WIDaT 2017

I WORKSHOP DE INFORMAÇÃO, DADOS E TECNOLOGIA

UNIVERSIDADE FEDERAL DE SANTA CATARINA - USFC

DE 04 A 06 DE SETEMBRO DE 2017, FLORIANÓPOLIS - SANTA CATARINA

ANAIS DO WIDAT'2017

FLORIANÓPOLIS | SANTA CATARINA

SETEMBRO - 2017

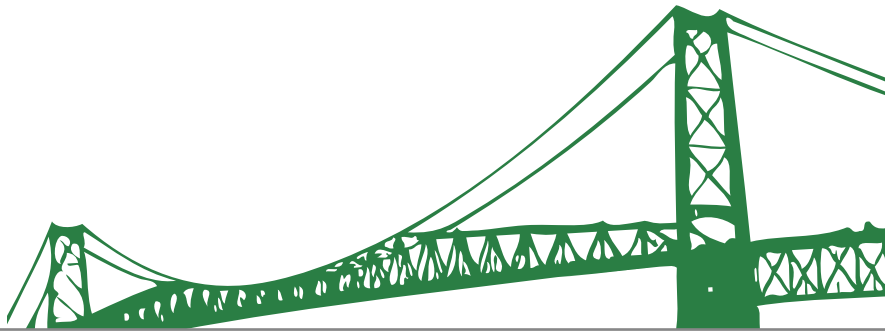
ORGANIZADORES:

MOISÉS LIMA DUTRA (UFSC)

DOUGLAS DYLLON JERONIMO DE MACEDO (UFSC)

ISBN: 978-85-64093-70-6

WIDaT 2017
I WORKSHOP DE INFORMAÇÃO, DADOS E TECNOLOGIA
Inteligência, Tecnologia e Informação - Research Group (ITI-RG) - UNIVERSIDADE FEDERAL DE SANTA CATARINA - USFC
DE 04 A 06 DE SETEMBRO DE 2017, FLORIANÓPOLIS - SANTA CATARINA



ANAIS DO WIDAT'2017

I WORKSHOP DE INFORMAÇÃO, DADOS E TECNOLOGIA

Organização:

Moisés Lima Dutra (UFSC)
Douglas Dyllon Jeronimo de Macedo (UFSC)

Realização:

Universidade Federal de Santa Catarina (UFSC)

Apoio:

Departamento de Ciência da Informação (CIN)
Programa de Pós-Graduação em Ciência da Informação (PGCIN)

Catalogação na fonte pela Biblioteca Universitária
da
Universidade Federal de Santa Catarina

W926a Workshop de Informação, Dados e Tecnologia (1 : 2017 : Florianópolis, SC).
Anais do WIDAT'2017 / I Workshop de Informação, Dados e Tecnologia ; organização: Moisés Lima Dutra, Douglas Dyllon Jeronimo de Macedo ; Universidade Federal de Santa Catarina, Programa de Pós-Graduação em Ciência da Informação. – Florianópolis : CIN/CED/UFSC, 2017.
127 p. : gráfs., tabs.

Inclui bibliografia.
ISBN: 978-85-64-093-70-6
Evento realizado 04 a 06 de setembro de 2017.

1. Ciência da informação. 2. Organização da informação. I. Dutra, Moisés Lima. II. Macedo, Douglas Dyllon Jeronimo de. III. Universidade Federal de Santa Catarina, Programa de Pós-Graduação em Ciência da Informação. IV. Título.

CDU: 007

Apresentação do Workshop de Informação, Dados e Tecnologia

O I Workshop de Informação, Dados e Tecnologia (WIDAT'2017) foi idealizado com o intuito de unir – através de abordagens interdisciplinares, as comunidades acadêmicas e industriais que trabalham com dados no Brasil, por meio da oferta de um espaço de discussão e interação entre industriais e pesquisadores das áreas de Ciência da Informação, Ciência da Computação, Engenharias e áreas afins. Este workshop pretende promover a aproximação de grupos que trabalham com esta temática de natureza interdisciplinar, que possui atualmente diversas iniciativas dispersas, mas possuidoras de grande potencial de integração entre si. O WIDAT'2017 será realizado entre os dias 04 e 06 de setembro de 2017, na cidade de Florianópolis (SC), no Auditório do Espaço Físico Integrado (EFI) da Universidade Federal de Santa Catarina (UFSC).

Num contexto de centralidade crescente do papel do dados nos cenários das mais diversas áreas do conhecimento humano, temas como Big Data, Data Science, Ambientes Inteligentes, Machine Learning, Segurança de Dados, Visualização da Informação, Modelos de Desenvolvimento Baseados em Dados, entre outros do mesmo porte, tornam-se estratégicos e fundamentais para o desenvolvimento de novos serviços/produtos e de novas aplicações inteligentes, que sejam capazes de entregar valor agregado para suprir as necessidades informacionais surgidas nestes cenários emergentes.

Além disso, acreditamos que deve ser destacada a forte utilização das Tecnologias da Informação e da Comunicação em todos os setores produtivos e na estruturação das novas tramas do tecido social brasileiro, na busca para reduzir as assimetrias existentes no acesso e utilização de dados e, por conseguinte, ampliando a produtividade e as condições de participação cidadã. Na dimensão acadêmica, existe ainda a necessidade de instrumentalização que emerge em todas as áreas de conhecimento que trabalham com as temáticas propostas para o WIDAT'2017, além da criação e do aperfeiçoamento de novas expertises associadas.

Sejam bem-vindos ao WIDAT!

Organizadores do WIDAT'2017

Coordenação de Programa do WIDAT'2017

O I Workshop de Informação, Dados e Tecnologia (WIDaT'2017) recebeu contribuições a respeito de reflexões teóricas, tecnologias, aplicações e estudos caso, mas não limitado para os seguintes temas: **Web Semântica:** Ontologias, Metadados, Linguagens, Linked Data. **Dados:** Ciclo de Vida dos dados, Dados Abertos, Qualidade dos Dados, Recuperação da Informação, Visualização. **Big Data:** Data Science, Ambientes e modelos, Big Data Analytics, Impacto e Evolução. **Aplicações em Tecnologia da Informação:** e-Health, e-Commerce e e-Learning, Internet das Coisas, Internet do Futuro, Ambientes Inteligentes (Cidades, Casas, Escolas, Rodovias, etc), Visualização da Informação. **Internet e a Sociedade da Informação:** Copyright na era digital, Curadoria de conteúdos, Humanidades Digitais, Privacidade, Serviços aplicados a Unidades de Informação, Social Media e Vigilância Tecnológica.

O WIDAT recebeu um total de 36 submissões de artigos completos para serem apresentados no evento. Foram recebidos artigos do Brasil (9 Estados) e do Canadá (1 Província). Gostaríamos de ressaltar que nesse ano tivemos muitos artigos de ótima qualidade, e só foram aprovados artigos com recomendações claras de aceitação. Logo, infelizmente, vários artigos com um bom potencial não estão nos anais. A programação do WIDAT neste ano conta com 3 sessões técnicas, com 18 apresentações orais de artigos completos, os quais serão publicados no Repositório Institucional da UFSC (https://repositorio.ufsc.br) e na página do evento (http://www.widat2017.ufsc.br).

Agradecemos imensamente a todos os membros do Comitê de Programa que colaboraram com a revisão dos artigos, a todos os colaboradores do Comitê de Organização Local, e também, em especial, aos autores que submeteram trabalhos neste ano, além de todos os convidados e participantes que vieram prestigiar o WIDAT.

Sejam bem-vindos ao WIDAT!

Douglas Dyllon Jeronimo de Macedo
Coordenador do Programa do WIDAT'2017

Comitês do WIDAT'2017

Coordenação Geral

Moisés Lima Dutra (UFSC)
Douglas Dyllon Jeronimo de Macedo (UFSC)

Coordenador do Comitê de Programa

Douglas Dyllon Jeronimo de Macedo (UFSC)

Membros do Comitê de Programa:

Adilson Luiz Pinto (UFSC)
Jose Eduardo Santarém Segundo (USP)
Moisés Lima Dutra (UFSC)
Márcio José Moutinho da Ponte (UFOPA)
Guilherme Ataíde Dias (UFPB)
Ricardo Sant'Ana (UNESP)
Celson Pantoja Lima (UFOPA)
Marcio Matias (UFSC)
Enrique Muriel-Torrado (UFSC)
Fábio Manoel França Lobato (UFOPA)
Sandro Rautenberg (UNICENTRO)

Comitê Local de Organização:

Priscila Basto Fagundes (PGCIN/UFSC)
Vitor Rozsa (PGCIN/UFSC)
Jean Fernandes Brito(PGCIN/UFSC)
Eduardo Silveira (PGCIN/UFSC)
Graciela Sardo Menezes (PGCIN/UFSC)

Sessões Técnicas do WIDAT'2017

Sessão Técnica I

Sessão Técnica II

Sessão Técnica III

WIDAT'2017 – Sessão Técnica I

Dia: 05/09/2017 – Horário: 10h30 às 12h30

UM *WORKFLOW* AUTOMATIZADO PARA COMPARTILHAMENTO DE DADOS CIENTÍFICOS PRIMÁRIOS BASEADO EM DADOS ABERTOS CONECTADOS

AN AUTOMATED WORKFLOW FOR SHARING PRIMARY DATA BASED ON LINKED OPEN DATA

Sandro Rautenberg[(1)], Alessandra Cassiana Burda[(2)], Lucélia de Souza[(3)]

(1) Universidade Estadual do Centro-Oeste (UNICENTRO), R. Simeão Varela de Sá, 03 - Vila Carli, Guarapuava - PR, 85040-080, srautenberg@unicentro.br.

(2) Universidade Estadual do Centro-Oeste (UNICENTRO), R. Simeão Varela de Sá, 03 - Vila Carli, Guarapuava - PR, 85040-080, alessandra.burda@gmail.com.

(3) Universidade Estadual do Centro-Oeste (UNICENTRO), R. Simeão Varela de Sá, 03 - Vila Carli, Guarapuava - PR, 85040-080, lucelia@unicentro.br.

Resumo: Investiga a automatização dos processos para a publicação de dados abertos científicos na *Web* de Dados. Metodologicamente, o trabalho é baseado no ciclo de vida *Linked Data Lifecycle* e suas tecnologias. Como resultado, apresenta-se um *workflow* automatizado para compartilhar dados primários. Conclui-se que o *workflow* é importante na preservação de dados científicos primários, suportando tanto as pesquisas científicas quanto o reuso de recursos sob os princípios de Dados Abertos Conectados.

Palavras-chave: Dados Abertos Conectados; *Workflow*; *Workflow* para Dados Abertos Conectados; Dados Primários.

Abstract: We investigate the automation of the processes for publishing scientific open data on the Web of Data. This work is based on the Linked Data Lifecycle and its technologies. As a result, a workflow is established for sharing primary datasets. As conclusion, we stand that this establishment is important for digital preservation of scientific data and can support scientific researches, considering the reuse of resources based on the Linked Open Data principles.

Keywords: Linked Open Data; Workflow; Workflow for Linked Open Data; Raw Data.

1 Introdução

A base constitutiva deste trabalho é alinhada ao que se entende por Dados Abertos Conectados (*Linked Open Data*) e as Melhores Práticas para a publicação desse tipo de recurso na *Web* de Dados. Em suma, os dados classificados como Dados Abertos Conectados são aqueles disponibilizados na *web* e regidos por licenças que advogam seu reuso por aplicações e em diversos contextos (OPEN KNOWLEDGE INTERNATIONAL, 2017; HEATH e BIZER, 2011).

Considerando as pesquisas científicas, principalmente, as que são financiadas com recursos públicos, pressupõe-se que seus dados primários devem ser compartilhados conforme os preceitos de Dados Abertos Conectados, primando pelo (re)uso de recursos em demais investigações.

Em consonância a essa visão, neste trabalho considera-se o esquema de implementação das 5 Estrelas para abertura de dados proposto por Tim Bernes-Lee. Objetiva-se o desenvolvimento de um *workflow* automatizado para publicação de dados científicos na *Web* de Dados, incrementando o grau de abertura. Neste prisma, os dados científicos devem ser publicados ao nível da 5ª Estrela (grau máximo de abertura de dados), tendo como característica principal o livre relacionamento a outros dados primários da pesquisa científica distribuídos na Internet.

Para apresentar o *workflow* proposto, além dessa seção introdutória, este artigo compreende as seguintes seções: (i) fundamentação teórica, a qual discorre sobre o conceito Dados Abertos Conectados; (ii) materiais e métodos, apontando as bases constitutiva e tecnológica do *workflow* proposto e os conjuntos de dados abertos cientométricos considerados na verficação; (iii) a apresentação do *workflow* e seus passos; (iv) verificação do *workflow*, reportando os esforços despendidos na publicação e na exploração dos índices cientométricos como Dados Abertos Conectados; e por fim (iv) considerações finais, discutindo as conclusões e trabalhos futuros.

2 Dados Abertos Conectados

Os Dados Abertos Conectados são aqueles publicados de acordo com licenças abertas, possibilitando que sejam reutilizados sem restrições, por pessoas ou aplicações e em diversos contextos. Constitutivamente, esta percepção é vinculada a dois entendimentos: (a) o que são dados abertos; e (b) o como os dados são conectados.

Os dados são considerados abertos quando "podem ser livremente usados, reutilizados e redistribuídos por qualquer pessoa - sujeitos, no máximo, à exigência de atribuição da fonte e compartilhamento pelas mesmas regras" (OPEN KNOWLEDGE INTERNATIONAL, 2017).

Ressalta-se que os dados abertos são classificados de acordo com seu nível de abertura e sua conexão a outros dados. Representada na Figura 1 (Apêndice A), essa classificação é denominada 5-Estrelas e é organizada como segue (5-STAR, 2017):

1ª **Estrela** - é atribuída aos dados que são publicados sob uma licença aberta (*Open License* - OL), entretanto, em um formato proprietário. Os dados somente podem ser manipulados (lidos, visualizados ou impressos) por determinados *softwares*.

2ª **Estrela** - é conferida à publicação de dados estruturados legíveis por máquinas (*Readable Machine* - RE). Os dados são processados por *softwares* proprietários e podem ser exportados em outros formatos.

3ª **Estrela** - é concedida aos dados que são publicados em formato aberto (*Open Format* - OF). A manipulação dos dados não necessita o uso de um *software* proprietário.

4ª **Estrela** - é designada à utilização dos Identificadores Uniforme de Recursos (*Universal Resource Identifier* - URI) para rotular os dados, permitindo que os usuários criem ligações e façam o reuso dos dados disponibilizados.

5ª **Estrela** - é atribuída aos dados que são conectados (*Linked Data* - LD) a outros dados em uma infraestrutura de rede. Isso permite a navegação entre dados e a descoberta de informação. Dessa forma, acrescenta-se valor aos dados ao fornecer uma contextualização mais ampliada.

Considerando a classificação anterior, a união de dados abertos com dados conectados é estabelecida ao se atingir a 5ª Estrela. Isso representa o ideal de publicação de dados na *web*. Ou seja, na *web*, os dados abertos podem estar conectados a outros dados, constituindo os Dados Abertos Conectados. Ressalta-se que essa união constitui a base informacional de um imenso grafo RDF, a *Web* de Dados. Neste sentido, a publicação de Dados Abertos Conectados tem como objetivo usar a arquitetura da *web* para compartilhar dados estruturados em uma escala global. Assim, incentiva-se o (re)uso do conjunto de dados universal por diferentes pessoas e aplicações.

No contexto deste trabalho, busca-se investigar os processos para o incremento dos níveis abertura dos dados, alcançando a 5ª Estrela. Para tanto, propõem-se um *workflow* automatizado baseado no ciclo de vida *Linked Data Lifecycle* e suas tecnologias (AUER, 2014), como descrito a seguir.

3 Materiais e Métodos

A definição do *workflow* proposto é inspirada em um subconjunto das atividades do ciclo de vida *Linked Data Lifecycle* e tecnologicamente suportado pelo *Linked Data Stack* (AUER, 2014). No *workflow*, são consideradas as atividades de Extração, Armazenamento, Enriquecimento e Exploração de Dados Abertos Conectados.

No que tange a verificação do *workflow*, três bases de dados abertos do domínio da Cientometria são consideradas, sendo elas:

- **Qualis**. Segundo WebQualis (2013), "Qualis é o conjunto de procedimentos utilizados pela CAPES para estratificação da qualidade da produção intelectual dos programas de pós-graduação". O Qualis afere a qualidade de produções científicas a partir da análise da qualidade dos periódicos científicos. Sua classificação compreende oito estratos em ordem decrescente de valor: A1, A2, B1, B2, B3, B4, B5 e C. O índice Qualis foi coletado ao longo dos últimos doze anos, a partir do Sistema WebQualis (WEBQUALIS, 2013) e da Plataforma Sucupira (SUCUPIRA,

2017). Cabe ressaltar que a preservação do índice Qualis como Dados Abertos Conectados foi discutida em Rautenberg e Burda (2016) e Rautenberg et al. (2016).

- **SJR (*SCImago Journal & Country Rank*)**. O *Journal SCImago & Country Rank* é um portal que disponibiliza informações cientométricas a partir de dados contidos na base de dados *Scopus*. Dentre as informações disponibilizadas, está o índice SJR, o qual pode ser utilizado para avaliar a qualidade e a reputação de periódicos científicos (JOURNAL METRICS, 2017). Este índice foi coletado no referido portal, em formato XLS (*eXceL Spreadsheet* - formato de planilha eletrônica da Microsoft), com o período de referência de 2005 a 2015.
- **SNIP (*Source Normalized Impact per Paper*)**. O índice SNIP é uma métrica que mede o impacto de citação contextual de uma comunicação científica, normalizando a distância interna das citações das comunicações de um periódico perante o universo das citações em uma área de conhecimento (JOURNAL METRICS, 2017). Em outras palavras, o SNIP é definido como a razão do impacto bruto de um jornal/revista por publicação e o potencial de citação nas áreas de conhecimento. Isto permite, por exemplo, a avaliação de uma revista em comparação com seus pares e fornece informações mais contextualizadas, dando uma melhor imagem do impacto em determinado domínio. O SNIP também foi coletado nos anos 2015 e 2017. A partir do Portal *Journal Metrics*, os dados primários são extraídos em formato XLS, com o período de referência de 2005 a 2015.

4 *Workflow*

Para elevar os conjuntos dados Qualis, SNIP e SJR ao nível de abertura da 5ª Estrela, um *workflow* (Figura 2 - no Apêndice B) é constituído com os seguintes passos:

- **Atividade 01 - Extração** – os arquivos em formato original são convertidos para arquivos texto. Alguns *scripts* de pré-processamento (na linguagem de programação PHP[1]) são empregados para organizar e criticar os dados.
- **Atividade 02 - Armazenamento** – os dados são armazenados em um Sistema Gerenciador de Banco de Dados Mysql[2] para serem usados por sistemas legados, por exemplo.
- **Atividade 03 – Enriquecimento** – os dados são extraídos de suas bases legadas e convertidos para arquivos no formato CSV (*Comma Separated Value*), alcançando a 3ª Estrela. Os dados também são mapeados para o formato RDF (*Resource Description Framework*) com o auxílio da ferramenta Sparqlify[3], atingindo a 4ª Estrela.
- **Atividade 04 - Armazenamento** – ao primar pelo (re)uso de dados científicos na *web*, os dados primários são compartilhados em um *endpoint* na *Web* de Dados implementado em um servidor *Open Link Virtuoso*[4] no endereço http://lod.unicentro.br/sparql (vide a Figura 3 no Apêndice C).
- **Atividade 05 – Exploração** - na *Web* de Dados, ao consultar os Dados Abertos Conectados, geralmente, objetiva a aquisição de informação contextualizada. Ao se relacionar recursos oriundos dos vários grafos RDF disponibilizados, alcança-se a 5ª Estrela.

5 Verificação

Originalmente, os dados abertos dos índices Qualis, SNIP e SJR são compartilhados na *web* em formatos proprietários. Neste sentido, considerando a Classificação 5-Estrelas, destaca-se que:

[1] É uma linguagem de uso geral, especialmente adequada para o desenvolvimento de aplicações *Web*. Disponível em: <http://www.php.net/>

[2] É um Sistema Gerenciador de Banco de Dados relacional *open-source* que pode ser usado em aplicações para gerir bases de dados. Disponível em: <https://www.mysql.com/>.

[3] É uma ferramenta *open-source* do Instituto *Agile Knowledge and Semantic Web* que enriquece os dados primários, convertendo os dados em triplas RDF. Disponível em: <http://aksw.org/Projects/Sparqlify.html>.

[4] Um sistema universal para acesso, integração e gerenciamento de dados baseados no modelo RDF. Disponível em: <http://virtuoso.openlinksw.com/>.

- os dados do índice Qualis capturados do Sistema WebQualis estavam na 1ª Estrela, no formato PDF;
- os dados dos índices SNIP e SJR são disponibilizados conforme a 2ª Estrela, no formato XLS; e
- a partir da Plataforma Sucupira, o índice Qualis é consumido no formato XLS.

Para a verificação do *workflow* proposto, procedeu-se da seguinte forma. A cada índice cientométrico considerado, uma instância do *workflow* é configurada com vistas à publicação dos recursos de dados na *Web* de Dados. Nas execuções das referidas instâncias: 829.577 avaliações Qualis; 514.828 avaliações SNIP; e 485.795 avaliações SJR foram disponibilizadas. Na Tabela 1 (Apêndice D) são sumarizados os recursos de dados compartilhados, ano a ano.

