docente	nome, código da subunidade, descrição, formação, áreas de interesse, lattes, email, telefone, link da imagem do sigaa	1477
/subunidade	código da subunidade, nome	176
/curso	código do curso, modalidade, nome do coordenador, município, nome	290
/discente	código do curso, matricula, nome do discente, nome do curso	27578
/monografia	ano, código do trabalho, código do curso, nome do discente, nome do orientador, siape do orientador, título	15620
Total		45141

4.2. Transformação

A transformação consiste no mapeamento dos dados públicos para o formato de dados conectados. Esse mapeamento pode ser denominado de *object-triple*, que é uma abordagem análoga ao mapeamento objeto relacional como discutido em (GROVE, 2010). Atualmente já existem diversas bibliotecas que fazem esse mapeamento e uma análise delas pode ser encontrada em (LEDVINKA; KREMEN, 2018). Porém, nenhuma das opções se adequaram às tecnologias utilizadas no projeto. Desse modo, optou-se por desenvolver uma biblioteca em Python com esse fim.

A biblioteca para o mapeamento *object-triple* foi denominada de SIMPOT[3] (*Simple Object-triple Mapping*). Inspirada no

[1] Veja https://www.crummy.com/software/BeautifulSoup
[2] Veja https://palletsprojects.com/p/flask/
[3] O repositório da biblioteca desenvolvida é http://github.com/inovacampus/simpot

SQLAlchemy[4], ela permite associar os vocabulários aos atributos de um dado objeto. Deste modo, usa-se uma abordagem declarativa para o mapeamento, favorecendo a leitura, compreensão e manutenção do código. Um exemplo de um mapeamento é apresentada na Listagem 1.

```
base = "https://sigaa.ufma.br/sigaa/public"

class Docente:
  nome = FOAF.name
  email = FOAF.mbox

 @RdfsClass(FOAF.Person,
  base + "/docente/portal.jsf?siape=")

 @BNamespace("dc", DC)
 @BNamespace("foaf", FOAF)

 def __init__ (self,none, email, mat):
    self.nome = Literal (nome)
    self.email = URIRef(email)
    #convenção para gerar o URI
    self._id = mat

d=Docente("Joao","joao@gmail.com",685)
print (graph (d)) # imprime o grafo
```

Listagem 1: Exemplo de uso da biblioteca SIMPOT para mapeamento object-triple.

Na Listagem 1 é apresentada apenas um recorte da modelagem de um docente. Observe que o mapeamento está sendo feito através de dois *decorators*: `@RdfsClass` e `@BNamespace`. O primeiro define que a classe `Docente` é do tipo `FOAF.Person`. FOAF (*Friend of a Friend*) é um vocabulário muito utilizado para representar informações sobre pessoas e organizações. Ele também foi usado nas propriedades `nome` e `email`, que foram associadas aos conceitos FOAF.name e FOAF.email respectivamente. O segundo *decorator* foi usado para associar *namespaces* aos dados conectados, nesse caso, os namespaces FOAF e DC (*Dublin Core*). Esse é um dos pontos que impossibilita um processo totalmente automático. A definição desses vocabulários requer um estudo do domínio que será mapeado. Neste caso, a biblioteca SIMPOT se tornou muito útil por facilitar a manutenção e atualização dos vocabulários. Então, antes de fazer o mapeamento, é preciso pesquisar os vocabulários já existentes que representam melhor os dados. Neste trabalho foram usados os seguintes vocabulários:

Tabela 2: Vocabulários utilizados

Vocabulário	Descrição
FOAF	Usado para representar o vínculo de pessoas, sejam por informações formais como documentos físicos e digitais ou por relacionamentos não formais.
VCard	Usado para descrever pessoas e organizações utilizando de técnicas da web semântica.
DBO	Projetado a partir do mapeamento feito de dados do DBPedia para prover a semântica da extração dos dados do Wikipedia.
DC	Usado para descrever metadados genéricos.
VIVO	Usado para representar as habilidades de pessoas envolvidas na criação, transmissão e preservação de conhecimento e trabalhos criativos.
BIBO	Usado para representar citações e referências bibliográficas.

4.3. Carregamento e publicação

Para a automatização da carga dos dados, foram utilizadas as seguintes tecnologias: Fuseki, Docker[5] e Heroku[6]. Para o armazenamento de dados, foi utilizado o *triple store* denominado Fuseki. Esse sistema foi instanciado dentro de um container Docker e publicado como uma aplicação no Heroku. Durante a criação do container o servidor Fuseki é baixado, instalado e

[4] SQLAlchemy é uma biblioteca Python para o mapeamento objeto-relacional (https://www.sqlalchemy.org/)

[5] *Docker* é uma tecnologia de containers Linux (https://www.docker.com/).