Ademais, o relacionamento dos recursos disponibilizados, fomentando o alcance da 5ª Estrela, é exemplificado no Apêndice E. Na Listagem 1 do referido Apêndice, encontra-se codificada uma consulta em SPARQL que relaciona os *scores* de determinado periódico. Já a Listagem 2 exemplifica parcialmente os recursos recuperados. Ressalta-se que consultas similares ao exemplo da Listagem 1 podem ser desenvolvidas e submetidas ao *endpoint* disponibilizado. Desta forma, por exemplo, com o auxílio de APIs (*Application Programming Intefaces*), permite-se a integração dos dados abertos em outras aplicações *web*.

6 Considerações Finais

Com os estudos de caso desenvolvidos, verifica-se a adequação do *workflow* proposto para publicar dados abertos científicos na *Web* de Dados. Admite-se que este estabelecimento colabora à preservação digital de demais dados científicos primários. Inspirando-se no *workflow* desenvolvido como um modelo tecnológico, pode-se compartilhar outros recursos de dados primários, baseando-se nos preceitos de Dados Abertos Conectados.

Por isso, como trabalho futuro vislumbra-se o uso do *workflow* proposto como base para: (i) o compartilhamento de outros conjuntos de dados científicos; e (ii) a curadoria digital dos dados já disponibilizados.

Agradecimentos

O autor principal agradece à Fundação Araucária pelo suporte financeiro (Projeto n° 601/2014 - Modelo para Compartilhamento de Informações sobre Pesquisas baseado em *Linked Open Data* para Estudos Cientométricos).

Referências

5-STAR. **5-Star OPEN DATA**. Disponível em: <http://5stardata.info/en>. Acesso em: 16 abr 2016 09:00.

AUER, S. Introduction to lod2. In AUER, S.; BRYL, V.; TRAMP, C (ed). **Linked Open Data – Creating Knowledge Out of Interlinked Data**. Springer-Verlag, 2014. 215p.

HEATH, T.; BIZER, C. **Linked Data Evolving the Web into a Global Data Space**. Londres: Morgan & Claypool, 2011. 136p.

JOURNAL METRICS. **Journal Metrics - Scopus.com**. Disponível em: <https://www.journalmetrics.com/>. Acesso em: 16 de Abril de 2017.

OPEN KNOWLEDGE INTERNATIONAL. O que são Dados Abertos? Disponível em: <http://opendatahandbook.org/guide/pt_BR/what-is-open-data/>. Acesso em: 14 jun 2017 21:00.

RAUTENBERG, S.; BURDA, A. C. Linked Open Data para Cientometria: Compartilhando e Mantendo o índice Qualis na *Web* de Dados In: ENCONTRO BRASILEIRO DE BIBLIOMETRIA E CIENTOMETRIA, 5., 2016, São Paulo. **Anais**... São Paulo: USP, 2016. p. A34.

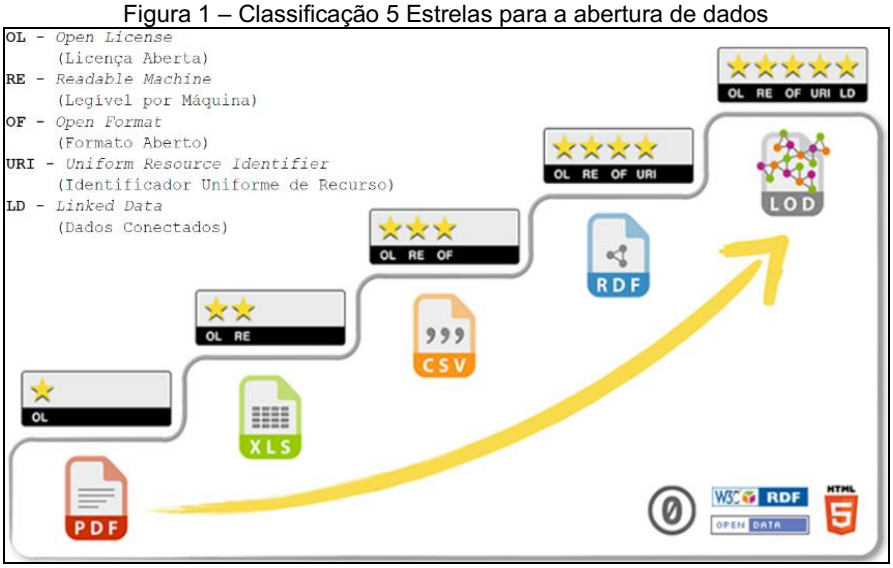
RAUTENBERG, S.; *et al*. Linked Data Workflow Project Ontology: uma Ontologia de Domínio para Publicação e Preservação de Dados Conectados. **Tendências da Pesquisa Brasileira em Ciência da Informação**, v. 9, p. 1-19, 2016.

SUCUPIRA. **Plataforma Sucupira**. Disponível em: <https://sucupira.capes.gov.br/sucupira/public/consultas/coleta/veiculoPublicacaoQualis/listaConsultaGeralPeriodicos.jsf>. Acesso em: 03 abr 2017 21:00.

WEBQUALIS. **Sistema WebQualis - Portal Capes**. Disponível em: <http://qualis.capes.gov.br/webqualis/principal.seam>. Acesso em: 25 ago 2013 10:00.

Apêndice A – Classificação 5 Estrelas para a abertura de dados

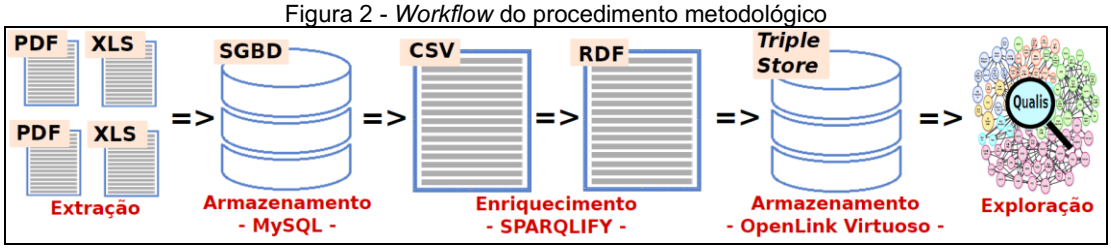
Figura 1 – Classificação 5 Estrelas para a abertura de dados



Fonte: adaptado de (5-STAR, 2017).

Apêndice B – Representação do *Workflow* automatizado

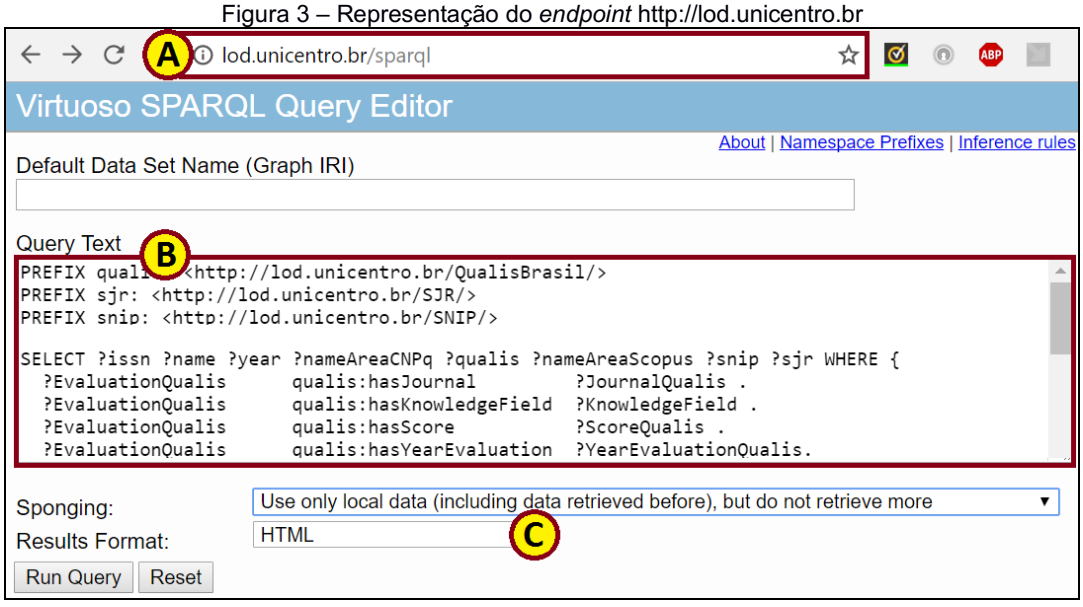
Figura 2 - *Workflow* do procedimento metodológico



Fonte: Dados da Pesquisa, 2017.

Apêndice C – Interface do *endpoint* http://lod.unicentro.br para consumo de dados

Figura 3 – Representação do *endpoint* http://lod.unicentro.br



Fonte: Dados da Pesquisa, 2017.

Apêndice C – Sumarização dos recursos compartilhados

Tabela 1 - Dados primários cientométricos compartilhados como *Linked Open Data* para pesquisas no domínio da Ciência da Informação

ANO	# AVALIAÇÕES QUALIS	# AVALIAÇÕES SNIP	# AVALIAÇÕES SJR
2005	35.020	32.932	26.881
2006	35.020	34.971	28.446
2007	35.020	37.183	30.049
2008	54.233	39.684	31.758
2009	54.233	42.984	34.074
2010	54.233	46.834	36.721
2011	107.429	51.448	54.577
2012	107.429	54.253	57.688
2013	107.429	56.360	60.019
2014	108.622	58.125	61.963
2015	44.463	60.054	63.619
2016[5]	86.446	--	--
TOTAL	829.577	514.828	485.795

Fonte: Dados da Pesquisa, 2017.

Apêndice D – Listagens da consulta e de resultado de processamento

Listagem 1 - Exemplo de consulta SPARQL que relaciona o periódico *Information Sciences* e seus índices cientométricos no ano 2015.

```
PREFIX qualis: <http://lod.unicentro.br/QualisBrasil/>
PREFIX sjr: <http://lod.unicentro.br/SJR/>
PREFIX snip: <http://lod.unicentro.br/SNIP/>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX bibo: <http://purl.org/ontology/bibo/>

SELECT DISTINCT ?issn ?name ?year ?nameAreaCNPq ?qualis ?nameAreaScopus ?snip ?sjr WHERE {
  ?EvaluationQualis      qualis:hasJournal         ?JournalQualis .
  ?EvaluationQualis      qualis:hasKnowledgeField  ?KnowledgeField .
  ?EvaluationQualis      qualis:hasScore           ?ScoreQualis .
  ?EvaluationQualis      qualis:hasYearEvaluation  ?YearEvaluationQualis.
  ?JournalQualis         bibo:issn                 ?issn .
  ?JournalQualis         foaf:name                 ?name .
  ?KnowledgeField        dc:title                  ?nameAreaCNPq .
  ?ScoreQualis           rdf:value                 ?qualis.
  ?YearEvaluationQualis  rdf:value                 ?year .

  ?EvaluationSJR      sjr:hasJournal         ?JournalSJR .
  ?EvaluationSJR      sjr:hasScore           ?ScoreSJR .
  ?EvaluationSJR      sjr:hasYearEvaluation  ?YearEvaluationSJR.
  ?JournalSJR         bibo:issn              ?issn .
  ?YearEvaluationSJR  rdf:value              ?year .
  ?ScoreSJR           rdf:value              ?sjr.

  ?EvaluationSNIP      snip:hasJournal         ?JournalSNIP .
  ?EvaluationSNIP      snip:hasScore           ?ScoreSNIP .
  ?EvaluationSNIP      snip:hasYearEvaluation  ?YearEvaluationSNIP.
  ?EvaluationSNIP      snip:hasSubAreaScopus   ?SubAreaScopus .
  ?SubAreaScopus       dc:title                ?nameAreaScopus .
  ?JournalSNIP         bibo:issn               ?issn .
  ?YearEvaluationSNIP  rdf:value               ?year .
  ?ScoreSNIP           rdf:value               ?snip.
  FILTER (?year = "2015" && ?issn = "0020-0255")
}
```

Fonte: Dados da Pesquisa, 2017.

Listagem 2 - Resultado parcial do processamento da consulta da Listagem 1.

```
"issn","name","year","nameAreaCNPq","qualis","nameAreaScopus","snip","sjr"
"0020-0255","Information Sciences","2015","MATEMÁTICA / PROBABILIDADE E
ESTATÍSTICA","B1","Artificial Intelligence","2.4890","2.5130"
[...]
```

Fonte: Dados da Pesquisa, 2017.

[5] Quando da escrita deste artigo, as avaliações SNIP e SJR do ano 2016 não estavam disponíveis.

Uso de software livre para disseminação e análise de dados abertos governamentais

Use of open source software for dissemination and analysis of open government data

Lucas Rodrigues Costa, Lucas Ângelo Silveira, Ronnie Fagundes Brito e Milton Shintaku
Instituto Brasileiro de Informação em Ciência e Tecnologia - Ibict, lucasrodrigues, lucasangelo, ronniebrito, shintaku {@ibict.br}

Resumo:
O estudo apresenta um modelo voltado à disseminação e análise de dados de órgãos governamentais e desenvolvido com software livre. Devido a grande oferta de ferramentas foi adotada uma metodologia voltada a seleção de tecnologias robustas que atendessem às demandas destas agências. Desta forma, o modelo baseou-se nos softwares *Comprehensive Knowledge Archive Network* (CKAN) e Pentaho, os quais possibilitam serviços de depósito, recuperação, visualização e análise dos dados. O caso de estudo foi aplicado na Secretaria Nacional da Juventude, onde dados oriundos de várias fontes foram recolhidos, tratados e publicados para o público em geral. O modelo atende aos objetos dos órgãos governamentais, como a oferta de acesso a dados brutos e sua análise pelo público em geral, contribuindo assim com a transparência de seus dados.
Palavras-chave: Softwares livres; dados abertos; *data warehouse*.

Abstract:
This paper presents a model for dissemination and analysis of government agencies data and which is developed with free software. Due to the great offer of tools, it followed a methodology focused on the selection of robust technologies that met agencies requirements. Thus, it resulted in a model based on the Comprehensive Knowledge Archive Network (CKAN) and Pentaho software, which offer deposit, retrieval, visualization and data analysis services. A case study was developed at the National Youth Secretariat, where data from various sources were collected, processed and published for the general public. The model attends government agencies, offering access to raw data and its analysis by the general public, contributing to transparency of the government data.
Keywords: Open source; Open data; data warehouse.

1. Introdução

No Brasil a transparência do estado tem sido promovida com várias ações, entre os quais situa-se a iniciativa para dados abertos governamentais, processo pelo qual os governantes disponibilizam informações aos seus cidadãos (OLIVERIO, 2011). Essa orientação governamental é regida pela Lei nº 12.527, que regula o acesso livre à informação governamentais, reservadas às questões de segurança e proteção dos dados sensíveis. Essa lei engloba desde dados brutos à documentos completos, tratando questões como sigilo, autenticidade, integridade e primariedade, entre outros pontos (BRASIL, 2011).

Em uma análise detalhada, Dados governamentais são descritos como resultados de atividades dos órgãos públicos e podem estar contidos em bases de dados, documentos impressos ou digitais, entre outros (OLIVEIRA, 2016), figurando assim uma grande variedade de formas e formatos. Por sua vez, dados abertos podem ser lidos, utilizados e disseminados de forma livre, sendo requisitos a citação da fonte e se for o caso o compartilhando sob mesmo tipo de licença (ISOTANI; BITTERNCOURT, 2015). Assim, dados abertos governamentais são dados abertos gerados pelo governo e que devem ser disseminados com a sociedade.

Palazzi e Tygel (2014) afirmam que parte desses dados gerados no governo são de cunho estatístico, possibilitando a geração de indicadores, que podem ser utilizados para análises e tomada de decisão. Dessa forma, para a disseminação e análise dos dados governamentais, pode-se utilizar sistemas informatizados que possibilitem a implementação de políticas relacionadas iniciativa de dados abertos governamentais para a disseminação e análise dos dados armazenados e produzidos do governo.

Com isso, tem-se oportunidades e desafios relacionados à pesquisas que atendam às necessidades das instituições e

órgãos às suas especificidades ao tratamento dos dados abertos governamentais.

Neste trabalho, apresenta-se o resultado de pesquisa efetuado na Secretaria Nacional de Juventude voltado à criação de um modelo de dados abertos governamentais utilizando software livre. Isso, contribui com a discussão sobre dados abertos governamentais e a disseminação de informação por meio de software livre.

2. Metodologia

O presente estudo tem aspectos aplicados, com utilização de técnicas ligadas à ciência da computação, na seleção e uso de tecnologias, para criação e aplicação de um modelo voltado a dados abertos com o uso de software livre. Assim, utiliza-se conceitos e técnicas voltados à avaliação de ferramentas informatizadas, alinhado à técnicas ligadas a qualidade de software. Nesse sentido, qualidade é entendida na forma da Norma ISO 8402, na qual se refere ao conjunto de características relacionadas ao atendimento das necessidades dos usuários, sejam explícitas ou implícitas. Atendendo os requisitos registrados ou não, a qualidade do software é uma avaliação de aspectos quantitativos e qualitativos, no qual contempla o processo e produto, como advogado por Tsukumo et al (1997), na medida em que o processo oferta certas garantias ao produto.

Para Silva (2007), a avaliação de softwares livres não deve ser embasada apenas na gratuidade da ferramenta, mas nos benefícios gerais que a nova tecnologia pode trazer. Seguindo tal raciocínio, o critério básico de seleção de ferramentas a serem utilizadas, era ser software livre e que provesse funcionalidades tais como: suporte por comunidade internacional atuante, mantida por instituição confiável, e que fornecesse certas garantias de sustentabilidade.

3. Ferramentas utilizadas

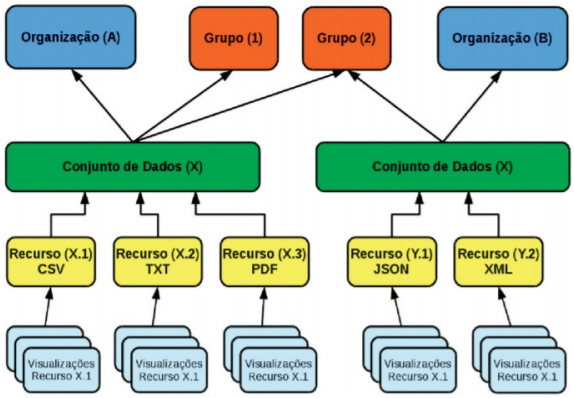
A escolha adequada das ferramentas utilizadas para a disseminação e análise dos dados do governo tem o intuito de auxiliar a iniciativa de dados abertos governamentais com o propósito de divulgar dados e informação.

Seguindo as linhas de software livres e os benefícios fornecidos chegou-se ao uso do sistema de repositórios *Comprehensive Knowledge Archive Network* - CKAN e o Pentaho. Nas próximas seções, uma descrição detalhada de ambos os softwares será realizada.

3.1. CKAN

O CKAN (CKAN, 2017) possibilita o depósito de bases de dados de forma organizada. Conforme a hierarquia descrita por Costa et al (2017) na figura 1, tem-se os conjuntos de dados como ponto central ligados às organizações. O conjunto de dados é caracterizado e descrito por metadados, podendo conter um ou mais recursos, em formato diversos. Com isso, caso tenha-se uma base de dados com arquivos em formato textuais, planilhas entre outros. Estes dados podem ser agrupados em uma única base de dados ligados a uma organização, tendo em vista que um único CKAN pode ser utilizado por uma ou mais instituições. Além disso, pode-se criar organizações artificiais para agrupar uma ou mais bases de dados denotadas como grupos e com isso facilitar a recuperação dos dados.

Figura 1: Possíveis hierarquias de bases de dados no CKAN



Este modelo alinha-se às indicações de Correa et al. (2017) as quais descrevem o uso do CKAN como uma ferramenta que apoia a disseminação de dados

governamentais, contribuindo em parte com a transparência do governo. Cabe ressaltar que o CKAN está de acordo com as orientações dos dados abertos governamentais e a Lei nº 12.527, que regula o acesso livre à informação de governo.

3.2. Pentaho

No que se refere a análise dos dados, optou-se pela ferramenta Pentaho, visto que é uma ferramenta robusta para *Business Intelligence* (BI) (AHISHAKIYE, 2017) com foco no tratamento e análise de dados (COSTA et al., 2017; MARINHEIRO; BERNARDINO, 2015).

O Pentaho oferece ferramentas de *Online Analytical Processing* (OLAP) que permitem analisar grandes volumes de dados de forma online e sob diferentes cruzamentos e dimensões dos dados, bem como a realização de cálculos complexos. O principal artefato envolvido em uma análise por meio de OLAP é o cubo multidimensional. Vale ressaltar, que a ferramenta é uma solução composta por vários módulos a fim de atender diferentes etapas de estruturação de uma base analítica. Entre os módulos disponíveis no Pentaho está o *Pentaho Data Integration* (PDI), o *Pentaho Schema Workbench* (PSW- MONDRIAN) e o SAIKU.

- **PDI -** tem por finalidade a integração de dados, possibilitando cruzar dados proveniente de várias fontes, com o uso de técnicas de ETL (extract-transform-load). O PDI oferece uma interface gráfica para a conexão das várias fontes e possibilita apresentar os resultados do processamento em formas de grafos.
- **PSW - MONDRIAN -** é uma ferramenta para o desenvolvimento de um esquema xml que descreve o cubo multidimensional dos dados.
- **SAIKU -** é um módulo para visualização dos dados do cubo de uma forma amigável e dinâmica.

4. Resultados e Discussões

A Secretaria Nacional da juventude (SNJ) possui um fluxo de dados no qual coleta ou gera uma grande quantidade de dados brutos sobre juventude, nos mais diversos temas. Conforme descreve Cury (2007), essa secretaria nasceu de uma ação interministerial, devido ao seu caráter multifacetado. O mesmo autor, relata que o tema juventude é novo na política no mundo, revelando a inovação desta secretaria e por consequência suas ações.