[6] O Heroku é uma plataforma como serviço (www.heroku.com).

configurado. Além disso, os dados são baixados da API de dados abertos para o formato RDF. Então, para cada carregamento, estes dados serão importados para dentro do *triple store*. Na fase atual a base de dados já possui 169.228 triplas.

4.4. Consumo

O consumo dos dados conectados pode ser realizado diretamente através do SPARQL *endpoint* http://bit.ly/31qbJLS. Além de ser possível escrever aplicativos ou *scripts* de análise de dados que enviam requisições para esse *endpoint*. A Figura 3, apresenta uma visão de como alguns desses dados estão conectados.

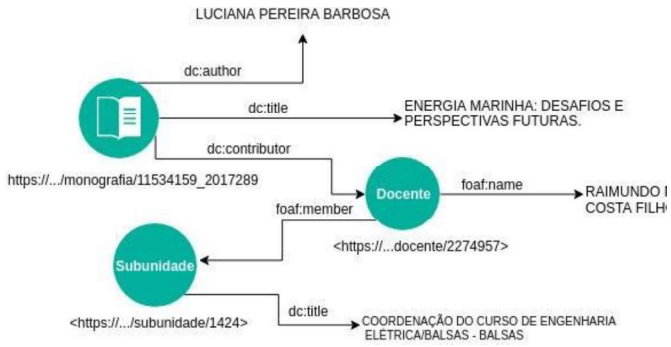


Figura 3: A conexão entre os recursos Monografia, Docente e Subunidade.

Observe que a monografia é um recurso, que se conecta a um discente e um docente, enquanto um docente se conecta a uma subunidade acadêmica. A Listagem 2 apresenta uma consulta que retorna o título da monografia e o nome do discente, curso e orientador.

```
SELECT
?mono ?autor ?nome_curso ?nom_orient
WHERE {
 ?s rdf:type bibo:Thesis.
 ?s dc:creator ?autor.
 ?s dc:title ?mono.
 ?s dc:publisher ?curso.
 ?s dc:contributor ?orient.
 ?curso dc:title ?nome_curso.
 ?orient foaf:name ?nom_orient.
}
```

Listagem 2: Exemplo de uma consulta SPARQL.

A consulta na Listagem 2 pode ser enviada por um script escrito em qualquer linguagem. Para um experimento rápido, foi escrito um script em Python (http://bit.ly/2kujpgP) que envia essa consulta, e ao receber os dados gera uma nuvem de palavras a partir do título de cada monografia. Tendo como resultado o gráfico da Figura 3, demonstrando visualmente como a palavra "análise" é muito comum nesses títulos. O nome da cidade e do estado são bem recorrentes também.

Figura 3: Nuvem de palavras a partir dos títulos de 15620 monografias.

Este artigo não teve como objetivo uma avaliação rigorosa de eficiência, mas mesmo assim foi realizado um teste simples através do seguinte script Python (http://bit.ly/2k9YtM1). Onde, neste contexto, foram necessários em média 3,4 segundos para retornar 169.228 triplas executando no ambiente Google Colab.

5. Considerações Finais

Este artigo apresentou uma metodologia semi-automática para extração dos dados públicos e a sua disponibilização como dados abertos conectados. Podendo ser replicado sem a necessidade do acesso direto a base de dados dos sistemas das instituições. Deste modo, mesmo pessoas externas às instituições são capazes de disponibilizar estes dados abertos e conectados.

As tecnologias utilizadas se mostraram eficientes para este projeto, destacando o SIMPOT, Fuseki, Docker e Heroku. Contudo, atualmente o serviço está sendo testado em uma conta gratuita do Heroku, o que impacta no tempo de resposta inicial. Então, com a arquitetura validada, o próximo passo será buscar parcerias e financiamentos para a evolução da base de dados.

A definição e utilização de vocabulários é um grande desafio para os dados conectados. Em um trabalho futuro, será realizada uma pesquisa para uma melhor definição e utilização dos vocabulários existentes.

Referências

ACKOFF, R. L. From data to wisdom. **Journal of Applied Systems Analysis**, v. 16, n. 1, p. 3–9, 1989.

ALENCAR, A.; XAVIER, D.; SOUZA, D. **Publicação e consumo de dados abertos conectados acadêmicos.** Revista Principia, 2018.

BERNERS-LEE, **T. Linked Data.** 2006. Disponível em: <https://www.w3.org/DesignIssues/LinkedData.html>.

BIZER, C.; HEATH, T.; BERNERS-LEE, T. **Linked data-the story so far. Semantic Services, Interoperability and Web Applications: Emerging Concepts**, p. 205–227, 2009.

COBDEN, M. et al. **A Research Agenda for Linked Closed Dataset. Proceedings of the Second International Workshop on Consuming Linked Data (COLD)**. Bonn, Germany, 2011

DAGA, E.; D'AQUIN, M.; ; ADAMOU, A.; BROWN, S. **The open university linked data – data. open.ac.uk.** Semantic Web, v. 7, n. 2, p. 183–191, 2015. Disponível em: <http://www.semantic-web-journal.net/system/files/swj973.pdf>.