Após o levantamento de requisitos, verificou-se que a principal necessidade da SNJ era de um modelo que contemplasse a gestão de bases de dados estatísticas, com o fluxo informacional de Coleta ou Geração; Catalogação; e finalmente sua Recuperação/Análise. Os dados em sua maioria, são provenientes de outras instituições como o Instituto Brasileiro de Geografia e Estatística (IBGE) e Instituto de Pesquisa Estatística Aplicada (IPEA), ou gerado para o SNJ, por instituições como a Caixa Econômica Federal. Assim no primeiro caso considera-se como coleta e no segundo uma geração. Requerendo, dessa forma, uma ferramenta que possibilite o armazenamento organizado das bases de dados, permitindo a recuperação e análise dos dados.

A prospecção focou em duas etapas: a primeira consiste na migração dos dados de seus diferentes formatos para uma plataforma comum, e uma segunda que trata de disponibilizar uma ferramenta de visualização de dados e elaboração de relatórios. Pela prospecção de tecnologias foi utilizado o uso integrado do CKAN e do Pentaho.

Com isso, tem-se o CKAN como um repositório de dados, com todas as funcionalidades voltadas à gestão de bases de dados e o Pentaho como uma ferramenta com foco em analisar, organizar e apresentar tais dados.

4.1. CKAN na SNJ

Na SNJ as organizações artificiais (grupos) do CKAN tornaram-se temas de interesse da secretaria, como: saúde, educação, lazer, onde as organizações são organizadas como sendo diferentes fontes para base de dados. Com isso, tem-se um modelo de repositório para dados governamentais, organizado de forma temática, com três níveis, sendo: Tema → base de dados → recursos, possibilitando

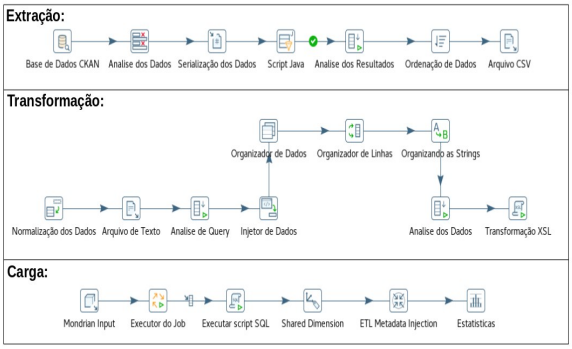
catalogar todas as bases de dados de forma organizada, facilitando a recuperação.

O CKAN neste modelo apresenta-se como repositório de dados tornando-se útil para a catalogação das bases de dados em uma estrutura organizada e integrável com outros sistemas CKAN na forma de um ecossistema de dados governamentais. Todavia, por ter um único órgão depositante de dados, tornou-se um repositório de dados institucional e temático.

4.2. Pentaho na SNJ

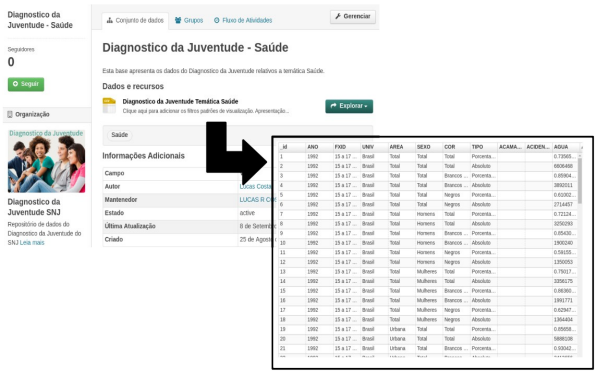
Inicialmente foi desenvolvido uma estratégia para a extração dos dados armazenados no CKAN e uma transformação dos mesmos para um formato comum, tendo em vista que diferentes bases e tipos de dados foram armazenados no repositório. Para essa tarefa foi utilizado o módulo PDI que possibilitou cruzar dados proveniente de diversas fontes do CKAN por meio das técnicas de ETL. A interface gráfica para a conexão de tais fontes auxiliou na apresentação das tarefas resultando nos processos descritos na figura 2.

Figura 2: ETL utilizado no processo da SNJ.



Como pode ser visto, foram separados as três etapas do processo ETL realizadas no PDI. A extração começa na base de dados do CKAN a qual possuem inúmeras tabelas com informações de diversas áreas relacionadas à juventude. A figura 3 mostra um exemplo do conjunto de dados relacionado a saúde da juventude no Brasil e como se encontra o formato dos dados.

Figura 3: Exemplo de conjunto de dados da SNJ.

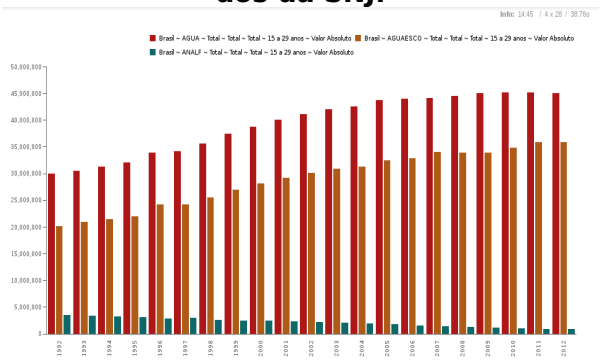


Ainda na figura 2, após a extração, tem-se o processo de transformação dos dados. Nesta etapa todas os dados são convertidos para um formato e estrutura comum. Em seguida, tem-se a parte da carga em que realiza-se a inserção de dados em um banco de dados relacional com uma estrutura específica para a criação do cubo.

Posteriormente, o PSW-MONDRIAN foi utilizado para a criação do cubo, o qual permite análise multidimensional de dados podendo-se desenvolver diferentes soluções para exploração do repositório de dados por meio de uma ferramenta de análise para a produção de informações em diversos formatos, tais como, tabela, gráficos e relatórios.

Por fim, utiliza-se o SAIKU para a visualização dos dados conforme a orientação do cubo possibilitando a busca de informação de forma dinâmica. A figura 4 mostra um exemplo em que foram relacionados três bases de dados do CKAN, relacionando jovens com acesso adequado a água, saneamento e sua relação com a taxa de analfabetismo. A informação extraída deste gráfico mostra que jovens com acesso a água e saneamento apresentam taxa de analfabetismo menor.

Figura 4: Exemplo de conjunto de dados da SNJ.



Vale ressaltar que as consultas podem ser realizadas de forma dinâmica cruzando vários tipos de dados em diversos tipos de gráficos, relatórios e planilhas. Dessa forma, com o SAIKU, pode-se reutilizar os dados do CKAN para geração de novas informações, incrementando as possibilidades de uso das bases de dados mantidas no repositório.

O modelo apresentado no presente estudo encontra semelhanças na experiência de Tygel (2012), na medida em que utiliza o Pentaho para análise de dados abertos governamentais. De Faria Cordeiro et al (2011) utilizaram o Pentaho para integrar uma base de dados para uso de ferramentas semânticas. Mendonça, Cruz e Campos (2014) utilizaram o Pentaho como parte de um modelo para tratamento de dados governamentais. Dos Santos e da Silva (2014) comparam o sistema I-GOV com o uso do Pentaho para integrar dados governamentais. Portanto, o Pentaho revela-se apropriado ao apoio à análise de dados governamentais de características estatísticas.

O resultado do caso de estudo pode ser encontrado em: http://magonia.ibict.br/ckan/consultas-livres.

5. Considerações Finais

O artigo apresenta um estudo de caso com os dados da Secretaria Nacional da Juventude (SNJ) para a colaboração da proposta.

A contribuição do modelo utilizado no presente estudo está no uso de duas ferramentas livres (CKAN e o Pentaho) para atender as necessidades da SNJ para com às orientações dos dados abertos governamentais e ofertar aos gestores de políticas de juventude bem como a sociedade em geral, um sistema web com oferta de dados brutos ou pré-analisados. O modelo desenvolvido e implementado na SNJ composto pelo CKAN e Pentaho se apresentou eficaz, dinâmico podendo ser implementado em outros órgãos de governo, contribuindo com a discussão do uso de ferramentas livres em órgãos públicos. Além de se apresentar eficaz no atendimento aos objetivos do estudo. O CKAN possibilita a criação de um repositório, com a catalogação de bases de dados de forma organizada. Para facilitar a recuperação o Pentaho possibilita cruzar dados das bases armazenadas no CKAN para a integração e elaboração de novas informações.

Apresentou recursos tecnológicos que apoiam a gestão estratégica e eficiente das organizações. Para isso é proposto uma metodologia de integração baseada nos softwares livres CKAN e Pentaho para ajuda na tomada de decisões organizacionais.

A metodologia dos dois sistemas combinados é capaz de fortalecer o plano de atuação das organizações através da geração de informações rápidas, precisas e personalizáveis garantindo uma estruturação de gestão diferenciada para a melhora no processo de tomadas de decisões pelos gestores organizacionais.

6. Referências

ALBANO, Cláudio Sonaglio. Dados governamentais abertos: proposta de um modelo de produção e utilização de informações sob a ótica conceitual da cadeia de valor. 2014. Tese de Doutorado. Universidade de São Paulo.

AHISHAKIYE, Emmanuel et al. Comparative Analysis of Open source Business Intelligence tools for Crime Data Analytics.

BRASIL. Lei nº 12.527, de 18 de novembro de 2011 .

CKAN. Documentation. 2013. Disponível em: <http://docs.ckan.org/en/latest/>. Acesso em: 18 jul. 2017.

COSTA, Lucas Rodrigues et al. Guia do usuário CKAN. IBICT. 2017.

COSTA, Evandro B. et al. Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. Computers in Human Behavior, v. 73, p. 247-256, 2017.

CORRÊA, Andreiwid Sh et al. Transparency and open government data: a wide national assessment of data openness in Brazilian local governments. Transforming Government: People, Process and Policy, v. 11, n. 1, p. 58-78, 2017.

CURY, BETO. Admirável mundo novo. In: KEHL, Maria Rita. A juventude como sintoma da cultura. Revista de saberesmandato vereador Arnaldo Godoy, p. 44-55, 2007.

DE FARIA CORDEIRO, Kelli et al. An approach for managing and semantically enriching the publication of Linked Open Governmental Data. In: Proceedings of the 3rd workshop in applied computing for electronic government (WCGE), SBBD. 2011. p. 82-95.

DE OLIVEIRA, Carolina. A gestão arquivística de documentos como apoio à publicação de dados governamentais abertos. Acervo, v. 29, n. 2 jul-dez, p. 168-178, 2016.

DA SILVA, José Fernando Modesto; SUBSÍDIO À GESTÃO BIBLIOTECÁRIA. CBBD.

DOS SANTOS, João Paulo Clarindo; DA SILVA, Fábio José Coutinho. IGOV: um sistema de integração de dados governamentais. Revista Brasileira de Administração Científica, v. 5, n. 2, p. 8-16, 2014.

ISOTANI, Seiji; BITTENCOURT, Ig Ibert. Dados Abertos Conectados: Em busca da Web do Conhecimento. Novatec Editora, 2015.

MARINHEIRO, Antonio; BERNARDINO, Jorge. Experimental evaluation of open source business intelligence suites using OpenBRR. IEEE Latin America Transactions, v. 13, n. 3, p. 810-817, 2015.

DE MENDONÇA, Rogers Reiche; CRUZ, S. M. S.; CAMPOS, Maria Luiza M. Gerência de proveniência multigranular em linked data com a abordagem etl4linkedprov. Anais do Simpósio Brasileiro de Bancos de Dados. Paraná, Brazil, 2014.

OLIVERIO, Marcio Araujo. Governo aberto como ferramenta de comunicação entre o governo e o cidadão. In: XXXIV Congresso Brasileiro de Ciências da Comunicação. Recife, PE. 2011.

PALAZZI, Daniele; TYGEL, Alan. Visualização de Dados Estatísticos Representados como Dados Abertos Ligados. 2014.

TSUKUMO, Alfredo N. et al. Qualidade de software: visões de produto e processo de software. II ERI–SBC, Piracicaba, São Paulo, Brasil, 1997.

TYGEL, Alan. Representação e Visualização de dados estatísticos: os desafios dos dados abertos ligados. 2012

UMA REVISÃO PRELIMINAR SOBRE A DIFUSÃO DE DADOS CONECTADOS NO ÂMBITO EMPRESARIAL

A PRELIMINARY REVIEW ON THE DIFFUSION OF LINKED DATA IN THE ENTERPRISE

Murilo Silveira Gomes [(1)], Lidiane Visintin [(1)], Fernando Álvaro Ostuni Gauthier[(2)]

(1) Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento, Universidade Federal de Santa Catarina, {lilo.flp lidiane.visintin}@gmail.com.

(2) Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento, Universidade Federal de Santa Catarina, fernando.gauthier@ufsc.br.

Resumo: As empresas necessitam estar atentas as mudanças no cenário mundial, para que possam responder rapidamente as mudanças de mercado. Para isso, a Web contribuiu possibilitando dinamismo, impactando com isso em um aumento significativo no volume de dados, que podem ser explorados, com o intuito de se obter benefícios. O Linked Enterprise Data é apresentado como um formato de dados que pode auxiliar os gestores na tomada de decisão, oportunizando explorar dados internos e externos a empresa, porém este conceito não é muito explorado no âmbito empresarial. Por este motivo o presente estudo objetiva-se em discutir os fatores que impedem a difusão de Linked Data neste contexto. Deste modo, observa-se que a cultura organizacional e alinhamento estratégico, entre outros, são alguns dos fatores encontrado para que não ocorra a difusão do Linked Enterprise Data. O que permite concluir que Linked Enterprise Data está começando a chamar atenção das empresas, mas se faz necessário mais trabalhos que abordem as reais contribuições ao meio empresarial, para difundir o Linked Data neste contexto.
Palavras-chave: Empresa; Dados Conectados; Fatores de Impacto.

Abstract: Companies need to be aware of changes in the global landscape, so they can respond quickly to market changes. For this purpose, the Web has contributed to the dynamism, thus impacting on a significant increase in the volume of data that can be exploited in order to obtain benefits. Linked Enterprise Data is presented as a concept that can help managers in decision making, allowing the exploration of internal and external data to the enterprise, but this concept is not much explored in the business scope. For this reason, the present study aims at discussing the factors that prevent the diffusion of Linked Data in this context. In this way, it is observed that the organizational culture and strategic alignment, among others, are some of the factors found so that the dissemination of Linked Enterprise Data does not occur. This allows us to conclude that Linked Enterprise Data is beginning to attract attention from companies, but more work is needed to address the real contributions to the business environment, to disseminate Linked Data in this context.
Keywords: Enterprise; Linked Data; Impact Factors.

1 Introdução

As empresas têm a necessidade de se tornarem dinâmicas, devido as mudanças que ocorrem nos mais diversos segmentos de mercado (HU; SVENSSON, 2010). Para isso otimizar o uso de seus dados e explora-los, pode possibilitar reagir as rápidas mudanças, gerando benefícios para a tomada de decisão.

Há algumas soluções que buscam auxiliar na integração de dados e informações, bem como no processo de tomada de decisão, sendo algumas destas soluções: MDM (Master Data Management), SOA (Service-Oriented Architecture) e BI (Business Inteligence), mas o grande desafio ainda se encontra em como integrar grandes quantidades de dados heterogêneos (ANTIDOT, 2017). Por sua vez, os dados conectados (Linked Data) podem minimizar tal desafio e dar semântica aos dados (BERNERS-LEE, 2001).

O uso dos princípios de dados conectados no contexto empresarial é interessante, pois está área de pesquisa promete integrar e consolidar fontes de dados heterogêneos, possibilitando analisar dados empresariais, com dados disponibilizados na Web, reutilizando de vocabulários conhecidos, como: *Friend-*

of-a-Friend (FOAF), *Dublin Core* (DC) e *Simple Knowledge Organization System* (SKOS) (WEICHSELBRAUN; STREIFF; SCHARL, 2015).

Analisado o contexto apresentado e verificando que o conceito de dados abertos conectados tem ganharam grande ênfase no setor público, este trabalho tem por objetivo discutir quais são os principais fatores que impedem a difusão de dados conectados no âmbito empresarial.

Na sequência, são apresentadas as seções deste trabalho: na seção 2 aborda-se o contexto de *Linked Enterprise Data*; na seção 3 é apresentada uma breve descrição dos procedimentos metodológicos; na seção 4 apresenta-se uma discussão preliminar dos resultados obtidos com a análise realizada; e por fim, na seção 5 são apresentados os encaminhamentos para trabalhos futuros e considerações finais.

2 Linked Enterprise Data

O Linked Enterprise Data (LED) é uma extensão do conceito de dados conectados, com foco empresarial. Que é definido como: "uma estrutura base para incorporar tecnologias semânticas em ambientes empresariais de TI" (GALKIN; AUER; SCERRI, 2016). Contudo, esse tema é apresentado como uma alternativa para que as empresas possam se libertar das tecnologias tradicionais. Assim como podem obter benefícios como: interoperabilidade e homogeneidade, proveniência, agilidade, coerência, acesso, identificação e governança (GALKIN; AUER; SCERRI, 2016).

O LED também possibilita que as empresas possam consumir os dados presentes na nuvem de dados abertos conectados (Linked Open Data), ou seja, podem fazer uso de *datasets* disponibilizados com licença aberta, com o intuito de obter uma nova percepção de análise de dados (GLACHS; SCHAFFERT; BAUER, 2012). Outro ponto destacado na literatura é que com o uso deste conceito as empresas podem obter um alto desempenho sobre os sistemas de informação em geral (LI; ZHAi, 2016).

3 Procedimentos Metodológicos

Para a realização desta revisão literatura preliminar, utilizou-se de 6 artigos que abordam o tema *Linked Enterprise Data*, sendo que os mesmos são apresentados entre os anos de 2010 a 2016. Os artigos foram selecionados com base no número de citações e priorizado alguns dos últimos anos, devido a fornecer uma visão mais atual das pesquisas na área. No Quadro 1 apresenta-se os artigos selecionados, destacando o título, ano de publicação, a quantidade de citações e os autores, apresentados na ordem de maior número de citações.

Quadro 1 – Artigos Selecionados para analise

Autores	Titulo	Ano	Qtd. Citação	Objetivo
HU, SVENSSON	A Case Study Of Linked Enterprise Data	2010	10	Apresenta um estudo piloto em uma organização internacional, tem por objetivo criar um espaço compartilhado que proporcione a integração de dados inter-organizacionais em Linked Data.
GRAUBE et al.	Flexibility Vs. Security In Linked Enterprise Data Access Control Graphs	2013	3	Apresentar o projeto ConVantage, focado na segurança dos dados em Linked Data em domínio empresarial.
BIANCHINI ANTONELLIS MELCHIORI	A Linked Data Perspective For Collaboration In Mashup Development	2013	3	Apresenta a abordagem e o protótipo LINKSMAN(Linked Data Supported Mashup Collaboration) baseando-se na integração de dados internos e externos a empresa utilizando de

				Linked Data, a fim de localizar novos colaboradores.
WEICHSELBRAUN STREIFF SCHARL	Consolidating Heterogeneous Enterprise Data For Named Entity Linking And Web Inteligence	2015	2	Apresenta o componente Recognyze, que tem por objetivo identificar entidades nomeadas com uso de bases de dados conectados, a fim de promover a interoperabilidade de conjunto de dados públicos abertos.
TANEJA et al.	Linked Enterprise Data Model And Its Use In Real Time Analytics And Context-Driven Data Discovery	2015	2	Aborda as deficiências para o gerenciamento de IoT e apresentam a abordagem desenvolvida pelos autores denominada de LEDM (Linked Enterprise Data Model) desenvolvida para atender o domínio de IoT com foco em Big Data. O LEDM é baseado nos princípios de Linked Data com foco na interoperabilidade entre sistemas e subsistemas.
GALKIN AUER SCERRI	Enterprise Knowledge Graphs - A Backbone Of Linked Enterprise Data	2016	0	Apresenta um estudo focado em EKG (Enterprise Knowledge Graphs) utilizando de Linked Data para a integração de diferentes bases de conhecimento, utilizando de três abordagens (Indefinido, Transição e Federado).

Fonte: Autores, (2017)

Na próxima seção serão discutidos os artigos selecionados, apresentando os fatores identificados que evitam a difusão do LED.

4 Resultados e Discussões Preliminares

Com a análise realizada, foi possível identificar que todos os autores deixam claro que a web semântica e suas tecnologias podem beneficiar o meio empresarial, perceptivelmente através da interoperabilidade.

Com a realização desta análise também é possível perceber que o foco de discussão dos artigos analisados de dados conectados no âmbito empresarial, englobam áreas como: IoT, reconhecimento de entidades nomeadas, grafos de conhecimento, integração de dados entre outros, como pode ser observado no Quadro 1.

A literatura apresenta que o uso do LED está focado no consumo de dados, no entanto, o baixo índice de estudos relacionados ao tema geram uma certa "insegurança" por parte das empresas que desejam fazer uso deste conceito. Esse é um dos motivos que faz com que o LED não seja tão difundido (GRAUBE et al., 2013, LI; ZHAI, 2016, GLACHS, SCHAFFERT; BAUER, 2012). Porém, há esforços em termos de pesquisa para difundir o LED, mas é possível perceber outros fatores que influenciam negativamente o aproveitamento deste contexto.

Com base nos artigos analisados, destaca-se e discute-se os seguintes fatores identificados: (1) cultura organizacional, (2) alinhamento estratégico, (3) trabalhos apenas com foco em domínio público e no meio acadêmico e (4) confusão quanto aos conceitos de dados abertos e dados conectados.

- Cultura Organizacional: Percebe-se que de nada adianta um ou dois indivíduos quererem implantar dados conectados em uma organização, caso não se tem uma conscientização e o

interesse das pessoas envolvidas no processo de implantação de dados conectados (principalmente os gestores da empresa) e dos benefícios que poderão ser obtidos com o uso deste conceito.