D'AQUIN, M., DIETZE, S., HERDER, E., DRACHSLER, H., & TAIBI, D. (2014). **Using linked data in learning analytics**. *E-Learning Papers, 36*, 1–9.

FIELDING, Roy T.; TAYLOR, Richard N. **Architectural styles and the design of network-based software architectures**. Doctoral dissertation: University of California, Irvine, 2000.

GROVE, M. **Empire: RDF & SPARQL Meet JPA**. Disponível em: <https://www.dataversity.net/empire-rdf-sparql-meet-jpa/>. 2010.

HEATH, T.; BIZER, C. Linked Data: Evolving the Web into a Global Data Space. **Synthesis Lectures on the Semantic Web: Theory and Technology**, v. 1, n. 1, p. 1–136, 2011.

KESSLER, C.; KAUPPINEN, T. **Linked open data university of münster–infrastructure and applications.** In: SPRINGER. Extended Semantic Web Conference. 2012. p. 447–451. Disponível em: <http://kauppinen.net/tomi/lodum-eswc-2012.pdf>.

LEDVINKA, M.; KREMEN, P. **A comparison of object-triple mapping frameworks**. Semantic Web. 2018

MITCHEAL, R. **Web Scraping with Python - Collecting data from the modern web.** 1nd. ed. [S.I.]: O'Reilly, 2015.

OLIVEIRA E.; GUIMARÃES J.; COSTA S. **Migrando dos dados abertos para dados conectados: uma proposta para a Universidade Federal do Maranhão**, 2018.

OPEN GOVERNMENT WORKING GROUP. **8 Principles of Open Government Data. Sebastopol, CA**: opengovdata.org, 2007. Disponível em: <https://public.resource.org/8_principles.html>.

OPEN KNOWLEDGE FOUNDATION. **The Open Definition November**, 2009. Disponível em: <http://opendefinition.org/>

PANTOJA, J. **Linked Open Data at the UPF**. 2013. Disponível em: <http: //data.upf.edu/upf/docs/2013/jorgepm/lodatupf.pdf >

VASSILIADIS, P.; SIMITSIS, A.; SKIADOPOULOS, S. **Conceptual Modeling for ETL Processes**. Proceedings of the 5th ACM International Workshop on Data Warehousing and OLAP. 2002

WORKFLOW DE AGREGAÇÃO DE DADOS: PROCESSOS PARA CRIAÇÃO DE UMA INTERFACE DE BUSCA INTEGRADA DO PATRIMÔNIO CULTURAL

DATA AGGREGATION WORKFLOW: PROCESSES FOR CREATING A CULTURAL HERITAGE INTEGRATED SEARCH INTERFACE

Joyce Siqueira[1], Dalton Lopes Martins[2]
(1) Universidade de Brasília, Campus Darcy Ribeiro, Asa Norte, Brasília – DF, joycitta@gmail.com
(2) Universidade de Brasília, Campus Darcy Ribeiro, Asa Norte, Brasília – DF, daltonmartins@unb.br

Resumo:
Nos últimos anos diferentes instituições culturais vêm envidando esforços para difundir a cultura por meio da construção de uma interface única de busca integrada para seus objetos digitais. Estes esforços resultaram em diferentes propostas para agregação de dados, para as quais foram construídos workflows, que apresentam as etapas necessárias a esse intento. Considerando a possibilidade de diferentes workflows, com etapas distintas, esta pesquisa objetiva destaca-los e discuti-los. Para tal, foi realizada pesquisa descritiva e bibliográfica, de natureza qualitativa, em bases de dados acadêmicas e na literatura cinzenta. Como resultado, apresenta-se sete workflows de agregação propostos pela: American Art Collaborative, Biblioteca Nacional da Nova Zelândia, Fundação Europeana, Instituto de Ciência e Tecnologia da Informação, Secretaria de Cultura do México, Biblioteca Nacional da Austrália e Universidade de Nevada. A análise do conjunto de workflows resultou em oito diferentes etapas a serem executadas: 1. extrair, 2. usar ontologias, 3. transformar, 4. reconciliar, 5. armazenenar, 6. expor, 7. publicar e 8. possibilitar novas aplicações. Além disso, também é visível a necessidade de maior detalhamento das etapas, a fim de que seja possível replicar o workflow, e usufruir de seus benefício em outras instituições.
Palavras-chave: workflow de agregação; busca integrada; patrimônio cultural; instituições culturais

Abstract:
In recent years, different cultural institutions have been making efforts to spread culture through the construction of a single, integrated search interface for their digital objects. These efforts resulted in different proposals for data aggregation, for which workflows were built, which present the necessary steps for this purpose. Considering the possibility of different workflows, with different steps, this research aims to highlight and discuss them. To this end, a descriptive and bibliographical research of qualitative nature was conducted in academic databases and in gray literature. As a result, we present seven aggregation workflows proposed by: American Art Collaborative, New Zealand National Library, Europeana Foundation, Institute of Information Science and Technology, Secretariat of Culture of Mexico, National Library of Australia, and University of Nevada. Analyzing the set of workflows resulted in eight different steps to perform: 1. extract, 2. use ontologies, 3. transform, 4. reconcile, 5. store, 6. expose, 7. publish, and 8. enable new applications. In addition, it is also visible the need for more detailed steps, so that it is possible to replicate the workflow, and enjoy its benefits in other institutions.
Keywords: aggregation workflow; integrated search; cultural heritage; cultural institutions

1. Introdução

Difundir a cultura por meio da oferta de uma interface de busca integrada, que possibilite aos usuários uma navegação eficiente pelos diversos objetos digitais que compõe o patrimônio cultural é um objetivo fortemente almejado, de tal forma que, nos últimos anos, grandes instituições culturais envidam esforços para construí-la.

Parte integrante deste procedimento se configura na construção de workflows que apresentam as etapas necessárias a agregação de dados, de forma que o mesmo contemple todos os processos e tecnologias, da extração a publicização dos dados agregados.

Neste ponto, é importante ter claro que um workflow se trata de uma maneira de organizar etapas em uma sequência produtiva e eficiente, sendo estas planejadas, modeladas e automatizadas de forma a atingir propósitos bem definidos (SANTOS, 2013).

O objetivo deste estudo é localizar diferentes workflows de agregação de dados, desenvolvidos por instituições culturais, para realizar uma análise qualitativa das etapas escolhidas por cada instituição.

Dessa forma, o primeiro passo dado foi encontrar quais instituições propuseram soluções para a agregação de dados. Procurou-se, inicialmente, por consolidadas instituições reconhecidas por realizar este trabalho, tais como a Europeana e a Mexicana, e na sequência, por meio de palavras-chaves, novas propostas foram selecionadas.

Ao final, foram localizados sete workflows de Instituições culturais que são brevemente apresentadas na seção de Resultados.

Este artigo está assim dividido: seção 2, Objetivos, seção 3, Procedimentos Metodológicos, seção 4, Resultados, e por último, na seção 5, as Considerações Finais.

2. Objetivos

Esta pesquisa tem o propósito de analisar workflows de agregação de dados, desenvolvidos por instituições culturais. Dessa forma os objetivos específicos são: 1. localizar as instituições culturais que propuseram workflows de agregação; 2. apresentar os workflows e 3. identificar as etapas propostas.

3. Procedimentos Metodológicos

Pesquisa de caráter descritivo e bibliográfico, de natureza qualitativa, realizada em bases de dados acadêmicas e na literatura cinzenta, como intuito de encontrar o workflow de agregação de dados de reconhecidas instituições, além novas iniciativas.

As buscas foram realizadas no Google, Google Acadêmico, EBSCOhost e BRAPCI, utilizando os termos: *"pipeline", "workflow", "architecture", "aggregation", "metadata ingest", "metadata aggregation", "europeana", "mexicana", "dpla", "digital public library of america", "trove", "digitalnz"*, "aggregative data infrastructures".

Optou-se pelo Google para localizar workflows na literatura cinzenta e as demais bases por serem agregadoras de outras bases, tornado a pesquisa mais ampla. No caso da BRAPCI, também é uma base específica da área de Ciência da Informação.

4. Resultados

Foram localizados sete workflows de agregação, apresentados nas Figura 01 a 08, dispostas no Apêndice A, que são tratados nessa seção. Inicialmente, apresenta-se um breve resumo de cada Instituição propositora dos workflows.

A *American Art Collaborative* – AAC, é um consórcio de 14 instituições de arte, nos Estados Unidos, que visam investigar e começar a construir uma massa crítica de *Linked Open Data* – LOD.

Para Fink (2018), LOD que se trata de um método para publicar dados estruturados na web de forma que as informações sejam interconectadas e, assim, tornadas amplamente úteis.

A Secretaria de Cultura do México desenvolveu a Mexicana, um Repositório do Patrimônio Cultural do México, livre e aberto, que tem o objetivo principal de difundir e vincular os acervos do patrimônio cultural do México (SECRETARÍA DE CULTURA, 2018).

A Universidade de Nevada, por meio da equipe do departamento de Coleções Digitais das Bibliotecas da Universidade, reuniu esforços para encontrar maneiras de tornar mais eficiente a descoberta e uso das informações, iniciando assim estudos para adoção do LOD culminando no desenvolvimento do *UNLV's Linked Data Project* (SOUTHWICK, 2015).