Nota-se ainda que a maioria dos usuários corporativos não tem uma "mente semântica" (HU; SVENSSON, 2010), fazendo com que o medo da mudança crie barreiras para aderir ao conceito de LED. Também se evidencia que em sua maioria os gestores das empresas estão preocupados apenas com que os dados sejam fornecidos de forma oportuna e precisa, independente das tecnologias utilizadas.

- Alinhamento estratégico: Constata-se que o gerenciamento de dados possui por finalidade o melhoramento da eficiência do *Core Business* da empresa (HU; SVENSSON, 2010), para isso se faz necessário conhecer os processos da empresa para que haja um alinhamento entre a estratégia empresarial e o uso dos dados. Da mesma forma, que os links entre dados não devem ser aleatoriamente criados, independentes dos processos de negócios (HU; SVENSSON, 2010).

Verifica-se ainda que alinhar dados a estratégia da empresa, possibilita o relacionamento dos dados empresariais internos ou externos, proporcionando assim novas perspectivas para a tomada de decisões.

- Trabalhos apenas com foco em domínio público e no meio acadêmico: Percebe-se que há pesquisas em meio acadêmico, que muitas vezes não chegam a ser testadas no âmbito empresarial, ou seja, aparenta haver um *gap* entre o que a academia desenvolve e o que as empresas necessitam para fazer uso (GOMES, 2017). Também se percebe que as pesquisas aplicadas em sua maioria estão voltadas para o domínio público.
- Confusão quanto aos conceitos de dados abertos e dados conectados: Para se fazer uso de dados conectados é preciso ter claro o seu conceito. Onde percebe-se que devido ao movimento de Governo Aberto (MEIJER; CURTIN; HILLEBRANDT, 2012), o conceito de dados abertos ganhou grande ênfase. No entanto quando se menciona dados conectados percebe-se que há uma confusão quanto aos conceitos apresentados, pois dados abertos podem ser disponibilizados a todos sem que estejam em um formato RDF. Ao mesmo tempo, que os dados podem ser conectados sem estarem livremente disponíveis para serem utilizados ou distribuídos, este é o caso de dados conectados. Assim, como se tem dados abertos conectados, que necessitam ter uma licença aberta bem como fazer uso do formato RDF, conforme pode ser observado no Quadro 2.

Quadro 2 – Justaposição de dados abertos e fechados

Representação/Grau de abertura	Fechado	Aberto
Modelo de dados estruturados Ex.: XML, CSV, SQL, etc...	Dados	Dados Abertos
Modelo de Dados RDF	Dados Conectados	Dados Abertos Conectados

Fonte: AUER et al., (2014)

Analisando ainda os artigos, observa-se que são apresentadas possibilidades de pesquisas futuras, sendo que a maioria dos autores apresentam possibilidades de pesquisas focando na continuidade de seu próprio trabalho. Conforme segue no Quadro 3.

Quadro 3 – Relação de trabalhos futuros

Titulo	Ano	Trabalhos Futuros
A Case Study Of Linked Enterprise Data	2010	Realizar uma avaliação em grande escala com usuários convidados de diferentes departamentos e regiões

		geográfica. Identificar outros cenários de uso que possa demonstrar valor a nova diversidade de usuários.
Flexibility Vs. Security In Linked Enterprise Data Access Control Graphs	2013	Criar regras políticas de acesso a fim de minimizar as barreiras impostas pelas empresas para começar a obter informações de ambientes de dados conectados.
A Linked Data Perspective For Collaboration In Mashup Development	2013	Identificar novos padrões de colaboração adicionais usando de novas fontes externas, estender a abordagem para considerar a experiência externa dos desenvolvedores e realizar a avaliação com base no protótipo.
Consolidating Heterogeneous Enterprise Data For Named Entity Linking And Web Inteligence	2015	Melhorar a performance de desambiguação da Recognyze, considerando bases estruturadas mais complexas no processo de vinculação. Otimizar e avaliar perfis de desambiguação que funcionem com fontes de dados públicas, explorar as opções de combinação de dados empresariais conectados com dados abertos conectados, fornece avaliações a outros tipos de entidades e ampliar a abordagem para outros idiomas.
Linked Enterprise Data Model And Its Use In Real Time Analytics And Context-Driven Data Discovery	2015	Os autores não evidenciam as oportunidades de trabalhos futuros.
Enterprise Knowledge Graphs - A Backbone Of Linked Enterprise Data	2016	Melhorar o conceito de EKG em domínios técnicos e empresariais e especificar melhor as características do EKG, a fim de aprofundar o estudo de modo que as empresas possam fazer uso.

Fonte: Autores, (2017)

Verifica-se que para a difusão do LED, há a necessidade de trabalhos que abordem mais claramente os benefícios e os ganhos reais que são obtidos ao utilizar de dados conectados. Deste modo, constata-se também que o fator primordial para a difusão de dados conectados no âmbito empresarial são as pessoas (HU; SVENSSON, 2010), pois sem elas não há a compreensão dos conceitos, bem como o interesse em utiliza-lo.

5 Considerações Finais

Com a realização deste trabalho observou-se que dados conectados agregam um valor notável no domínio público (BIZER; HEATH; BERNERS-LEE, (2009). No entanto quando se fala do domínio empresarial nota-se que os seguintes fatores impedem sua difusão: cultura organizacional, alinhamento estratégico, trabalhos apenas com foco em domínio público e no meio acadêmico e uma confusão quanto aos conceitos de dados abertos e dados conectados.

Identificou-se que os primeiros trabalhos que abordam LED possuem foco em publicar dados, com o intuito de favorecer o compartilhamento e a interoperabilidade dos dados, em um espaço inter-organizacional, porém o foco dos trabalhos mais recentes está na interoperabilidade de sistemas, para o consumo de dados.

Notou-se também que o artigo mais citado encontrado na literatura apresenta uma abordagem mais abrangente sobre as possibilidades no uso do LED, pois ele engloba além questões técnicas, questões estratégicas, que são cruciais para as empresas.

Outros pontos que foram observados é que todos os trabalhos relatam a possibilidade de fazer uso de dados internos e externo em um mesmo ambiente e também há relatos de uma estrutura de nuvem de dados privada. No entanto não são relatadas propostas para materializar estas possibilidades e percebe-se que estas propostas são viáveis de serem desenvolvidas em pesquisas futuras.

Referências

ANTIDOT (França). **Enterprise Data Principles, uses and benefits.** 2012. Disponível em: <http://www.antidot.net/wp-content/uploads/2012/11/LinkedEnterpriseData-WP-en-v2.2.pdf>. Acesso em: 03 ago. 2017

AUER, Sören; LEHMANN, Jens; NGOMO, Axel-Cyrille Ngonga. Introduction to linked data and its lifecycle on the web. In: **Reasoning Web. Semantic Technologies for the Web of Data**. Springer Berlin Heidelberg, 2011. p. 1-75.

BERNEERS-LEE, Tim. The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. **Scientific American**, v. 284, n. 5, p. 34-43, 2001.

BIANCHINI, Devis; DE ANTONELLIS, Valeria; MELCHIORI, Michele. A linked data perspective for collaboration in mashup development. In: **Database and Expert Systems Applications (DEXA), 2013 24th International Workshop on**. IEEE, 2013. p. 128-132.

BIZER, Christian; HEATH, Tom; BERNERS-LEE, Tim. Linked data-the story so far. **Semantic services, interoperability and web applications: emerging concepts**, p. 205-227, 2009.

GALKIN, Mikhail; AUER, Sören; SCERRI, Simon. Enterprise Knowledge Graphs: A Backbone of Linked Enterprise Data. In: **Web Intelligence (WI), 2016 IEEE/WIC/ACM International Conference on**. IEEE, 2016. p. 497-502.

GLACHS, Dietmar; SCHAFFERT, Sebastian; BAUER, Christoph. Interlinking Media Archives with the Web of Data. In: **I-SEMANTICS (Posters & Demos)**. 2012. p. 17-21.

GOMES, Murilo Silveira. **PROPOSTA DE ARQUITETURA PARA ECOSSISTEMA DE INOVAÇÃO EM DADOS ABERTOS.**2017. 104 f. Dissertação (Mestrado) - Curso de Engenharia e Gestão do Conhecimento, Centro Tecnológico, Universidade Federal de Santa Catarina, Florianópolis, 2017. Disponível em: <http://btd.egc.ufsc.br/wp-content/uploads/2017/04/Murilo-Gomes.pdf>. Acesso em: 05 ago. 2017.

GRAUBE, Markus et al. Flexibility vs. security in linked enterprise data access control graphs. In: **Information Assurance and Security (IAS), 2013 9th International Conference on**. IEEE, 2013. p. 13-18.

HU, Bo; SVENSSON, Glenn. A case study of linked enterprise data. **The Semantic Web–ISWC 2010**, p. 129-144, 2010.

LI, Hongqin; ZHAI, Jun. Constructing Investment Open Data of Chinese Listed Companies Based on Linked Data. In: **Proceedings of the 17th International Digital Government Research Conference on Digital Government Research**. ACM, 2016. p. 475-480.

MEIJER, Albert J.; CURTIN, Deirdre; HILLEBRANDT, Maarten. Open government: connecting vision and voice. **International Review of Administrative Sciences**, v. 78, n. 1, p. 10-29, 2012.

TANEJA, Kunal et al. Linked enterprise data model and its use in real time analytics and context-driven data discovery. In: **Mobile Services (MS), 2015 IEEE International Conference on**. IEEE, 2015. p. 277-283.

WEICHSELBRAUN, Albert; STREIFF, Daniel; SCHARL, Arno. Consolidating heterogeneous enterprise data for named entity linking and Web Intelligence. **International Journal on Artificial Intelligence Tools**, v. 24, n. 2, p. 1540008, 2015.

UM ESTUDO DOS CICLOS DE VIDA DE DADOS ABERTOS CONECTADOS

A STUDY OF LINKED OPEN DATA LIFE CYCLES

Lidiane Visintin[(1)], Murilo Silveira Gomes[(1)], José Leomar Todesco[(2)]

(1) Programa de Pós Graduação em Engenharia e Gestão do Conhecimento, Universidade Federal de Santa Catarina, {lidiane.visintin, lilo.flp}@gmail.com.

(2) Programa de Pós Graduação em Engenharia e Gestão do Conhecimento, Universidade Federal de Santa Catarina, titetodesco@gmail.com.

Resumo: Houve um aumento significativo do volume de dados e informações produzidas e disponibilizadas na web, nos últimos anos. Como consequência alguns conceitos têm emergido, sendo um destes: Dados Abertos Conectados. Este conceito tem ganhado ênfase em pesquisas, devido aos benefícios que os dados podem oferecer, tanto para indivíduos que publicam, quanto para aqueles que consomem estes dados. Este artigo objetiva-se em apresentar e discutir dois trabalhos que abordam ciclos de vida de publicação de dados abertos conectados, assim como fornecer uma análise preliminar do que a literatura aborda sobre as práticas e tratamentos de dados, com o intuito de encontrar características nestas pesquisas, assim como identificar possibilidades para pesquisas futuras. Com está analise foi possível discutir e concluir que os ciclos de vida são descritos como um processo sequencial e unidimensional de fases, que um grupo muitas vezes não especificado realiza para fornecer dados a um público geral, sem uma preocupação em relação as fases de pós publicação, sendo que, ainda foi possível identificar que os ciclos de vida de dados abertos conectados analisados são complementares.
Palavras-chave: Dados; Ciclo de Vida; Dados Conectados Abertos

Abstract: There has been a significant increase in the volume of data and information produced and made available on the web in recent years. As a consequence, some concepts have emerged, being one of these is: Linked Open Data. This concept has gained an emphasis on research because of the benefits that data can offer both to individuals who publish and who consume this data. This article aims at analyzing and discussing two papers that address linked open data life cycles, in order to find characteristics in these surveys, as well as to identify possibilities for future research. With this analysis, it was possible to discuss and conclude that life cycles are described as a sequential and one-dimensional process of phases, which an often-unspecified group performs to provide data to a general audience without a concern regarding post-publication phases. And it was possible to identify that the linked open data life cycles analyzed are complementary.
Keywords: Data; Life cycle; Linked Open Data

1 Introdução

O volume de dados produzidos nos últimos anos, no mundo teve um aumento significativo. Estima-se que cerca de 90% dos dados presentes na Web foram criados em anos recentes sendo que estes não obtiveram um aumento apenas no volume, mas também no nível de detalhamento, tudo isso devido às novas tecnologias (DATA REVOLUTION GROUP, 2014).

As novas tecnologias são oriundas de movimentos como: IoT (BARNAGHI, 2012), Big Data (HITZLER; JANOWICZ, 2013), Web Semântica (BERNERS-LEE et al, 2001), Governo Aberto (MEIJER; CURTIN; HILLEBRANDT, 2012), entre outros. Estes movimentos geraram também necessidades, como o maior nível de detalhamento dos dados, este contexto alavancou a Web semântica e com isso surgiram conceitos como: dados conectados (Linked Data) e dados abertos conectados (Linked Open Data), fortalecendo e aprimorando diversas pesquisas.

O movimento de Governo Aberto possibilitou que alguns países adotassem a abertura de dados principalmente no setor público, em busca de transparência, *accountability*, participação social e colaboração (MEIJER; CURTIN; HILLEBRANDT, 2012). Através deste movimento e

influenciado pelo conceito de *Openess*, surgiu o conceito de dados abertos (Open Data) que por sua definição são "dados que qualquer um pode livremente acessá-los, utilizá-los, modificá-los e compartilhá-los para qualquer finalidade, estando sujeito a, no máximo, exigências que visem preservar sua proveniência e sua abertura" (OPEN KNOWLEDGE, 2017b), como consequência houve um crescimento considerável na quantidade de dados que são disponibilizados na Web.

Já se tem diversos *datasets* publicados na Web, em formato aberto (ABELE; MCCRAE, 2017). No entanto, o grande desafio de publicar estes dados está em como publicar, para que indivíduos possam fazer uso de dados confiáveis e facilmente ligar a outros dados. Para isso se faz necessário conhecer quais são as fases ou estágio da publicação de dados como um todo para descrever o desenvolvimento deste fenômeno, assim como para prever os próximos passo a serem realizados quando deseja-se publicar dados.

Observando o contexto evidenciado, o objetivo deste artigo é apresentar e discutir dois ciclos de vida de publicação de dados abertos conectados encontrados na literatura, assim como fornecer uma análise preliminar do que a literatura aborda sobre as práticas e tratamentos de dados.

Na sequência, são apresentadas as seções deste trabalho: na seção 2 aborda-se o contexto de dados abertos; na seção 3 é apresentado o contexto de dados abertos conectados; na seção 4 é apresentada uma breve descrição dos procedimentos metodológicos; na seção 5 apresenta-se uma breve discussão sobre dois dos ciclos de vida de dados abertos conectados, assim como uma análise preliminar do que já foi identificado na literatura até o momento; e por fim, na seção 6 são apresentados os encaminhamentos para trabalhos futuros e conclusões.

2 Dados Abertos

Um grupo de trabalho da Califórnia definiu oito princípios, para a abertura dos dados (DADOS.GOV 2017), são estes: completos, primários, atuais, acessíveis, processáveis por máquina, acesso não discriminatório, formato não proprietário e licença livre. Sendo que, compreender como os dados são licenciados e como os mesmos podem ser disponibilizados são fatores fundamentais para se publicar dados.

O Open Knowledge (2017a) apresenta três regras para aqueles que desejam abrir seus dados:

- Manter simples: Começar pequeno, simples e rápido, pois não há exigências quanto ao tamanho do conjunto de dados (*dataset*) a ser aberto.
- Envolver-se cedo e com frequência: Envolver-se com os usuários dos dados sejam estes cidadãos, empresas, desenvolvedores, entre outros.
- Abordar medos e mal-entendidos comuns: pois, ao abrir os dados sempre surgirão dúvidas e medos.

Estas três regras viabilizam avaliar, assim como apreender com maior agilidade, com base na experiência. Para isso, ainda são apresentadas quatro etapas chave para a abertura dos dados, sendo que estas podem ser realizadas simultaneamente muitas vezes (OPEN KNOWLEDGE, 2017a):

Etapa 1: Escolha do conjunto de dados ou quais partes deste conjunto de dados pretende-se disponibilizar.

Etapa 2: Aplicar uma licença aberta especificando os direitos de propriedade intelectual existem sobre os dados.

Etapa 3: Disponibilizar os dados brutos e em formato que possibilite o processamento por máquina.

Etapa 4: Torna os dados visíveis, disponibilizando-os na Web.

No entanto, quando busca-se inserir dados na web de modo que máquinas ou pessoas possam explorar estes dados através das ligações entre os mesmos, deve-se seguir os princípios de dados conectados (BERNERS-LEE, 2006). Sendo que, dados conectados podem ser liberados sobre uma licença aberta (dados conectados abertos), fomentando assim a sua livre reutilização.

3 Dados Abertos Conectados

O conceito de dados conectados foi estendido, com o objetivo de construir a Web de Dados (HEATH; BIZER, 2011). A Web de dados é constituída de *datasets* com licença aberta, que fazem uso do formato RDF seguindo as recomendações de dados conectados, dando origem a dados conectados abertos (BIZER; HEATH; BERNERS-LEE, 2009).

Dados Conectados e Dados Conectados Abertos fazem uso de tecnologias presentes na arquitetura da Web Semântica (RDF e URIs), com o intuito de padronizar os dados, possibilitando assim o processamento por máquina, e a conexão dos dados e dos *datasets*. Deste modo, os *datasets* em RDF podem ser associados a outros conjuntos de dados para formar uma grande base de dados conectados, ou seja, materializando assim a Web de Dados (ALSHEHHI et al. 2013). Desde o início do projeto dados abertos conectados nota-se um aumento significativo no número de *datasets* que são disponibilizados, bem como um aumento no número de indivíduos que publicam seus dados (ABELE; MCCRAE, 2017).

A motivação por trás de dados abertos conectados é fornecer dados brutos, estruturados e ligados através da Web que possam ser acessados universalmente e ser facilmente compartilhados (BIZER; HEATH; BERNERS-LEE, 2009). Porém, o que é encontrado na literatura para guiar e auxiliar os indivíduos que desejam abrir seus dados, seja isto para dados abertos, dados conectados ou ainda para dados abertos conectados.

4 Procedimentos Metodológicos

No estudo de Broek et al. (2014) os autores apresentam os ciclos de vida existentes, para a publicação de dados. Desta forma, foi realizada uma busca por estes trabalhos na literatura, sendo que os trabalhos que foram encontrados são apresentados no Quadro 1, incluindo o trabalho de Broek et al. (2014).

Quadro 1: Trabalhos que abordam ciclos de vida de publicação de dados.

Autor / Ano	Título
Hyland; Wood (2011)	The joy of data-a cookbook for publishing linked government data on the web.
Villazón Terrazas et al. (2011)	Methodological guidelines for publishing government linked data
Hausenblas (2011)	Linked data lifecycles.
Auer et al. (2011)	Introduction to linked data and its lifecycle on the web.
Janssen; Zuiderwijk (2012)	Open data and transformational government.
Scharffe et al. (2012)	Enabling linked-data publication with the datalift platform.
Broek et al.(2014)	Walking the extra byte: A lifecycle model for linked open data.

Fonte: Autores, (2017)

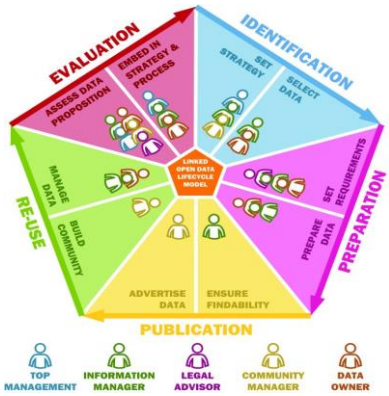
Com base nos artigos apresentados no Quadro 1, dois destes ciclos de vida foram selecionados para serem discutidos por este artigo, sendo estes: o ciclo de vida proposto por Auer et al. (2011) e o ciclo de vida proposto por Broek et al. (2014), devido a apresentarem focos distintos quanto a publicação de dados abertos conectados, sendo que os demais ciclos não abordam claramente o contexto de dados abertos conectados.

5 Resultados e Discussões Parciais

Esta seção analisa e discute dois ciclos de vida, que conceituam práticas e tratamentos para a publicação de dados abertos conectados.

A Quadro 2 apresenta os autores, o objetivo proposto por cada trabalho, bem como seu foco e o diagrama dos ciclos de vida analisados.

Quadro 2 – Ciclos de Vida de publicação de dados abertos conectados

Autor/Ano	Objetivo do trabalho	Foco	Diagrama
Auer et al. (2011)	O objetivo é facilitar a distribuição e instalação de ferramentas e componentes de software que suportem o ciclo de vida de publicação de dados conectados.	Tecnológico	
Broek et al. (2014)	Com base na literatura apresentam as etapas de um ciclo de vida onde busca-se considerar a visão estratégica de publicação de dados conectados.	Nas orientações básicas para a publicação de dados a nível estratégico.	

Fonte: Autores, (2017)

O primeiro ciclo analisado, foi proposto por Auer et al. (2011), o objetivo do projeto que construiu este ciclo é desenvolver e elencar ferramentas presentes na literatura, para apoiar a criação de dados conectados, sendo que o ciclo é composto por oito estágios, sendo estes: Armazenamento, Autoria, Fusão, Enriquecimento, Qualidade, Evolução, Exploração e Extração; Os autores deixam claro que estas etapas não devem ser tratadas separadamente, porém salientam que devem ser investigados métodos que facilitem benefícios mútuos para os estágio desenvolvidos, com o intuito de resolver os desafios que são apresentados.