A Fundação Europeana, desenvolveu a Europeana, que reuniu mais de 55 milhões de objetos digitais das coleções on-line de mais de 3.500 galerias, bibliotecas, museus, coleções audiovisuais e arquivos de toda a Europa (SCHOLZ, 2018).

O Istituto di Scienza e Tecnologie dell'Informazione desenvolveu o D-NET, um software que oferece um kit de serviços para a construção de Infraestruturas de dados (BARDI, MANGHI E ZOPPI, 2012).

A Biblioteca Nacional da Nova Zelândia junto a Rede do povo Aotearoa Kaharoa desenvolveu, no início de 2006, o DigitalNZ, que utiliza o software Supplejack para agregação de dados (DIGITAL NEW ZEALAND, 2019).

A Biblioteca Nacional da Austrália desenvolveu o Trove, que tem o objetivo de fornecer recursos relacionados à Austrália. Além de um mecanismo de busca, reúne conteúdo de bibliotecas, museus, arquivos e outras organizações de pesquisa e fornece um conjunto de serviços (TROVE HELP CENTRE, 2019).

Considerando a análise dos workflows, foram encontradas oito fases para agregação, sendo elas: extração, uso de ontologias, transformação, reconciliação, armazenamento, exposição, publicação e novas aplicações.

De forma sintética, estas etapas significam:

1. Extrair: extração dos dados em sua forma bruta, que podem estar, por exemplo, em pdf, em planilhas eletrônicas, documentos de texto, XML (*eXtensible Markup Language*), em bancos de dados relacionais, dentre outras opções.
2. Utilizar ontologias: selecionar vocabulários controlados pré-existentes para aplicação nos dados.
3. Transformar: realizar a normalização, limpeza e correção sintática dos dados.
4. Reconciliar: enriquecer os metadados por meio de outros dados existentes na web.
5. Armazenar: se trata da escolha de onde os dados coletados serão armazenados.
6. Publicar: se trata da interface única de busca integrada.
7. Expor: disponibilizar os dados agregados por meio de API, que exponham os dados em formato RDF, OAI-PMH ou JSON.
8. Possibilitar novas aplicações: a partir dos arquivos disponibilizados na etapa 'Expor' novas aplicações podem ser criadas.

O Quadro 01 mostra um panorama do uso de cada fase.

Quadro 01. Etapas dos Workflow de Agregação

Projeto/ Etapas	Extrair	Utilizar ontologias	Transformar	Reconciliar	Armazenar	Publicar	Expor	Possibilitar novas aplicações
AAC	X	-	X	X	X	X	X	X
DigitalNZ	X	-	-	X	X	X	-	-
D-NET Software	X	-	X	X	-	X	X	-
Europeana	X	X	X	X	X	-	-	-
Mexicana	X	-	X	X	X	X	X	X
TROVE	X	-	-	-	X	X	-	-
UNLV's Linked Data Project	X	X	X	X	X	X	X	-

Fonte: elaborado pelos autores

A documentação na qual os workflows estão inseridos apresentam alguns dados que não constam do fluxograma. Além disso, percebe-se pouca preocupação com a qualidade dos dados inseridos, ou seja, os dados coletados na etapa de extração.

Além das etapas, as publicações apresentação algumas ferramentas de softwares utilização para execução do workflow, e contata-se que não há escalabilidade, ou seja, à medida que o fluxo de dados cresce, o workflow torna-se impraticável.

De forma geral, os workflows são genéricos demais e não apresentam o fluxo real de processos necessários, contrariando assim, um dos princípios básicos de um workflow, que é a possibilidade de ser replicado.

Além disso, percebe-se a necessidade de um conhecimento técnico avançado e extremamente especializado para compreensão de todas as etapas.

5. Considerações Finais

A análise dos diferentes workflows de agregação de dados permitirá aos pesquisadores compreender quais etapas estão sendo executadas, quais estão sendo postas em segundo plano e quais precisam ser incluídas.

Como trabalho futuro, pretende-se realizar pesquisas direcionadas a cada etapa do workflow, além de propor um novo workflow, ainda mais completo.

Referências

BARDI, Alessia; MANGHI, Paolo; ZOPPI, Franco. **Aggregative data infrastructures for the cultural heritage.** In: Research Conference on Metadata and Semantic Research. Springer, Berlin, Heidelberg, 2012. p. 239-251.

DIGITAL NEW ZEALAND. **This is Digital New Zealand**. YouTube, 20 dez 2018. Disponível em: https://www.youtube.com/watch?v=UWbIDwsaA4o. Acesso em 14 set 2019.

DIGITAL NEW ZEALAND. **Our History**. Disponível em: https://digitalnz.org/about/our-history. 2019. Acesso em 14 set 2019.

FINK, Eleanor E. **American Art Collaborative (AAC) Linked Open Data (LOD) Initiative: overview and recommendations for good practices**. *Am. Art Collab.*, 2018.