Já o ciclo de vida proposto por Broek et al. (2014), foi desenvolvido baseado nos ciclos de vida encontrados pelos autores na literatura, sendo que este ciclo foi aperfeiçoado com base nas lições aprendidas em um estudo de caso desenvolvido em uma empresa semi-pública na Holanda. O ciclo de vida apresenta cinco fases, sendo estas: (1) identificação, (2) preparação, (3) publicação, (4) reuso, (5) avaliação.

Observa-se nos demais trabalhos presentes na literatura, assim como no ciclo de vida proposto por Auer et al. (2011), que os mesmos levam em conta exclusivamente os processos operacionais de publicação de dados abertos conectados (como por exemplo, extração, limpeza, publicação e manutenção de dados), ignorando os processos estratégicos (como tomada de decisão, alinhamento com a alta gerência e fatores legais). Assim, as decisões sobre quais dados serão publicados, quem irá extrair os dados, como os dados são editados, como os dados podem ser acessados, quais licenças estão disponíveis e questões como: privacidade e responsabilidade não são tratadas na maioria dos trabalhos encontrados na literatura. Estes processos estratégicos mais gerais, são abordados apenas pelo ciclo

de vida proposto por Broek et al. (2014), sendo que os autores também definem *stakeholders* que auxiliam no processo de publicação de dados abertos conectados.

O trabalho de Broek et al. (2014), apresenta orientações a nível estratégico. No entanto, observa-se que os autores não explicitam quais critérios que foram utilizados no agrupamento das fases genéricas identificadas na literatura para o desenvolvimento das fases do ciclo de vida proposto. Outro ponto observado no trabalho é que não está claro como foram identificados os cinco *stakeholders* apresentados, sendo que ao longo do texto observa-se um sexto *stakeholder* o gerente de projeto, que não é identificado junto ao ciclo de vida.

Observa-se ainda no mesmo trabalho que o ciclo proposto possui um foco diferenciado. Sendo que, através das fases de reuso e avaliação tem-se a possibilidade de obter feedbacks, possibilitando melhorar os processos estratégicos das organizações, referente aos dados que são disponibilizados.

Nota-se através desta análise que os dois ciclos de vida, são complementares, ou seja, o ciclo proposto por Auer et al. (2011) fornece ferramentas, focado em como tratar os dados, com o intuído de minimizar o árduo trabalho de publicar dados abertos conectados e o ciclo proposto por Broek et al. (2014) apresenta as orientações a nível estratégico, assim como os *stakeholder* envolvidos em todo o processo.

Realizando uma análise preliminar dos demais trabalhos encontrados na literatura, percebe-se que em sua maioria os ciclos de vida de publicação de dados são descritos como um processo sequencial e unidimensional de fases, que um grupo muitas vezes não especificado de *stakeholders* realiza frequentemente para fornecer uma quantidade de dados à um público geral. Percebe-se também que estes trabalhos em sua maioria focam nos processos operacionais, ignorando muitas vezes a mensuração do consumo dos dados e o feedback dos usuários, sendo que isto possibilitaria não somente melhorias quanto a qualidade dos dados disponibilizados, mas também poderia potencializar novos negócios. Visto que, a qualidade, a disponibilidade e a utilização dos dados têm um papel central no sucesso daqueles que publicam seus dados (HAIDER; HAIDER, 2013).

Outro item observado é que os ciclos de vida para a publicação de dados conectados e dados conectados abertos engloba a maioria dos trabalhos citados por este artigo. No entanto só um dos trabalhos abrange o contexto somente de dados abertos, sendo que não é abordado como um ciclo de vida, mas sim como um processo (JANSSEN; ZUIDERWIJK, 2012). Também observa-se que pouco se fala em licença aberta quando abordado o contexto de dados abertos e dados abertos conectados.

6 Conclusão ou Considerações Finais

Este artigo apresentou resultados parciais de um estudo, sobre os trabalhos encontrados na literatura que discutem o tratamento dos dados para publicação. O mesmo foi desenvolvido com o propósito de encontrar características referente as pesquisas da área, assim como fornecer possibilidade de futuras pesquisas.

Com a realização deste estudo observou-se que a maioria dos ciclos de vida encontrados na literatura abordam apenas processos operacionais, ou seja, seu foco é a nível tecnológico, e que em sua maioria ignoram as fases de pós-publicação, sendo que a mensuração dos dados utilizados, assim como os *feedbacks* fornecidos podem gerar benefícios. Com base neste itens elencado, também percebe-se que há uma falta de alinhamento estratégico do "Por que e para quem estamos publicando dados?", o que leva a pensar que as organizações apenas publicam dados por publicar, sem pensar nos benefícios que podem ser obtidos.

Conclui-se também que não há estudos que guiem ou auxiliem tanto em nível estratégico, quanto em nível tecnológico, aqueles que desejam começar a abrir seus dados, sem pensar

em dados conectados. Para isso sugere-se como possibilidade de pesquisas futuras a estruturação de um ciclo de vida para a publicação de dados abertos, com foco tecnológico e estratégico, baseado nas recomendações do Open Knowledge e dos ciclos de vida de publicação de dados.

Referências

ABELE, A., MCCRAE, J. Linking open data cloud diagram. **LOD Community**. 2017. Disponível em: <http://lod-cloud.net/. Acesso em: 04 ago. 2017.

ALSHEHHI, Maryam et al. Visual analytics in the web of data. In: **Electronics, Circuits, and Systems (ICECS), 2013 IEEE 20th International Conference on**. IEEE, 2013. p. 102-103.

AUER, Sören; LEHMANN, Jens; NGOMO, Axel-Cyrille Ngonga. Introduction to linked data and its lifecycle on the web. In: **Reasoning Web. Semantic Technologies for the Web of Data**. Springer Berlin Heidelberg, 2011. p. 1-75.

BARNAGHI, Payam et al. Semantics for the Internet of Things: early progress and back to the future. **International Journal on Semantic Web and Information Systems (IJSWIS)**, v. 8, n. 1, p. 1-21, 2012.

BERNEERS-LEE, Tim. The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. Scientific American, v.284, n. 5, p. 34-43, 2001.

BERNERS-LEE, Tim. Linked data-design issues. 2006. Disponível em: *<http://www.w3.org/DesignIssues/LinkedData.html>.* Acesso em: 04 ago. 2017.

BIZER, C.; HEATH, T.; BERNERS-LEE, T. Linked data-the story so far. Semantic Services, Interoperability and Web Applications: Emerging Concepts. **Information Science Reference (an imprint of IGI Global), USA**, p. 205-227, 2009.

BROEK, Tijs Adriaan Van Den; VAN VEENSTRA, A. F. E.; FOLMER, Erwin Johan Albert. Walking the extra byte: A lifecycle model for linked open data. 2013.

DADOS.GOV. Portal Brasileiro de Dados Abertos. **O que são dados abertos**. 2017. Disponível em: http://dados.gov.br/dados-abertos/ Acesso em: 28 ago. 2017.

DATA REVOLUTION GROUP. **A World That Counts:** Mobilising the Data Revolution for Sustainable Development. 2014. Disponível em: <http://www.undatarevolution.org/wp-content/uploads/2014/11/A-World-ThatCounts.pdf>. Acesso em: 01 ago. 2017.

AIDER, Waqar; HAIDER, Abrar. Governance structures for engineering and infrastructure asset management. In: **Technology Management in the IT-Driven Services (PICMET), 2013 Proceedings of PICMET'13:**. IEEE, 2013. p. 1229-1238.

HAUSENBLAS, M. Linked data lifecycles, presentation from DERI research institute, Galway, Ireland, July 2011.

HEATH, Tom; BIZER, Christian. 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.

HITZLER, Pascal; JANOWICZ, Krzysztof. Linked Data, Big Data, and the 4th Paradigm. **Semantic Web**, v. 4, n. 3, p. 233-235, 2013.

HYLAND, Bernadette; WOOD, David. The joy of data-a cookbook for publishing linked government data on the web. **Linking government data**, p. 3-26, 2011.

JANSSEN, M. and ZUIDERWIJK, A. Open data and transformational government. In **Transforming Government Workshop**. Brunel University, United Kingdom. 2012.

MEIJER, Albert J.; CURTIN, Deirdre; HILLEBRANDT, Maarten. Open government: connecting vision and voice. **International Review of Administrative Sciences**, v. 78, n. 1, p. 10-29, 2012.

OPEN KNOWLEDGE. How to Open Data. Disponível em: < https://okfn.org/opendata/how-to-open-data/>. Acesso em: 29 jul. 2017a.

OPEN KNOWLEDGE. **Open Definition.** Disponível em: <http://opendefinition.org/od/2.1/en/>. Acesso em: 02 ago. 2017b.

SCHARFFE, François et al. Enabling linked-data publication with the datalift platform. In: **Proc. AAAI workshop on semantic cities**. 2012.

VILLAZÓN-TERRAZAS, Boris et al. Methodological guidelines for publishing government linked data. Linking government data, p. 27-49, 2011.

CONSTRUÇÃO COLABORATIVA DE REPRESENTAÇÕES PARA A DISSEMINAÇÃO DE DADOS AGRÍCOLAS: Um estudo no Portal CoDAF

Collaborative construction of representations for agricultural data dissemination: a study in the CoDAF Portal.

Fábio Mosso Moreira[1], Diana Vilas Boas Souto Aleixo [2], Pedro Henrique Santos Bisi[3], Leonardo F. Franchi[4], Ricardo César Gonçalves Sant'Ana[5]

(1) UNESP, Marília, fabiomoreira@tupa.unesp.br
(2) UNESP, Marília, dianavbsouto@gmail.com
(3) UNESP, Marília, phbisi@gmail.com
(4) UNESP, Tupã, ffranchileonardo@gmail.com
(5) UNESP, Marília, ricardosantana@marilia.unesp.br

Resumo:
O ingresso das novas tecnologias informacionais na sociedade viabilizou o compartilhamento de dados e informações de forma dialógica, podendo ser considerado uma alternativa para aproximar os produtores às informações que os auxilia no desenvolvimento de suas atividades. Neste sentido, o objetivo deste trabalho é descrever o processo de construção colaborativa de representações de dados agrícolas, por meio de um estudo realizado no Portal Competências Digitais para Agricultura Familiar (Portal CoDAF), de modo a identificar os atributos que compõem a estrutura das representações e apontar os principais agentes e atividades envolvidos no processo. A metodologia baseou-se no Ciclo de Vida dos Dados como modelo para estruturação do fluxo informacional estudado. Como resultado parcial, pôde-se identificar um conjunto de atributos para compor a estrutura de disponibilização dos dados. Conclui-se que a participação dos agentes internos e externos de forma colaborativa é essencial para possibilitar a recuperação dos dados, assim como a definição de uma estrutura padrão para representação das fontes.
Palavras-chave: Fonte de dados. Acesso a dados. Ciclo de Vida dos Dados. Portal CoDAF. Dados agrícolas.

Abstract:
The admission new informational technologies in society, made possible data and information dialogical sharing, and may be considered an alternative for helping to bring producers to informations that helps you development of their activities. In this sense,the objective of the present work is describe the collaborative construction process of agricultural data representations, by study realized in Digital Skills for Family Farming Portal (CoDAF Portal) to identify the attributes which compose the representations structure and to indicate the key actors and activities involved in process. The methodology was based on the Data Life Cycle as a model for structuring of informational flow. As a partial result, it could be identified an set of attributes to provide data availability structure. The conclusion is that the participation of internal and external agents in a collaborative way is essential to allow for retrieving and recovery of the data, as well the definition a standard framework for representation of sources.
Keywords: Data sources. Access to data. Data Life Cycle. CoDAF Portal. Agricultural Data.

1 Introdução

As novas Tecnologias da Informação e Comunicação (TIC), em especial a Internet e a Web, podem ser consideradas importantes alternativas para viabilizar o compartilhamento de dados e informações dialogicamente, reduzindo a unidirecionalidade emissor-receptor, como, por exemplo, quando se observa na comunicação por meio de mídias tradicionais (ex: TV, rádio e jornais).

Um dos setores que pode se beneficiar com o uso destas tecnologias é a Agricultura. O seguimento é composto por pequenos, médios e grandes produtores rurais, que juntos tem importante participação na geração de capital para o país, representando 23,5% do PIB Nacional (CENTRO DE ESTUDOS AVANÇADOS EM ECONOMIA APLICADA, 2016).

Vieiro e Silveira (2011) ressaltam que as possibilidades proporcionadas pelas TIC para o meio rural são diversas, com destaque ao seu papel na ampliação de horizontes e incorporação de novas expectativas; constituição de grupos de comercialização;

novas políticas públicas; estimativas de safras e desempenhos nas bolsas de valores e *commodities*; serviços bancários; cooperativas de crédito e de produção; educação à distância e assistência técnica. Neste sentido, o meio rural não pode mais ser visto como um local distante e atrasado pelos que vivem no meio urbano e industrial, mas sim como um ícone de diversidade que está em constante desenvolvimento, demandando cada vez mais informações atualizadas e constantes.

Segundo Moreira et al. (2015), os dados e informações provenientes da agricultura estão distribuídos entre diversos agentes e instituições, como exemplo, produtores, universidades, institutos de pesquisa, serviços de extensão rural, iniciativa pública e privada, e organizações não governamentais (ONGs). Muitos destes conteúdos são disponibilizados em formato digital e podem ser acessadas via Internet e da Web, contudo, podem existir problemas na recuperação, tais como a heterogeneidade na definição dos domínios e a ambiguidade léxica.

Além dos problemas intrínsecos ao contexto da agricultura, Akerlof (1970) ressalta que a obtenção de conteúdos informacionais pode ser prejudicada quando a disponibilização das bases de dados não ocorre por meio de uma linguagem de fácil compreensão pelo usuário. Este fato se agrava quando a disponibilização é de responsabilidade de instituições públicas, as quais podem vir a fornecer serviços com acesso parcial, superficial e de difícil compreensão dos dados (LOPES; SANT'ANA, 2013), além de recursos com características heterogêneas e conjuntos de dados pulverizados em diferentes ambientes informacionais (SANT'ANA; RODRIGUES, 2013).

Quando se trata de um processo de compartilhamento de dados, deve-se considerar que o objeto em questão possui algumas particularidades, como o fato de um dado, por si só, não transmitir uma mensagem ou representar algum conhecimento (MOREIRA; SANT'ANA, JORENTE, 2016). O dado, como elemento básico, é formado por signos e não contém intrinsicamente, um componente semântico, mas somente elementos sintáticos (SANTOS; SANT'ANA, 2002).

As características relacionadas à baixa carga semântica e à alta estruturação inerentes aos dados tornam o processo de recuperação diferente dos processos de recuperação da informação via mecanismos de buscas (JANOWICZ et al., 2012). Segundo Van Rijsbergen (1979), no processo de obtenção de dados busca-se por uma correspondência exata à necessidade, enquanto na recuperação da informação espera-se uma correspondência aproximada (a melhor correspondência possível); na recuperação de dados a inferência utilizada é a dedução lógica, ao passo que na recuperação da informação é possível realizar uma inferência indutiva (possui graus de certeza ou incerteza); na recuperação de dados a linguagem utilizada é uma linguagem artificial restrita por sintaxes específicas, enquanto na recuperação da informação utiliza-se de uma linguagem natural que pode ser expressa apenas com as especificações necessárias; ainda sobre a linguagem de consulta utilizada, na recuperação de dados a linguagem, por ser amarrada a uma sintaxe, possui maior sensibilidade ao erro, enquanto na recuperação da informação, pequenos erros podem ser contornados.

Uma possível solução ao problema da recuperação de dados agrícolas seria a construção de representações, a fim de facilitar a localização das bases e uso dos dados. Entretanto, devido à sua baixa carga semântica (SANTOS; SANT'ANA, 2013), os dados requerem esforços adicionais ainda maiores para sua representação, já que não bastam elementos que os descrevam como um todo e que propiciem sua recuperação.

Para Sant'Ana (2017), são necessários, ainda, elementos que permitam a sua interpretação por quem os acesse, com informações que detalhem sua estrutura e possibilitem a interpretação de cada atributo que os compõe. É importante considerar neste processo, aspectos como o formato das representações, o momento em que os dados devem receber tratamento e o responsável pela representação.

Estas questões indicam a necessidade da elaboração de estudos orientados a

buscar alternativas que venham a amenizar os problemas encontrados na recuperação de dados agrícolas, tais como a criação de camadas de representação de forma colaborativa. Ressalta-se que estas iniciativas podem ser realizadas tendo como base teorias e princípios da Ciência da Informação, que por sua vez deve ultrapassar a fronteira da informação interpretada para atingir o nível rígido dos dados (MOREIRA; VALENTIM; SANT'ANA, 2016).

A construção de representações para dados agrícolas pode ser realizada sob a responsabilidade do detentor, ou mesmo com a participação de outrem que venha a realizá-la de forma colaborativa. Para tanto, adota-se como objeto desta pesquisa um ambiente informacional digital especializado na disponibilização de conteúdos agrícolas e que disponibiliza representações para viabilizar a disseminação dos dados.

O Portal CoDAF[1] (website desenvolvido pelo Projeto Competências Digitais para Agricultura Familiar, da Faculdade de Ciências e Engenharia UNESP/Tupã) é um ambiente que proporciona áreas para que pequenos produtores divulguem conteúdos sobre sua produção de maneira on-line, estes podem qualificar seus produtos, fornecer características sobre seu empreendimento rural e indicar locais onde ocorre a comercialização de seus produtos. No Portal também são disponibilizados conteúdos como notícias sobre agricultura e tecnologia, e obtidos conjunto de dados agrícolas, sendo o último foco desta pesquisa.

2 Objetivos

Esta pesquisa tem como objetivo descrever o processo de construção colaborativa de representações de dados agrícolas por meio de um estudo realizado no Portal CoDAF, identificando a estrutura das representações e apontando os principais agentes e atividades envolvidas neste processo.

3 Procedimentos Metodológicos

Para atender ao objetivo proposto na pesquisa, o procedimento metodológico utilizado baseou-se em um modelo para estruturar fluxos informacionais que envolvem processos de compartilhamento de dados. Segundo Sant'Ana (2016), o Ciclo de Vida dos Dados (Figura 1) pode ser utilizado para estudar fatores e características que propiciem ampliação do equilíbrio entre os atores envolvidos no processo e a máxima otimização do uso dos dados. O modelo se baseia em uma estrutura básica para contextualizar momentos, características e requisitos em um aspecto cíclico de fluxo de dados.

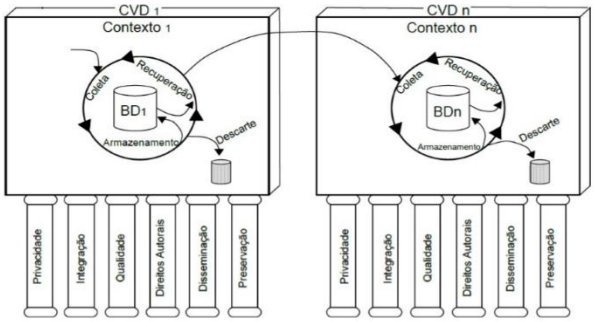
A estrutura do Ciclo de Vida dos Dados é composta pelas seguintes fases (SANT'ANA, 2016):

- COLETA: identificação das necessidades informacionais que irão nortear as escolhas dos dados necessários e a articulação de estratégias para localizar e avaliar estes dados, definindo as ferramentas necessárias para coleta.
- ARMAZENAMENTO: definição de quais dados serão disponibilizados e quais serão apenas armazenados, e qual estrutura física e lógica será utilizada para o armazenamento.
- RECUPERAÇÃO: formulação de estratégias para viabilizar que estes dados sejam encontrados, acessados e passíveis de interpretação (preferencialmente, e em muitos casos obrigatoriamente, por máquinas).
- DESCARTE: consiste na limpeza ou desativação da base, ou apenas a atividade de apagar o registro inteiro ou algum atributo específico.

Para Sant'Ana (2016) cada uma das fases são permeadas por seis objetivos específicos envolvidos no fluxo informacional: Privacidade, Integração, Qualidade, Direitos Autorais, Disseminação e Preservação. Esta pesquisa se insere principalmente no objetivo da Disseminação, uma vez que a construção de representações de dados pode aperfeiçoar seu processo de recuperação.

[1] Disponível em: <codaf.tupa.unesp.br>. Acesso em: 11/07/2017.

Figura 1: Ciclo de Vida dos Dados para Ciência da Informação.



Fonte: Sant'Ana (2016).

Após delimitado o modelo no qual a descrição do fluxo informacional será embasada, foram coletadas informações no Portal CoDAF por meio da observação da área contendo as representações de dados divulgados pelo portal. Ao total foram analisadas oito fontes[2] de dados agrícolas que estavam acessíveis até a data em que ocorreram as coletas (Julho/2017): "Dados sobre produção, importação e exportação de produtos agrícolas"; "Centro produtores de leite no estado de São Paulo"; "IBGE como fonte de Dados para a Agricultura"; "Dados sobre o repasse de recursos financeiros do Programa Nacional de Fortalecimento da Agricultura Familiar (PRONAF)"; "Dados sobre preço recebido pelos agricultores (Laranja)"; "Base Cartográfica Contínua do Brasil - Escala 1:1.000.000"; "Fonte de Dados sobre preço médio de Hortifrutícolas"; e "Dados sobre recursos do MDS/MDA aplicados no Programa de Aquisição de Alimentos".

A partir do conteúdo observado, foram definidos os principais atributos que compõem a estrutura das representações. Para apresentar as informações analisadas, elaborou-se um fluxograma que ilustra o processo de construção colaborativa de representações de dados agrícolas no Portal CoDAF.