KOLLIA, Ilianna, TZOUVARAS, Vassilis, DROSOPOULOS, Nasos and STAMOU, Giorgos. **A Systemic Approach for Effective SemanticAccess to Cultural Content**. Semantic Web, v. 3, n. 1, p. 65-83, 2012.

NATIONAL LIBRARY OF AUSTRÁLIA. Trove Help Center. **Trove System Architecture Diagram**. 2010. Disponível em: https://www.nla.gov.au/trove/marketing/Trove%20architecture%20diagram.pdf. Acesso em 14 set 2019.

SANTOS, Daniel Soares. **Automatização de Processos de Negócios Utilizando BPM/BPMS.** Monografia (Ciência da Computação) - Universidade Estadual do Sudoeste da Bahia. Vitória da Conquista – Bahia, p. 109. 2013.

SCHOLZ, Authors Henning; FANTONE, Federica. Europeana Publishing. n. September, p. 1–31, 2018.

SECRETARÍA DE CULTURA. **Mexicana Repositorio del Patrimonio Cultural de México**. Dirección General de Tecnologías de la Información y Comunicaciones, Agenda Digital De Cultura. Colonia Cuauhtémoc. Ciudad de Mexico. 2018.

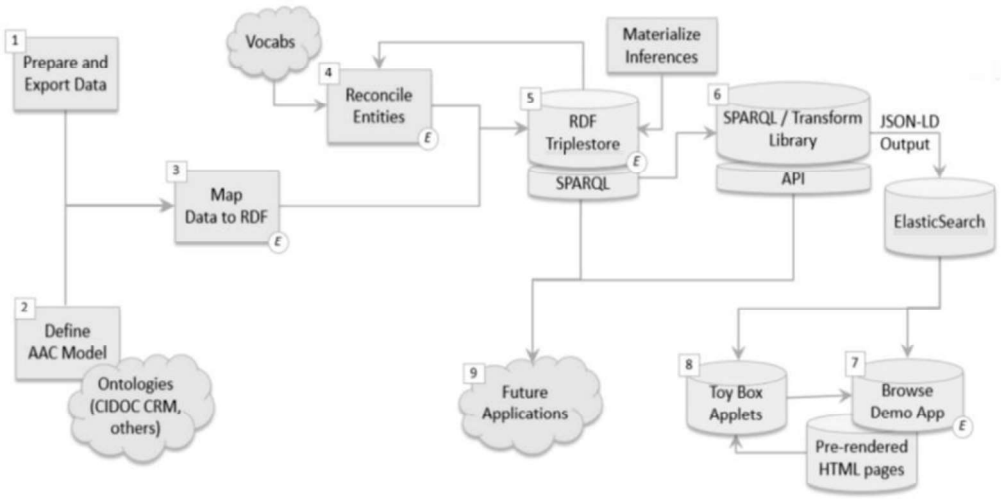
SOUTHWICK, Silvia B. **A guide for transforming digital collections metadata into linked data using open source technologies**. Journal of Library Metadata, vol. 15, no. 1, pp. 1–35, 2015.

TROVE HELP CENTRE. **About Trove.** 2019. Disponível em: https://help.nla.gov.au/trove/using-trove/getting-to-know-us. Acesso em 22 set 2019

SUPPLEJACK. **Architecture.** Documentation (Version 0.1). 2019. Disponível em: http://digitalnz.github.io/supplejack/architecture.html. Acesso em 22 set 2019.