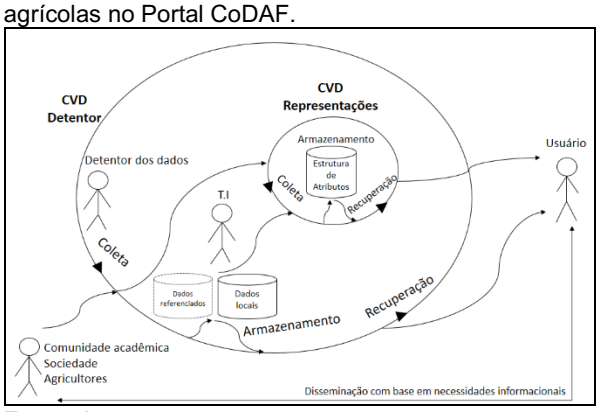
4 Resultados

O conteúdo disponibilizado no Portal CoDAF é gerado tanto pelos integrantes do projeto quanto pela participação de agentes externos, como no cadastro de propriedades rurais, no envio das notícias para publicação, e na construção das representações de dados agrícolas, que buscam viabilizar a disseminação de fontes e utilização dos dados. As atividades envolvidas na fase da coleta deste fluxo informacional baseiam-se na premissa da colaboração, e são realizadas por meio de atividades junto à comunidade acadêmica, à sociedade e aos produtores, por meio das funcionalidades do Portal ou através da realização de cursos e oficinas.

O conteúdo gerado na coleta é armazenado localmente por um banco de dados mantido pelo CoDAF. Também existem referências a conteúdos de fontes externas por meio de *hiperlinks,* apontados nas notícias publicadas e nas representações de dados disponíveis, configurando uma base referenciada. Na fase de armazenamento é importante destacar o papel do suporte tecnológico utilizado, garantido pelos recursos informáticos e pela mediação dos integrantes responsáveis pela gestão das ferramentas tecnológicas (gerenciador de conteúdo, manutenção do servidor).

Após a coleta e armazenamento, o conteúdo fica disponível no Portal CoDAF para ser recuperado pelo usuário. Como citado anteriormente, encontram-se problemas no processo de recuperação de dados que podem ser amenizados com a construção de representações. Este processo pode ser considerado um ciclo de dados interno ao ciclo de dados principal do detentor (Figura 2).

Figura 2: Descrição do processo de disponibilização de conteúdo e construção das representações de dados agrícolas no Portal CoDAF.



Fonte: Autores.

[2] Disponível em:<http://codaf.tupa.unesp.br/agricultura-familiar/fontes-de-dados>. Acesso em:11/07/2017.

Após analisar o conteúdo das oito representações de dados disponíveis no Portal CoDAF, identificaram-se os seguintes atributos que podem definir sua estrutura:

1. Título: denomina o conteúdo dos dados presentes na representação.
2. Introdução: resumo que sintetiza o contexto do conteúdo dos dados representados (ex: o que, como, onde, por quê).
3. Informações sobre a origem dos dados:
    a. Nome e descrição da instituição mantenedora: instituição detentora dos dados presentes na representação, tais como: área de atuação, empresa privada ou pública, quais seus objetivos, que informações estão disponibilizadas em seu portal;
    b. Título da fonte originária: Rótulo de cabeçalho da fonte originária;
    c. Endereço eletrônico: *hiperlink* de acesso para a fonte originária;
    d. Descrição do processo de navegação: trilha hierárquica para chegar até a fonte originária a partir da página inicial da instituição detentora dos dados.
4. Informação sobre os dados presentes no(s) recurso(s) informacional(ais) disponíveis na fonte originária:
    a. Formato(s) do(s) arquivo(s) acessíveis na fonte originária: (ex: ".xls", ".odt", ".html", ".csv");
    b. Atributos dos dados: descrição dos atributos (colunas), o que representam (registros), qual o tipo de variável que recebe (ex: data, texto, número, valor).
5. Demonstração de uma possível visualização para os dados presentes nos recursos utilizados.
    a. Exibição de gráficos ou elemento de visualização construídos a partir do uso dos dados acessíveis no recurso;
    b. Descrição das informações que podem ser obtidas através da interpretação das visualizações geradas e uso dos dados presentes em algum recurso.

5 Considerações parciais

A pesquisa baseou-se no Ciclo de Vida dos Dados para analisar a construção colaborativa de representações de dados agrícolas por meio de um estudo realizado no Portal CoDAF. Com a descrição do processo foi possível observar os principais agentes envolvidos neste ciclo, em quais fases atuam e as atividades relacionadas a cada um.

Com a definição de uma estrutura padrão, obtida a partir da análise das representações disponíveis no Portal CoDAF, as atividades de construção colaborativa podem ser ampliadas por meio do fornecimento de uma aplicação para coleta online, a fim de aumentar a participação de agentes externos e consequentemente viabilizar a disseminação do conteúdo.

A participação de agentes externos é fundamental uma vez que este público pode conhecer melhor as necessidades informacionais e competências de quem busca os dados, como ilustrado na Figura 2.

Considera-se que as ações descritas e a estrutura definida para construção das representações podem ser implementadas por outras iniciativas que busquem amenizar os problemas relacionados com a recuperação e disseminação de dados no contexto da agricultura.

Referências

AKERLOF, G A. The Market for "Lemons": qualitative uncertainty and the market mechanism. **The quarterly journals of economics**, v.84, n.3, p. 488-500, 1970. Disponível em:<https://www.iei.liu.se/nek/730g83/artiklar/1.328833/AkerlofMarketforLemons.pdf>. Acesso em: 29/08/2017.

CENTRO DE ESTUDOS AVANÇADOS EM ECONOMIA APLICADA [CEPEA-USP]. **PIB do Agronegócio BRASIL**. ESALQ/USP, 2016. Disponível em:<http://www.cepea.esalq.usp.br/upload/kc

editor/files/Relatorio PIBAGRO Brasil_DEZEMBRO.pdf>. Acesso em: 14/04/2017.
JANOWICZ, K.; SCHEIDER, S.; PEHLE, T.; HART, G. Geospatial semantics and linked spatiotemporal data: past, present, and future. **Semantic Web**, v. 3, n. 4, p. 321-332, 2012. Disponível em:<http://content.iospress.com/download/semantic-web/sw077?id=semantic-web%2Fsw077>. Acesso em: 28/08/2017.
LOPES, R. de C. C.; SANT'ANA, R. C. G. Percepção dos usuários sobre o processo de acesso a dados sobre saúde em sítios do governo federal. In: XIV - ENANCIB, 2013, Florianópolis. **Anais**..., 2013. Disponível em:<http://enancib.ibict.br/index.php/enancib/xivenancib/paper/view/4370/3493>. Acesso em: 29/08/2017.
MOREIRA, F. M.; SANT'ANA, R. C. G.; SANTAREM SEGUNDO, J. E.; VIDOTTI; S. A. B. G. Tecnologias da Web Semântica para a recuperação de dados agrícolas: um estudo sobre o International Information System of the Agricultural Science and Technology (AGRIS). **Em Questão**, v. 21, p. 173-192, 2015. Disponível em:<http://seer.ufrgs.br/index.php/EmQuestao/article/view/50317>. Acesso em: 29/08/2017.
MOREIRA, F. M.; SANTANA, R. C. G.; JORENTE, M. J. V. A Complexidade na disponibilização e acesso a dados governamentais na Web. **Perspectivas em Ciência da Informação**, v. 21, p. 70-88, 2016. Disponível em:<http://portaldeperiodicos.eci.ufmg.br/index.php/pci/article/view/2540>. Acesso em: 29/08/2017.
MOREIRA, F. M.; VALENTIM, M. L. P.; SANT'ANA, R. C. G. Interdisciplinaridades em Ciência da Informação: um estudo do compartilhamento de dados governamentais na web. In. III Encontro Internacional Dados, Informação e Tecnologia, 3, Marília, 2016. **Anais**... do III Encontro Internacional Dados, Informação e Tecnologia, 2016. Disponível em:<http://gpnti.marilia.unesp.br:8085/index.php/3DTI/3dti/paper/view/340/150>. Acesso em: 29/08/2017.
SANT'ANA, R. C. G. Ciclo de Vida dos Dados: Uma perspectiva a partir da Ciência da Informação. **Informação e Informação**, Londrina, v.21, n.2, p.116-142, 2016. Disponível em:<http://www.uel.br/revistas/uel/index.php/informacao/article/view/27940>. Acesso em: 29/08/2017.
SANT'ANA, R. C. G. Reflexões sobre a Representação no Ciclo de Vida dos Dados. In: I Encontro de Representação Documental (EnReDo), 1, São Carlos, 2017. **Anais**..., 2017. Disponível em:<http://www.telescopium.ufscar.br/index.php/enredo/enredo/paper/viewFile/109/96>. Acesso em: 29/08/2017.
SANTANA, R. C. G.; RODRIGUES, F. A. Visualização de afinidades entre parlamentares mediante dados de votações no Senado Brasileiro. **Informação & Sociedade: Estudos**, João Pessoa, v.23, n.1, p.49-59, 2013. Disponível em:<https://repositorio.unesp.br/handle/11449/75885>. Acesso em: 29/08/2017.
SANTOS, P. L. A. C.; SANT'ANA, R. C. G. Transferência da informação: análise para valoração de unidades de conhecimento. **DataGramaZero**, Rio de Janeiro, v.3, n.2, 2002. Disponível em:<http://basessibi.c3sl.ufpr.br/brapci/index.php/article/view/0000001259/f8d0251a874410d87c7a0bcb589fc725>. Acesso em:29/08/2017.
SANTOS, P. L. V. A. C.; SANT'ANA, R. C. G. Dado e Granularidade na perspectiva da Informação e Tecnologia: uma interpretação pela Ciência da Informação. **Ciência da Informação**, Brasília, v. 42, p. 199-209, 2013. Disponível em:<http://revista.ibict.br/ciinf/article/view/1382>. Acesso em: 29/08/2017.
VAN RIJSBERGEN, C. J. **Information Retrieval**. Ed. 2. Butterworth-Heinemann Newton, MA, USA, 1979. Disponível em:<http://openlib.org/home/krichel/courses/lis618/readings/rijsbergen79_infor_retriev.pdf>. Acesso em: 29/08/2017.
VIERO, V. SILVEIRA, A. Apropriação de Tecnologias de Informação e Comunicação no meio rural brasileiro. **Cadernos de Ciência e Tecnologia – Embrapa**, Brasília, v. 28, n. 1, p. 257-277, 2011. Disponível em:<https://seer.sct.embrapa.br/index.php/cct/article/view/12042>. Acesso em: 29/08/2017.

PROPOSTA DE ARQUITETURA DE MICROSSERVIÇOS PARA UM SISTEMA DE CRM SOCIAL

Microservice Architecture Proposal For A Social Crm System

Luiz Felipe Correa Chiaradia, Douglas Dyllon Jeronimo Macedo, Moisés Lima Dutra
Universidade Federal de Santa Catarina (UFSC)
Florianópolis, SC – Brasil
luiz.chiaradia@posgrad.ufsc.br, douglas.macedo@ufsc.br, moises.dutra@ufsc.br.

Resumo:
A explosão informacional, impulsionada, principalmente, pelo uso massivo de serviços da Web 2.0 é vista como um desafio para as noções tradicionais do CRM, considerando-se que o consumidor passa a ter um papel ativo no relacionamento com a empresa. Neste contexto, surge o CRM Social que é construído a partir da integração das estratégias tradicionais da Gestão de Relacionamento com o Cliente com ferramentas capazes de recuperar, armazenar e analisar informações coletadas das redes sociais. Por meio de uma pesquisa qualitativa e aplicada, este artigo busca abordar os conceitos das áreas de Gestão de Relacionamento com o Cliente, CRM Social e Web 2.0, enumerando as características e benefícios oferecidos. Baseando-se nestas definições, propõe uma arquitetura de microsserviços para um sistema de CRM Social, que, apesar de aplicável para o enquadramento em questão, deverá ser testada com o intuito de determinar se a mesma atende aos requisitos propostos, visando performance e acurácia nas análises.
Palavras-chave: CRM Social; Web 2.0; Inteligência Competitiva; Microsserviços

Abstract:
The informational explosion, driven mainly by the massive use of Web 2.0 services, is seen as a challenge to the traditional conceptions of CRM, considering that the consumer starts to play an active role in the relationship with the company. In this context, Social CRM emerges, which is built on the integration of traditional Customer Relationship Management strategies with tools capable of retrieving, storing and analyzing information collected from social networks. Athwart a qualitative and applied research, this article pursues to grapple the concepts of the areas of customer relationship management, Social CRM and Web 2.0, enumerating the characteristics and benefits offered. Based on these definitions, it proposes a micro-service architecture for a Social CRM system, which, although applicable to the context in question, should be tested in order to determine whether it meets the proposed requirements, aiming to reach the desired performance and accuracy levels in analysis tasks.
Keywords: Social CRM; Web 2.0; Competitive Intelligence; Microservices.

1 Introdução

Em março de 2017, a rede social Facebook contabilizava com aproximadamente 2 bilhões de usuários, dos quais 75% permaneciam conectados vinte minutos ou mais diariamente (HUTCHINSON, 2017). Este uso massivo das redes sociais possibilitou que novas formas de comunicação e colaboração entre empresas e clientes ocorressem. Além disso, permitiu que as organizações criassem relacionamentos mais pessoais com seus consumidores, enquanto estes, passaram a ter a possibilidade de compartilhar livremente suas opiniões a respeito dos produtos adquiridos.

O cenário supracitado cria um desafio para as noções tradicionais da Gestão de Relacionamento com o Cliente, tradução para Customer Relationship Management (CRM), tendo em vista que, de acordo com Malthouse et al. (2013), o consumidor passa a ter um papel ativo no relacionamento com a empresa, dado que sua opinião é facilmente propagada entre sua rede de contatos. O chamado CRM Social é construído sobre percepções que empresas obtêm por meio da análise de uma grande quantidade de informações produzidas por seus clientes (MALTHOUSE et al., 2013) e surge como uma nova abordagem que visa adequar os costumeiros modelos de CRM à realidade atual.

Embora seja perceptível o crescimento de pesquisas relacionadas ao tema CRM Social, observa-se que os resultados apresentados, muitas vezes, denotam interpretações diferentes acerca da definição do termo. Além disso, não existe um modelo definido no que

diz respeito a como a informação será obtida e posteriormente recuperada e analisada. Conforme consta no trabalho de Marolt, Pucihar e Zimmermann (2015) a definição proposta por Greenberg (2009, p. 34) para o conceito de CRM Social é a mais aceita e o trata como uma nova abordagem que integra os conceitos tradicionais de relacionamento com o cliente com as aplicações de mídias sociais.

Nos textos de Malthouse et al. (2013), Trainor (2012), dentre outros, também são citados alguns conceitos e estratégias, no entanto, em nenhum momento se debatem métodos para armazenar, recuperar e analisar as informações obtidas. Orenga-roglá e Chalmeta (2016) afirmam que na literatura sobre CRM Social não há nenhuma metodologia específica ou modelo de arquitetura de sistemas que possa auxiliar no desenvolvimento de um sistema de CRM Social. As principais pesquisas focam, principalmente, nas características, oportunidades e benefícios que a estratégia de negócios oferece.

2 Objetivos

Este artigo aborda os conceitos das áreas de Gestão de Relacionamento com o Cliente, CRM Social e Web 2.0, enumerando as características e benefícios oferecidos. Baseando-se nesta caracterização, será proposta uma arquitetura de microsserviços para um sistema de CRM Social, na qual serão apresentados os fluxos de comunicação, bem como sugestões de tecnologias que poderão ser utilizadas para a realização das tarefas. Não cabe ao escopo deste trabalho apresentar trechos de códigos ou avaliações de desempenho, todavia, pretende-se criar um modelo que possa servir como base para futuras implementações.

Baseando-se nas assertivas anteriores, intenta-se com este trabalho realizar um aprofundamento dos conceitos de CRM Social, Web 2.0 e Gestão de Relacionamento com o Cliente e, fundamentando-se nestas definições, propor uma arquitetura baseada em microsserviços de um sistema de CRM Social.

3 Procedimentos Metodológicos

A metodologia adota para a realização deste trabalho consiste no levantamento e na análise de trabalhos publicados, focando as publicações que abordam temas correlacionados ao deste trabalho. Para as buscas, foram utilizados os termos “CRM”, “Social CRM”, “CRM 2.0” e “Web 2.0” nas bases Web of Science, ScienceDirect e Google Scholar. No que diz respeito ao ano das publicações, optou-se por selecionar apenas os artigos publicados a partir de 2010, no entanto, caso houvesse alguma citação, considerada importante para o trabalho, de um artigo publicado antes desta data, ele seria utilizado.

Na sequência, as publicações selecionadas serviram como base para a definição dos conceitos considerados como chave para a próxima etapa, na qual é proposta uma arquitetura baseada em microsserviços de um sistema de CRM social visando solucionar o problema de pesquisa.

4 Referencial teórico

Nesta seção, inicialmente serão apresentados os conceitos relacionados aos temas de Gestão de Relacionamento com o Cliente, Web 2.0 e CRM Social para, na sequência, apresentar a arquitetura de sistemas baseados em microsserviços.

4.1 Gestão de Relacionamento com o Cliente, Web 2.0 e CRM Social

A segunda geração dos serviços web, ou Web 2.0, possui uma enorme variedade de definições, descrições e princípios na literatura acadêmica. O principal objetivo da grande rede sempre foi promover a conexão entre diferentes indivíduos, no entanto, devido a evolução das formas de interação, o padrão de uso foi modificado (FAASE; HELMS; SPRUIT, 2011, tradução nossa). As redes sociais, como o Facebook e o Instagram, promovem o compartilhamento de grandes volumes de dados na Web por usuários que acessam uma plataforma aberta, cuja arquitetura é baseada em cooperação.

Sob um ponto de vista tecnológico, Mishra e Mishra (2009) classificam a Gestão de Relacionamento com o Cliente como uma abordagem orientada ao cliente que faz uso de sistemas de informação para fornecer informações que suportam os processos de

operação, análise e colaboração. Por meio da união de estratégias de relacionamento com o cliente do Marketing com a TI, o CRM oferece uma série de oportunidades para se trabalhar com informações visando entender e criar valor com os clientes (PAYNE; FROW, 2005).

Os conceitos sobreditos deixam claro que o processo de comunicação será gerido totalmente pela empresa, afirmativa que pode ser fundamentada pelo trabalho dos autores Faase, Helms e Spruit (2011). Em uma sociedade na qual a informação compartilhada cresce de forma exponencial, as estratégias criadas nos anos 2000 podem encontrar desafios. Neste contexto, surge o chamado CRM Social.

O CRM Social se trata de uma filosofia e uma estratégia de negócios, suportada por uma plataforma tecnológica, regras de negócios, processos e características sociais, desenhado para engajar o consumidor em uma conversa colaborativa com o objetivo de prover mutuamente valor em um ambiente de negócios confiável e transparente (GREENBERG, 2009, tradução nossa).

4.2 Arquitetura de Microsserviços

Baseado no Princípio da Responsabilidade Única de Robert C. Martin, Newman (2015) propôs uma arquitetura de microsserviços, na qual serviços pequenos e autônomos trabalham de forma cooperativa. Cada serviço possui uma única responsabilidade, de acordo com a regra de negócio que ele implementa, por conseguinte, torna-se fácil evitar que o código aumente de forma a prejudicar a sua manutenabilidade. Por ser uma entidade dissociada, a implantação de cada um dos microsserviços deve ser realizada de forma independente, como um serviço isolado em uma plataforma como serviço (PaaS), e toda a sua comunicação deverá ser realizada via chamadas de rede.

Resiliência diz respeito a habilidade de um sistema de se recuperar em uma falha. Caso algum componente apresente uma falha, é possível isolá-lo sem comprometer a disponibilidade do sistema. A Netflix se trata de um excelente exemplo de uso desta arquitetura: se o serviço de recomendações, por exemplo, falhar, o serviço de streaming continua operando normalmente. Todavia, a utilização de microsserviços também oferece um conjunto de desvantagens, que podem ser caracterizadas como complexidade de desenvolvimento, chamadas remotas e gerenciamento de múltiplos bancos de dados e transações.

5 Proposta e Resultados

Tradicionalmente, os processos relacionados à coleta de dados por sistemas de Gestão de Relacionamento com o Cliente foram responsáveis por departamentos como o de vendas e marketing. Estes buscam os potenciais clientes e fornecem dados ao sistema para que estes sejam utilizados e complementados em interações posteriores (ROUSE, 2014). Todavia, a evolução da internet e a introdução do conceito de rede colaborativa, permitiu a criação de uma enorme quantidade de ferramentas para comunicação digital e publicações, como as redes sociais. Como resultado, os usuários podem compartilhar suas ideias e propagar entre amigos, familiares e seguidores (ENNAJI et al., 2015). Este cenário fez com que as abordagens tradicionais do CRM fossem atualizadas com o objetivo de utilizar as mídias sociais, como o Facebook, Twitter e LinkedIn, para engajar seus clientes.

Visando a criação de valor das interações em mídias sociais, grandes empresas têm utilizado ferramentas que monitoram trocas de mensagens em redes sociais, quantificam menções de uma marca e palavras-chave para determinar o seu público-alvo e qual a plataforma que é mais utilizada (ROUSE, 2014). Ennaji et al. (2015), em seu trabalho propõe um framework de arquitetura monolítica para a extração e análise de dados extraídos de redes sociais para avaliar a opinião do público a respeito de um determinado produto:

- Módulo de extração de dados: responsável por extrair os dados das redes sociais.
- Módulo de análise de dados: responsável por analisar os dados extraídos das redes sociais.
- Dados das redes sociais: repositório onde os dados são armazenados. Será utilizado para

carregar o data warehouse e o sistema de CRM.

- Módulo de análises: fornece o resultado das análises.

Utilizando as funcionalidades sobreditas é possível propor um diagrama de camadas de um sistema de CRM Social que utiliza uma arquitetura de microsserviços, conforme consta na Figura 1.

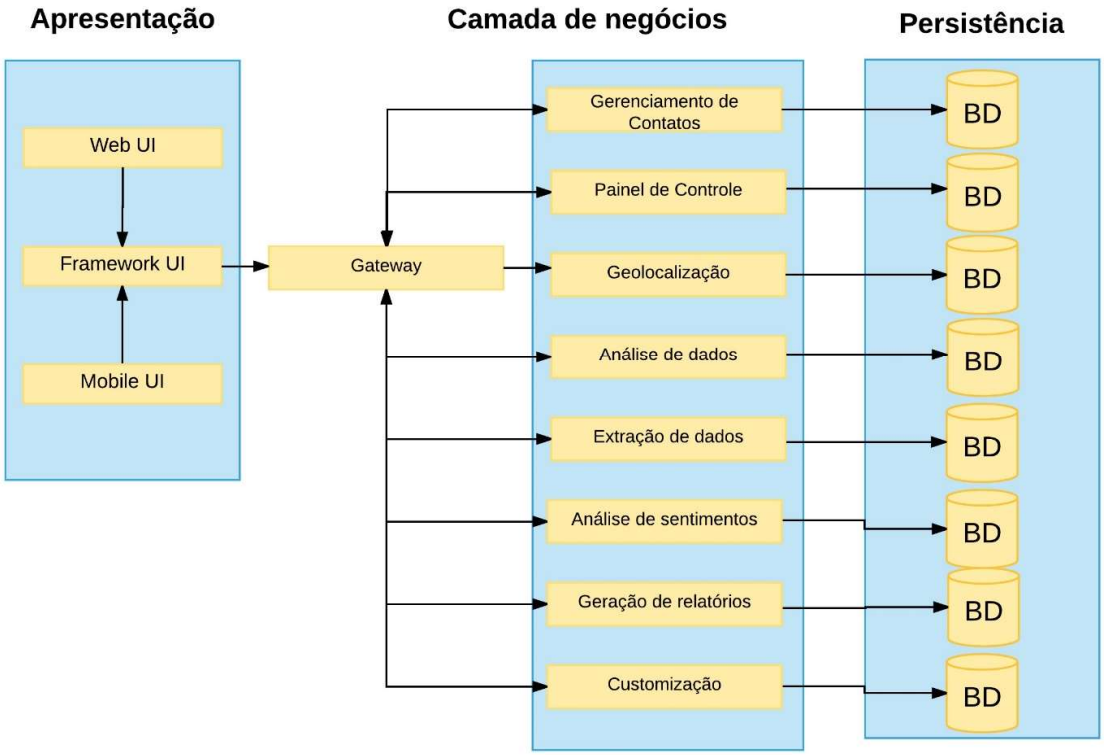


Figura 1 – Arquitetura de um sistema de CRM Social (Fonte: Elaborada pelo autor)

O elemento que representa o "gateway" será o responsável por receber as chamadas para os sistemas internos e será a única parte acessível externamente da API. Atuará como um filtro para o tráfego de chamadas e como um dispositivo de segurança para controle de acesso. Sua utilização auxilia na separação de camadas da aplicação e no controle de erros, tendo em vista que o mesmo será a interface para todas as requisições. Além disso, empregando-o é praticável o uso de um front-end único para o usuário final, acelerando a etapa de integração com a API (MALAVASI, 2016).

O CRM Social possui como uma de suas principais características o foco em todas as iterações das relações, seja entre companhias, parceiros ou clientes, por meio de diversos canais de comunicação. Dentre estes, é preciso destacar as redes sociais, tendo em vista as inúmeras oportunidades que estes ambientes oferecem para que as empresas obtenham informações mais detalhadas de seus consumidores. Dessa forma, em um sistema de CRM Social, os serviços de extração de dados e de análise de sentimentos devem ser cuidadosamente implementados. O primeiro, será responsável por buscar e coletar as informações consideradas pertinentes para a companhia e, considerando-se o grande volume dados contidos nestes ambientes, o uso do paradigma Big Data é mandatório para a realização desta tarefa (ENNAJI et al., 2015). Para tanto, será utilizado o framework de processamento distribuído Apache Spark, cujos mecanismos avançados para execução de grafos acíclicos dirigidos (DAG) e utilização do conceito de conjuntos de dados resilientes distribuídos (RDD) permitem a execução de programas cem vezes mais rápido que o Hadoop MapReduce em memória (GOPALANI; ARORA, 2015).

A análise de sentimentos tem como principal objetivo definir um conjunto de ferramentas capazes de extrair informação subjetiva a partir de textos em linguagem natural, como sentimentos e opiniões, para criar um conhecimento estruturado e acionável que possa ser utilizado por sistemas de suporte ao processo de tomada de decisão (POZZI et al., 2017, tradução nossa). No que tange as operações realizadas por este serviço, estas deverão incluir:

a. Classificação de subjetividade: o primeiro desafio encontrado quando se trabalha com a extração de informações subjetivas de textos em linguagem natural diz respeito a classificação das frases em objetiva ou subjetiva. Caso a frase expresse objetividade, nenhuma tarefa adicional é necessária. Todavia, caso expresse subjetividade, como visões e opiniões, será preciso determinar sua polaridade (positivo, negativo ou neutro).
b. Classificação de polaridade: esta tarefa é considerada a mais famosa na área de análise de sentimentos, pois visa determinar qual a polaridade (positivo, negativo ou neutro) do texto subjetivo que está sendo analisado.

A operação supracitada possui uma série de estratégias que vão desde a utilização de vocabulários previamente classificados até o uso de abordagens do aprendizado de máquina, por exemplo. Glorot, Bordes e Bengio (2011) propõe a utilização da abordagem de aprendizado profundo para a realização da classificação de polaridade de grandes volumes de dados. Este método, segundo os autores, prevê a aplicação de um conjunto de algoritmos que objetivam, por meio de grafos, criar abstrações de alto nível. Bastani (2014) afirma que esta tarefa também pode ser realizada utilizando abordagens como Máquinas de Vetores Suporte (SVM) e Entropia Máxima, no entanto, estas ainda são inferiores no que diz respeito a performance.

Os microsserviços aludidos merecem destaque por desempenharem tarefas críticas em um sistema de CRM Social, principalmente o que será responsável pela análise de sentimentos. Isto posto, evidência-se a necessidade da realização de um estudo mais aprofundado para determinar qual a abordagem mais eficiente para a execução do serviço. Feito isso, a arquitetura deste microsserviço será definida e, posteriormente, implementada.

6 Considerações Finais

A estratégia de negócios criada pelo CRM Social, apesar de ser um tema relativamente novo, já possui uma série de trabalhos acadêmicos publicados que buscam defini-la e listar seus inúmeros benefícios. Mesmo assim, poucos trabalhos fornecem uma metodologia para integrá-la com uma plataforma tecnológica, com o intuito de criar um modelo de arquitetura para um sistema de CRM Social. Por conseguinte, este trabalho buscou abordar os conceitos relacionados à Web 2.0, à gestão relacionamento com o cliente e ao CRM Social para propor um modelo de arquitetura de microsserviços.

A utilização de microsserviços oferece uma série de benefícios relacionados a performance e disponibilidade do sistema, no entanto, os custos para manter uma estrutura desta forma podem ser relativamente altos. Isto posto, posteriormente será necessário realizar uma série de testes de desempenho para avaliar se a arquitetura proposta é funcional e atende aos requisitos necessários. Além disso, o método que será utilizado para o serviço responsável pela análise de sentimento deverá ser estabelecido, visando precisão e desempenho, principalmente em cenários onde o volume de dados é alto.

Por meio desta pesquisa, foi possível conceituar os termos gestão de relacionamento com o cliente, CRM Social e Web 2.0 e propor uma arquitetura de microsserviços para um sistema de CRM Social. Levando-se em conta estes aspectos, entende-se que a referida arquitetura é aplicável para o enquadramento em questão, no entanto, serão necessários realizar testes de desempenho com o intuito de determinar se a mesma atende aos requisitos propostos, visando performance e precisão nas análises.

Referências

BASTANI, Kenny. **Deep Learning Sentiment Analysis for Movie Reviews using Neo4j.** 2014. Disponível em: <http://www.kennybastani.com/2014/09/deep-learning-sentiment-analysis-for.html?spref=tw>. Acesso em: 04 jun. 2017.

ENNAJI, Fatima Zohra et al. **Social intelligence framework: Extracting and analyzing opinions for social CRM**. In: Computer Systems and Applications (AICCSA), 2015 IEEE/ACS 12th International Conference of. IEEE, 2015. p. 1-7.

FAASE, Robbert; HELMS, Remko; SPRUIT, Marco. **Web 2.0 in the CRM domain: defining social CRM**. Ijecrm, [s.l.], v. 5, n. 1, p.1-21, 2011. Inderscience Publishers.

GOPALANI, Satish; ARORA, Rohan. **Comparing apache spark and map reduce with performance analysis using k-means**. International Journal of Computer Applications, v. 113, n. 1, 2015.

GLOROT, Xavier; BORDES, Antoine; BENGIO, Yoshua. **Domain adaptation for large-scale sentiment classification: A deep learning approach**. In: Proceedings of the 28th international conference on machine learning (ICML-11). 2011. p. 513-520.

GREENBERG, Paul. **Social CRM Comes of Age.** 2009. White Paper. Disponível em: <http://www.computerworlduk.com/cmsdata/whitepapers/3203130/social crm.pdf>. Acesso em: 16 de julho de 2017.

GREENBERG, Paul. **The impact of CRM 2.0 on customer insight.** Jnl Of Bus & Indus Marketing, [s.l.], v. 25, n. 6, p.410-419, 3 ago. 2010. Emerald.

HUTCHINSON, Andrew. **Top Social Network Demographics 2017.** 2017. Elaborada por Trackx.

LEHMKUHL, Tobias; JUNG, Reinhard. **Towards Social CRM:** Scoping the concept and guiding research. 2013.

MALAVASI, Eike. **API Gateway governando a arquitetura de Microservices.** 2016. Disponível em: <http://sensedia.com/blog/apis/api-gateway-governando-a-arquitetura-de-microservices/>. Acesso em: 04 jun. 2017.

MALTHOUSE, Edward C. et al. **Managing Customer Relationships in the Social Media Era: Introducing the Social CRM House**. Journal Of Interactive Marketing, [s.l.], v. 27, n. 4, p.270-280, nov. 2013. Elsevier BV.

MAROLT, Marjeta; PUCIHAR, Andreja; ZIMMERMANN, Hans-dieter. **Social CRM Adoption and its Impact on Performance Outcomes: a Literature Review**. Organizacija, [s.l.], v. 48, n. 4, p.260-271, 1 jan. 2015.

MISHRA, Alok; MISHRA, Deepti. **Customer Relationship Management: implementation process perspective**. Acta Polytechnica Hungarica, v. 6, n. 4, p. 83-99, 2009

NEWMAN, Sam. **Building microservices**. Sebastopol: O'reilly Media, 2015. 280 p. 1 v.

ORENGA-ROGLÁ, Sergio; CHALMETA, Ricardo. **Social customer relationship management: taking advantage of Web 2.0 and Big Data technologies**. Springerplus, [s.l.], v. 5, n. 1, p.1-17, 31 ago. 2016. Springer Nature

PAYNE, Adrian; FROW, Pennie. **A Strategic Framework for Customer Relationship Management**. **American Marketing Association,** [s. L.], v. 4, n. 69, p.167-176, out. 2005.

POZZI, Federico Alberto et al. Challanges of Sentiment Analysis in Social Networks: An Overview. In: POZZI, Federico Alberto et al. **Sentiment Analysisin Social Networks.** Cambridge: Elsevier, 2017. Cap. 1. p. 11-22.

ROUSE, Margaret. **Customer Relationship Management (CRM)**. 2014. Disponível em: <http://searchcrm.techtarget.com/definition/CRM>. Acesso em: 04 jun. 2017.

TRAINOR, Kevin J.. **Relating Social Media Technologies to Performance: A Capabilities-Based Perspective**. Journal Of Personal Selling And Sales Management, [s.l.], v. 32, n. 3, p.317-331, 1 jul. 2012. Informa UK Limited.

WIDAT'2017 – Sessão Técnica II

Dia: 05/09/2017 – Horário: 14h00 às 16h00

MODELO DE ATUALIZAÇÃO DE BASES DE CONHECIMENTO: um estudo de caso ONTO-AmazonTimber

Knowledge base update model: an ONTO-AmazonTimber case study

Ademilson de Almeida Barbosa[1], Márcio José Moutinho Da Ponte[2], Celson Pantoja Lima[3]

(1) Universidade Federal Do Oeste Do Pará / Instituto De Engenharia E Geociências
Programa De Computação, Av. Mendonça Furtado, 2946 - Fátima, Campus Universitário
Santarém - PA, 68040-470, ad.ufopa@gmail.com.

(2) Universidade Federal Do Oeste Do Pará / Instituto De Engenharia E Geociências
Programa De Computação, Av. Mendonça Furtado, 2946 - Fátima, Campus Universitário
Santarém - PA, 68040-470, marcio.ponte@ufopa.edu.br.

(3) Universidade Federal Do Oeste Do Pará / Instituto De Engenharia E Geociências
Programa De Computação, Av. Mendonça Furtado, 2946 - Fátima, Campus Universitário
Santarém - PA, 68040-470, celson.lima@ufopa.edu.br.

Resumo:
O contexto tecnológico apresenta novos desafios ao processo de produção e atualização de conhecimento. A constante manutenção de bases de conhecimento, permite a evolução no domínio do conhecimento, no entanto somente a consistência do conhecimento inserido, possibilita a eficácia dos resultados nas interações semânticas. Neste contexto, este trabalho objetiva desenvolver um modelo de atualização de bases de conhecimento, e será validado por um estudo de caso aplicado a ontologia ONTO-AmazonTimber, desenvolvida como suporte ao processo de identificação botânica em espécies florestais da Amazônia. O desenvolvimento da pesquisa, utiliza-se da metodologia e-COGNOS e o método de atualização de bases conhecimento, que obedece três processos: o cadastro e a validação do conhecimento e, a inserção deste na base de conhecimento. Com a aplicação do modelo de atualização na ontologia ONTO-AmazonTimber, verificou-se que é possível minimizar a inserção de conhecimentos redundantes e incorretos e, a consistência do conhecimento atribuído que repercute em toda a estrutura semântica.
Palavras-chave: Gestão do Conhecimento; Ontologia; Atualização de conhecimento; ONTO-AmazonTimber.

Abstract:
The technological context presents new challenges to the process of production and updating of knowledge. The constant maintenance of knowledge bases, allows the evolution in the domain of knowledge, however only the consistency of the inserted knowledge, makes possible the effectiveness of the results in the semantic interactions. In this context, this work aims to develop a knowledge base updating model, and will be validated by a case study applied to the ONTO-AmazonTimber ontology, developed as a support to the botanical identification process in Amazonian forest species. The development of the research uses the e-COGNOS methodology and the knowledge base updating method, which obeys three processes: the registration and validation of knowledge and its insertion into the knowledge base. With the application of the update model in the ONTO-AmazonTimber ontology, it was verified that it is possible to minimize the insertion of redundant and incorrect knowledge and the consistency of the attributed knowledge that has repercussions throughout the semantic structure
Keywords: Knowledge management; Ontology; Knowledge Update; ONTO-AmazonTimber.

1 Introdução

Segundo Luckesi e Passos (1996) o conhecimento é o resultado do processamento analítico de informações que fornecem os subsídios essenciais para tomadas de decisões. A Gestão do Conhecimento (GC) por sua vez, está intimamente relacionada ao fator sucesso no processo de tomada de decisões, o que tende a aumentar à medida que se intensifica a interação entre a produção de conhecimento e a tecnologia (ROSSETTI e MORALES, 2007).

Nesta perspectiva, o atual contexto tecnológico direciona-se a produção do conhecimento, o que permite expandir o alcance nos domínios de aplicação e potencializar as funcionalidades de um referencial semântico.

Ocorre, que domínios de conhecimento raramente são estáticos e, assim sendo, a ontologia deve acompanhar a evolução do domínio de conhecimento. Caso mudanças do domínio do conhecimento não sejam mapeadas e incorporadas nas bases de conhecimento, está se tornará estagnada,

ultrapassada, ineficaz e possivelmente incorreta (PONTE, 2017).

A atualização de bases de conhecimento, pode ser consequência da descoberta de erros na modelagem, ou um reparo no conhecimento de domínio. Para que a aquisição de novos conhecimentos sejam adicionadas, segundo (LÖSCH et.al., 2009), as pesquisas devem ser dedicadas empregando técnicas de revisão de crenças, aquisição de conhecimento, aprendizagem de ontologia e evolução de ontologia para citar apenas algumas. Em um nível mais geral, as especificações de atualização, permitem codificar o conhecimento do processo e associá-lo à ontologia, de modo que possa ser usado para atualizações.

É necessário, nesta senda, que as mudanças mediadas por especificações da atualização de ontologias possam ser diretamente evidentes para o construtor do conhecimento (LÖSCH et. al., 2009). Este, por sua vez é crucial para garantir que o comportamento do sistema seja totalmente transparente ao construtor desse conhecimento

No que se refere ao contexto semântico, o excesso ou escassez da abordagem do domínio pode torna-se volumoso ou obsoleto. Para Prado (2001), segundo os critérios para construção de uma base de conhecimento, toda a aquisição de conhecimento é destacada como uma prioridade e, todo o conhecimento deve ser validado por especialistas que consigam destacar a relevância do conhecimento para a base de domínio da ontologia.

2 Objetivos

Este trabalho visa propor um modelo de atualização para bases de conhecimento, e como estudo de caso instancia-se o modelo de atualização para a base de conhecimento da ontologia ONTO-AmazonTimber.

3 Procedimentos Metodológicos

Este tópico destina-se a apresentar as delimitações metodológicas deste trabalho, isto inclui as metodologias aplicadas, os métodos e procedimentos utilizados para o desenvolvimento do modelo de atualização de bases de conhecimento.

3.1 Metodologia para o desenvolvimento da Ontologia

A metodologia aplicada neste trabalho advém do projeto e-COGNOS (Lima, El-Diraby e Stephens, 2005) no qual foi desenvolvida uma plataforma de gestão do conhecimento baseada na web.

O método proposto para o desenvolvimento da ontologia, foi inspirado na abordagem usada pelo projeto e-COGNOS, motivado pela participação da construção e refinamento da metodologia e do histórico de bons resultados angariados até o momento.

Os principais conceitos que servem como a espinha dorsal da ontologia também foram inspirados na ontologia e-COGNOS. No entanto, para o propósito deste artigo, algumas adaptações e refinamentos do modelo ontológico tiveram de ser feitas.

O método aqui adotado usa uma abordagem iterativa (Figura 01), que é dividida em várias fases, com cada fase contendo um conjunto de tarefas relacionadas.[1]

Figura 01. Metodologia e-COGNOS para construção da Ontologia, fonte: Adaptado de Costa (2014).

3.2 Inserção na estrutura semântica da Ontologia

O método de atualização para a base de conhecimento proposto, se utiliza de uma ontologia de referência, nomeada como ontologia ONTO-AmazonTimber, para assim conseguir adicionar novos conceitos a ontologia básica.

[1] Para uma descrição mais detalhada da metodologia empregada analisar em PONTE, 2017.

Este método é realizado seguindo três etapas. Em primeiro lugar se utiliza um medida de similaridade semântica na ontologia de referência, afim de reconhecer ou encontrar o sentido mais similar ao conceito, de modo a adicionar a ontologia básica.

Na segunda etapa, se procura a posição correta em que este conceito condiz com o sentido e pode ser encaixado a ontologia básica. No fim o conceito é adicionado a ontologia básica sempre respeitando a estrutura hierárquica.

4 Resultados

A ONTO-AmazonTimber trata-se de um referencial semântico no âmbito da botânica que formaliza o conhecimento do especialista (taxonomista) na tarefa de identificação botânica. Dispondo de conhecimentos como: caracterização das espécies botânicas comercializadas no setor madeireiro da Amazônia; caracterização do ambiente em que estão inseridas; e caracterização do contexto ambiental.

Para validar o modelo apresentado nesse trabalho, introduziu-se uma abordagem de resultados baseados em uma proposta de modelo de atualização de bases de conhecimento a partir do estudo de caso da ontologia ONTO-AmazonTimber.

4.1 Modelo conceitual de atualização de base de conhecimento

Segundo Lisboa (2000), os modelos conceituais permitem representar de maneira abstrata, formal e não ambígua a realidade da aplicação, com objetivo de facilitar a comunicação entre os envolvidos no projeto.

O modelo conceitual apresentado neste trabalho, permite mensurar o escopo do planejamento de atualização da base de conhecimento, assim como a base teórica do domínio do conhecimento e a especificação da estruturação dos elementos envolvidos, possibilitando que um novo conhecimento possa compor a base de conhecimento.

Para que haja sustentação do modelo conceitual proposto, os seguintes elementos estruturais são fundamentais: Especialista, Conhecimento, Comissão de Validação e Base de Conhecimento (Figura 02).

O Especialista representa um grupo de profissionais que trabalham no domínio do conhecimento da problemática em questão. Sua formação e experiência são consideradas de alto valor quando se trata da formalização e atualização do conhecimento.

A comissão de validação se utiliza dos seguintes critérios: a Formação e, a Experiência do Especialista quando avalia a proposta de um novo conhecimento para ser inserido a base de conhecimento.

Note-se, que nesta etapa o conhecimento ainda não foi inserido, tendo em vista a atuação do especialista que no processo de validação, confere ao conhecimento mera expectativa de inserção, já que somente quando um novo conhecimento é validado, que este irá compor a base de conhecimento.