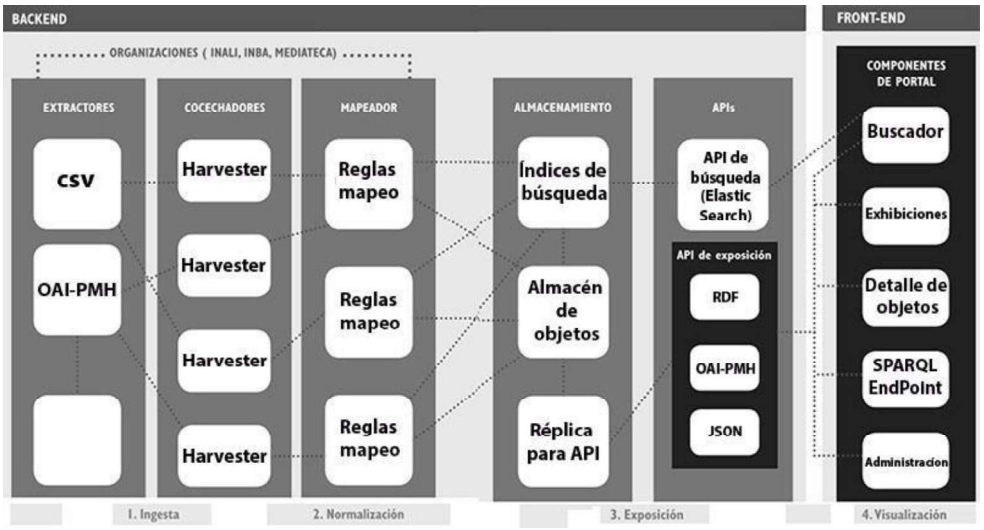
Apêndice A – Workflow de Agregação

Figura 01. Workflow de agregação proposto pela *American Art Collaborative*



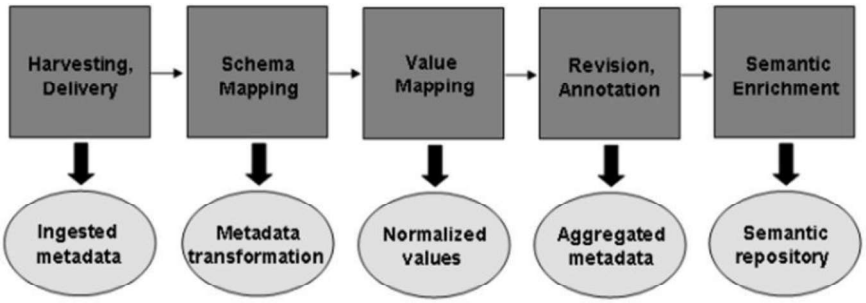
Fonte: Fink (2018), adaptada.

Figura 02. Workflow de agregação proposto pela Mexicana



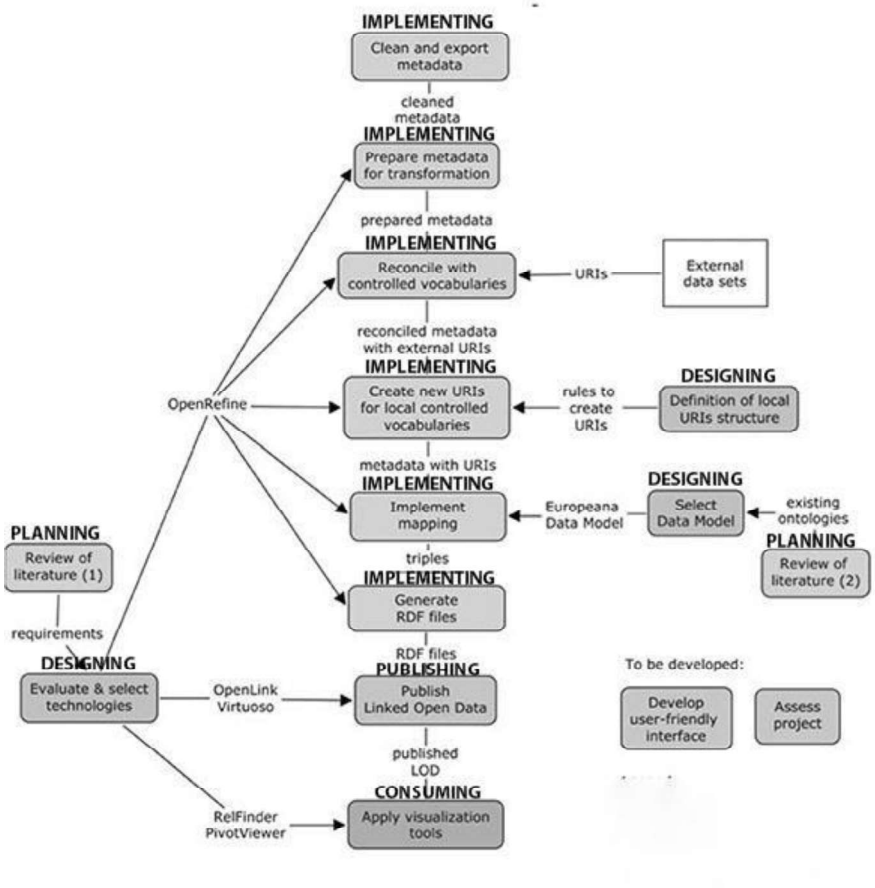
Fonte: Secretaría de Cultura (2018), adaptada

Figura 03. Workflow de agregação proposto pela Europeana



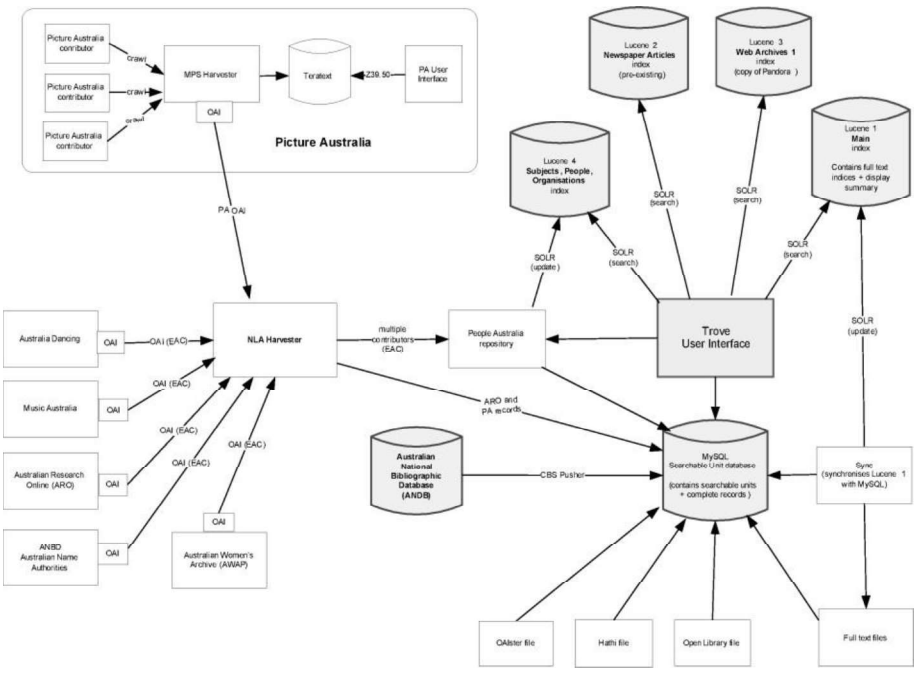
Fonte: Kollia et al (2012), adaptada

Figura 04. Workflow de agregação proposto pela University of Nevada



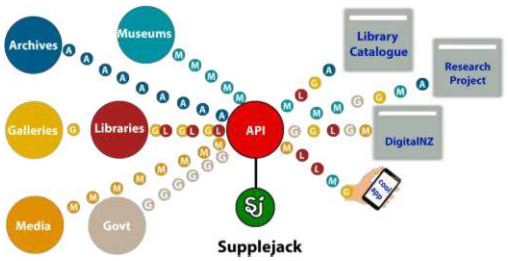
Fonte: Southwick (2015)

Figura 05. Workflow de agregação proposto pela TROVE



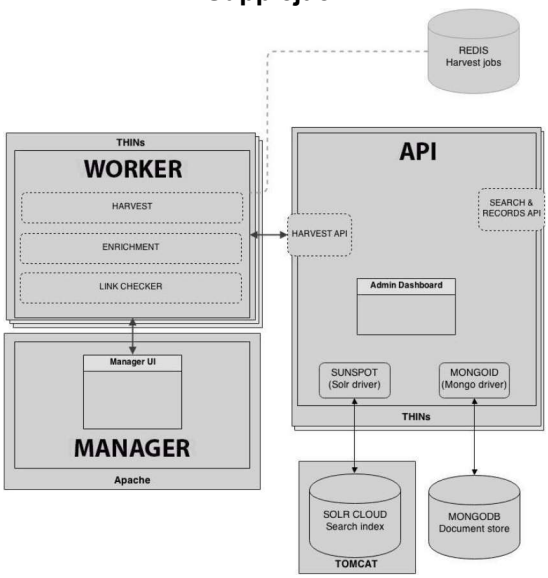
Fonte: National Library of Austrália (2010)

Figura 06. Workshop de agregação proposto pela DigitalNZ



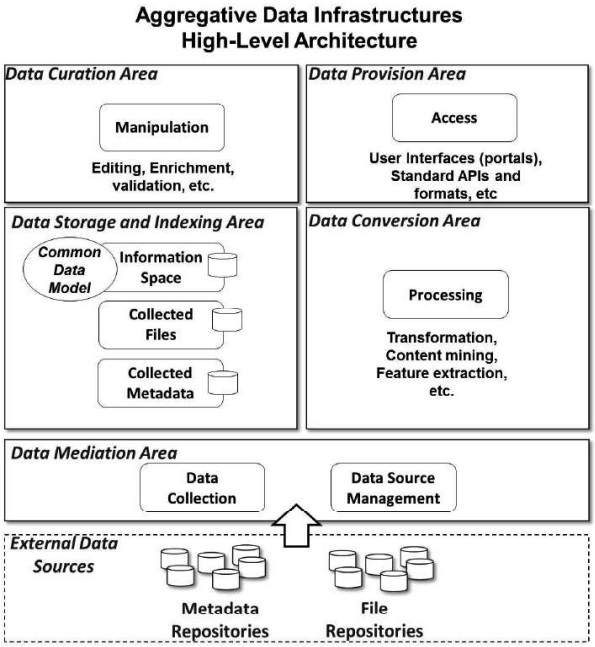
Fonte: Digital New Zealand (2018), adaptada

Figura 07. Arquitetura da plataforma Supplejack



Fonte: Supplejack (2019)

Figura 08. D-NET Software Toolkit, proposto pelo Istituto di Scienza e Tecnologie dell'Informazione



Fonte: Bardi, Manghi e Zoppi (2012), adaptada