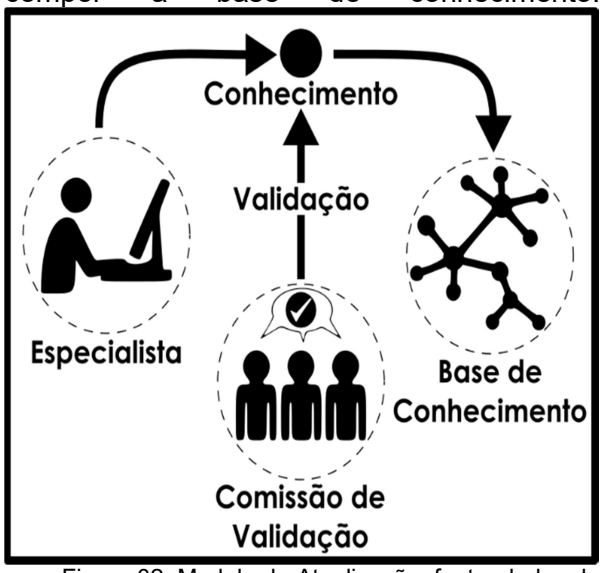


Figura 02. Modelo de Atualização, fonte: dados da pesquisa, (2017).

4.2 Instanciação do Modelo de atualização aplicado a ONTO-AmazonTimber.

A instanciação do modelo de atualização desenvolvido neste trabalho é representada por elementos estruturais que se fundamentam quando aplicados no âmbito da botânica (Figura 03).

O Especialista representa um grupo de profissionais que atuam no processo de identificação botânica, destacam-se: Taxonomistas, Parataxonomista, Botânicos e Engenheiros Florestais com experiência em inventário florestal. As especializações, habilidades, experiências e formações no domínio da botânica servem como quesitos

fundamentais na formalização de novos conhecimentos.

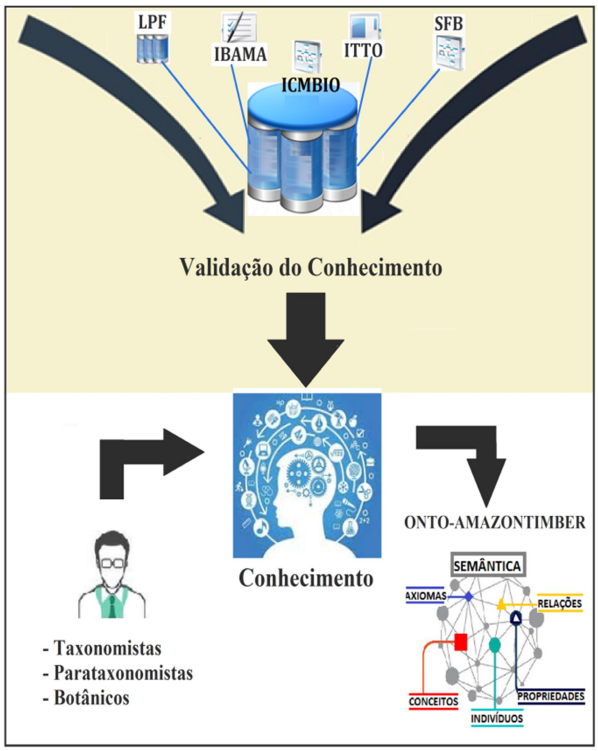


Figura 03. Instanciação do Modelo de Atualização, fonte: dados da pesquisa, (2017).

A comissão de validação é representada por indivíduos inseridos em entidades de fiscalização, pesquisa e promoção da sustentabilidade ambiental, a citar: Instituto Brasileiro do Meio Ambiente e dos Recursos Naturais Renováveis (IBAMA), Laboratório de Produtos Florestais (LPF) da Universidade Federal do Oeste do Pará (UFOPA), Serviço Florestal Brasileiro (SFB), Instituto Chico Mendes de Conservação da Biodiversidade (ICMBio).

O modelo de atualização proposto apresenta os seguintes processos: o especialista propõe um novo conhecimento. Tal conhecimento é avaliado pela comissão de validação, que devem utilizar os seguintes critérios de avaliação: formação e experiência do profissional, para assim analisar e validar o novo conhecimento.

A base de conhecimento da ONTO-AmazonTimber trata-se do terceiro elemento do modelo, formalizado por uma estrutura semântica composta por conceitos, relações, propriedades e axiomas, instanciados e inter-relacionados para compor o domínio de conhecimento em questão.

4.3 Interfaces do modelo de atualização da ONTO-AmazonTimber

Neste tópico apresentam-se protótipos de telas desenvolvidos na elaboração do modelo de atualização. O processo de interação do modelo de atualização inicia-se com a seleção da espécie de interesse (Figura 04).

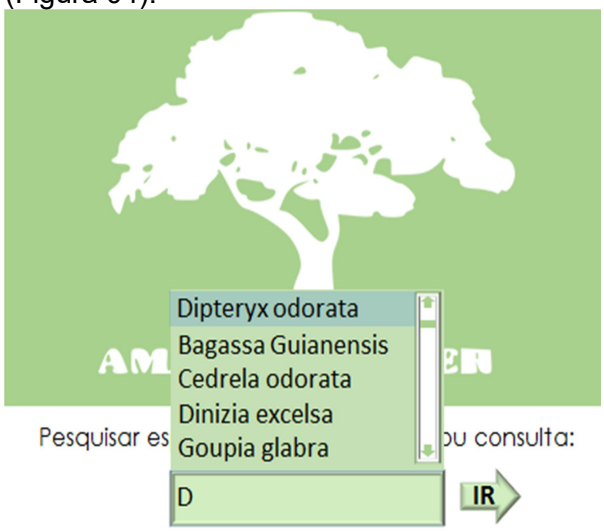


Figura 04. Tela de seleção, fonte: dados da pesquisa, (2017).

Posteriormente é exibido todas as relações semânticas de propiciam a caracterização da espécie selecionada (Figura 05). Permitindo desta forma incluir uma nova características.

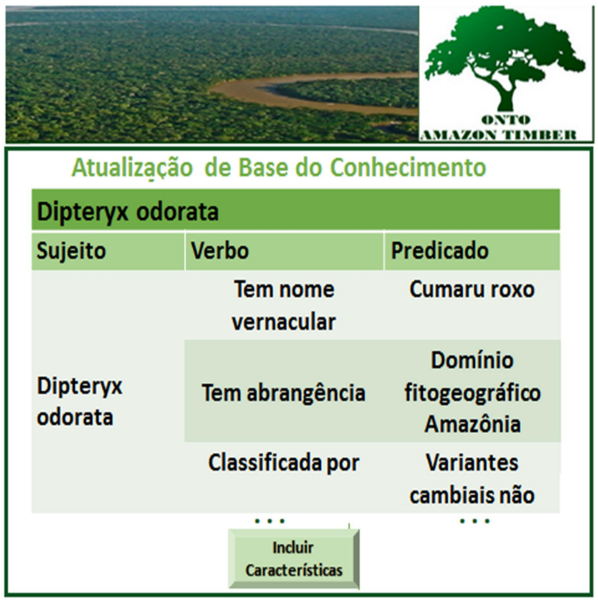


Figura 05. Tela de caracterização, fonte: dados da pesquisa, (2017).

Desta forma na tela de inclusão (Figura 06), é selecionado a estrutura semântica e incluir ou cadastrar novas espécies, verbos e predicados.

Figura 06. Tela de inclusão, fonte: dados da pesquisa, (2017).

4.4 Interação Ontológica com JENA

A JENA consiste em um framework JAVA que permite trabalhar em ambiente de programação com manipulação dinâmica de modelos RDF (Resource Description Framework), representadas pelos recursos, propriedades e literals, formando as tuplas (predicate, [subject], [object]) que originam os objetos criados pelo JAVA.

Este Framework disponibiliza um kit de funcionalidades para apoiar o desenvolvimento de aplicações no contexto de ontologias. Além das funcionalidades para manipulação da linguagem OWL e uso do *Simple Protocol And Rdf Query Language (*SPARQL).

A API JENA apresenta um conjunto de métodos que permitem acessar os elementos de uma Ontologia (classes, propriedades e indivíduos) podemos utilizar os métodos iniciados com *list* como listClasses(), listIndividuals() ou listSubClasses(). A partir deles pode-se chamar o método toList() para ter acesso aos elementos através de uma instância da classe java.util.List. Além disso, para identificar qual classe ou instância está sendo manipulada dentro das iterações, temos dois métodos básicos: getURI(), que retorna o nome completo ou a URI (prefixo + nome) do objeto; e getLocalName(), que retorna apenas o nome do objeto em questão.

Outros métodos permitem uma maior especificação quanto ao acesso a estrutura ontológica, o método *getObjectsFromObjectTriple* possibilita listar um conjunto de objetos da classe A que se relacionam através de uma propriedade específica com um outro objeto da classe B. Para melhor entendimento visualiza-se o código abaixo:

```
OntologyInteraction listaCaracteristica = new OntologyInteraction() ;

ArrayList<String> objects = listaCaracteristica.
getObjectsFromObjectTriple("Dipteryx_odorata", "ClassificadoPor");
```

Tal método *getObjectsFromObjectTriple* tem como função listar as características botânicas da espécie botânica Dipteryx_odorata que estão interligadas pela propriedade ClassificadaPor, tal relação semânticas podem ser observadas na (Figura 07), no qual a classe Species tem uma série de objetos dentre estes o Dipteryx_odorata, por sua vez apresenta algumas propriedade de objetos que criam relações com outros objetos, como por exemplo: *Heartwood_Distinct_Color* instancia da classe *Heartwood_Color.*

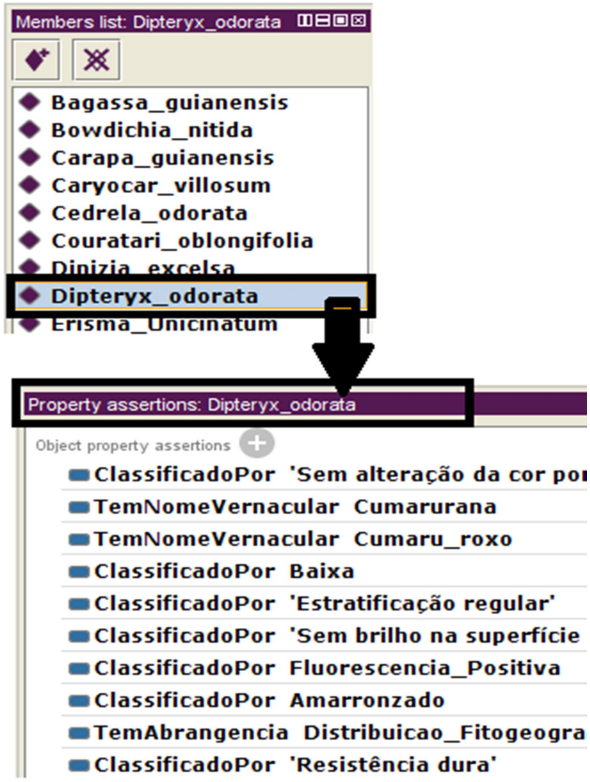


Figura 07. Relações semânticas obtidas pelo método *getObjectsFromObjectTriple,* fonte: PONTE (2017).

4 Considerações Finais

O modelo conceitual de atualização desenvolvido neste trabalho apresenta requisitos necessários para manutenção de bases do conhecimento nos mais diversos domínios do conhecimento. Premissa necessária para contínua evolução da estrutura semântica, item indispensável na Gestão do Conhecimento.

Considerando-se a extensa diversidade biológica da Amazônia e a vasta quantidade de espécies florestais existentes na Amazônia, aliado a escassez de profissionais que atuem no processo de identificação botânica, evidencia-se a extrema necessidade da constante atualização da base de conhecimento.

O modelo de atualização proposto permite minimizar a inserção de conhecimentos redundantes e incorretos e, a consistência do conhecimento atribuído que repercute em toda a estrutura semântica.

Referências

COSTA, R. D. D. Semantic Enrichment of Knowledge Sources Supported by Domain Ontologies. Tese de Doutorado em Engenharia Electrotécnica e de Computadores – Faculdade Ciências e Tecnologia, Universidade Nova de Lisboa, Portugal - Lisboa, 2014.

LIMA,C.; EL-DIRABY,T.; STEPHENS,J. Ontology-basedoptimisation of knowledge management in e-Construction. ITcon 10, 305–327. 2005.

LISBOA F., J. Modelagem de Banco de Dados Geográficos. In: LADEIRA, M.; NASCIMENTO, M. E. M. III Escola Regional de Informática do Centro-Oeste. Brasília – DF. SBC –Sociedade Brasileira de Computação, 2000.

LÖSCH U., RUDOLPH S., VRANDE˘CI´C D., and STUDER R., “Tempus fugit: Towards an Ontology Update Language”, Institut AIFB - Universit¨at Karlsruhe (TH): Karlsruhe, Germany: 2009.

LUCKESI, C. C. e PASSOS, E.S. “Introdução à filosofia: aprendendo a pensar.”, São Paulo: Cortez, 1996.

MOTA, M. R. A. “Mapeamento Sistemático Sobre o Uso de Ontologias em informática Médica”, João Pessoa: Agosto, 2013.

PONTE M. J. M, “Referencial Semântico no suporte da identificação botânica de Espécies Amazônicas”, Tese de doutorado – UFOPA/UNL -Santarém – Pará, Abril 2017.

PRADO, E. V. “Sistema Especialista para dimensionamento e seleção de equipamentos para pré-processamento de café ǁ”, “Tese de doutorado” – UFV, 2001.

ROSSETTI A. G., MORALES A. B. T. O papela da tecnologia da informação na gestão do conhecimento. C i. Inf., Brasília, v. 36, n. 1, p. 124-135, jan./abr. 2007.

RUSSEL, S. N. P. “Artificial Intelligence: A Modern Approach.” 3rd Edition, New Jersey: Prentice Hall; 2009.

Desenvolvimento de Ontologia Ciente de Qualidade de Informações para a Melhoria de Consciência Situacional no Domínio de Gerenciamento de Emergências

Development of Ontology Aware of Information Quality for the Improvement of Situational Awareness in the Field of Emergency Management

Jordana N. Silva[1], Jessica Souza[1], Állan César M. de Oliveira[1], Maria de Fátima Tavares[2], Leonardo C. Botega[1]

(1) Grupo de Interação Humano-Computador, Centro Universitário Eurípides de Marília, Marilia/SP,
(2) Instituto Brasileiro de Informação em Ciência e Tecnologia, Brasília/DF,
jordanasnogueira@gmail.com
osz.jessica@gmail.com, allan_oliveira@univem.edu.br, fatimatavares@ibict.br, botega@univem.edu.br

Resumo:
Uma questão desafiadora na comunidade da Avaliação da Situação é determinar como o processo de avaliação pode ser redesenhado para o aprimoramento da Consciência da Situação (SAW), que pode ser severamente degradada se dados de baixa qualidade propagarem pelo processo comprometendo assim a tomada de decisões. Em sistemas de gerenciamento de emergências o grande desafio de adquirir e manter a SAW em operadores humano, é o consumo em excesso de dados em um ambiente dinâmico. O presente trabalho tem por objetivo o desenvolvimento de uma ontologia de domínio para o gerenciamento de emergências mais especificamente sobre incêndios florestais visando contribuir para a representação de informações, contribuindo assim com processos de avaliação de situações de fogo. Para tal, foram desenvolvidas entrevistas com especialistas, modelagem conceitual das tarefas e objetivos e o emprego de ferramentas e metodologias para a construção da ontologia que ao final será incorporada ao sistema "Distrito Federal Sem Fogo" (DF100Fogo).

Palavras-chave: Consciência Situacional; Ontologia; Gerenciamento de Emergência; Qualidade de dados;

Abstract:
A challenging issue in the Situation Assessment community is to determine how the assessment process can be redesigned to improve Situation Awareness (SAW), which can be severely degraded if poor quality data propagate through the process, thereby compromising decision-making. In emergency management systems, the great challenge of acquiring and maintaining SAW in human operators is consuming too much data in a dynamic environment. The present work has the objective of developing a domain ontology for the emergencies management, more specifically about forest fires, aiming to contribute to the representation of information in this domain, thus contributing to fire assessment processes. To this end, interviews with specialists, conceptual modeling of tasks and objectives and the use of tools and methodologies for the construction of the ontology were developed, which in the end will be incorporated into the Distrito Federal Sem Fogo (DF100Fogo) system.

Keywords: Situational Awareness; Ontology; Emergency Management; Data quality;

1. Introdução

Consciência da situação (SAW) é um processo cognitivo caracterizado pela percepção, entendimento e projeção dos contextos que envolvem atividades de entidades de interesse em um dado local e em um determinado intervalo de tempo (Endsley, 2003).

No domínio de gerenciamento de emergências, adquirir e manter SAW é crucial para que operadores de sistemas para este fim sejam orientados com melhores subsídios informacionais e assim sustentar o processo de tomada de decisão.

Adquirir e manter a SAW de humanos operadores de sistemas críticos como os de gerenciamento de emergências mais especificamente em sistemas de monitoramento de queimadas é uma tarefa complexa. Para se construir a SAW de tais

operadores, sistemas de gerenciamento de emergências coletam, processam e representam dados provenientes principalmente de relatos humanos (HUMINT) que em geral são heterogêneos, imprevisíveis, complexos e dinâmicos.

Tais dados objetivam descrever ambientes reais monitorados e caracterizar o comportamento observável de cada entidade de interesse influenciando a forma com que o mundo real é descrito, entendido e processado por humanos e sistemas.

A qualidade inerente as informações HUMINT adquiridas de denúncias consistem em um fator fundamental para a aquisição de SAW. Uma má qualidade de dados pode influenciar negativamente humanos operadores de sistemas e processos computacionais que dependem de parâmetros corretos para produzir melhores informações e contribuir para uma tomada de decisão mais assertiva.

Neste contexto, foi identificado que o uso de modelos semânticos como a ontologias, são capazes de ajudar a representar informações que podem ser úteis ao sistema como um todo e ao próprio usuário final agregando significados métricas sobre a qualidade dos dados, além de relacioná-las em um contexto em constante mudança.

Assim, este trabalho tem como objetivo desenvolver uma ontologia de domínio que detenha conhecimento sobre a qualidade das informações capaz de auxiliar sistemas de gerenciamento de emergência em processos de inferência de situações, e consequentemente o combate a incêndios florestais.

2. O Projeto DF100Fogo

O projeto DF100Fogo (Saran et. al. 2017), tem por objetivo auxiliar o Corpo de Bombeiros do Distrito Federal e as Brigadas do Jardim Botânico de Brasília (JBB) no combate e controle de incêndios florestais. Atualmente o desenvolvimento e manutenção deste projeto está sob responsabilidade de uma parceria entre o Centro Universitário Eurípides de Marília (UNIVEM), o Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT) e a Universidade Federal de São Carlos (UFSCar).

O DF100Fogo (Saran et. al. 2017) é composto por um aplicativo voltado para a comunidade e tem por funções notificar os Bombeiros sobre focos de incêndio, essas notificações podem ser enviadas por áudio, texto ou foto.

Há também um servidor que processa os dados enviados pelas notificações que junto ao sistema administrador permite uma visualização das informações sobre as notificações e a real situação do incêndio, ainda é possível agregar a ocorrência de incêndio informações de georreferenciamento e clima, tudo isso impacta na gestão de informações de operadores humanos.

As notificações também podem ser visualizadas por bombeiros e brigadistas em patrulha no local da ocorrência através de um outro aplicativo desenvolvido, facilitando assim o deslocamento de equipes e alocações de recursos para combate e controle do incêndio.

3. Ontologias e a Qualidade de Dados no Auxílio à Obtenção de SAW

De acordo com Endsley (2003), Consciência da Situacional (SAW) é um processo cognitivo do humano composto por 3 níveis: percepção dos elementos no ambiente, compreensão do estado destes elementos compondo uma situação e a evolução destes em um estado futuro próximo.

Para Endsley (2003), a SAW tem sua aplicabilidade em situações operacionais e rotineiras como por exemplo dirigir um carro por exemplo, onde o motorista deve ter a consciência situacional por uma razão especifica, no caso prestar atenção ao trafego a sua volta. Portanto “[…] consciência situacional tem seus requisitos definidos como as necessidades dinâmicas de informação associadas aos principais objetivos ou subobjetivos do operador na realização de seu trabalho." (Endsley, 2003, p 269).

SAW tem sido utilizada como suporte em sistemas operados dentro do domínio de gerenciamento de emergência onde seus operadores precisão obter respostas rápidas e eficientes e serem capazes de interpretar tais respostas de maneira clara e objetiva, visando uma tomada de decisão mais assertiva mesmo que o ambiente ao redor seja de extrema pressão e com um grande volume de dados e informações sendo consumida em tempo real.

O modelo de SAW proposto por Endsley (2003), definiu ainda que este tem por objetivo medir o processamento e o consumo de informações por humanos.

De acordo com Kokar et al. (2009), a principal diferença do uso de SAW entre o processamento computacional e humano é que em humanos ele já é suportado e pode ser medido, enquanto em computadores esse processo precisa ser definido e implementado, sendo assim, é necessário desenvolver junto ao projeto uma especificação para que sua implementação seja correta. Neste contexto segundo Kokar et al. (2009), o uso de computação baseada em ontologias ajudaria a desenvolver um modelo no qual processos computacionais deteriam a conscientização da situação.

A ontologia para ser empregada dentro de um processo computacional deve ser "escrita em um idioma processável e normalmente suportado" (Kokar et al. 2009), visando criar uma representação de objetos, termos suas relações e restrições no contexto que está inserida. A maneira como os termos do domínio são declarados ontologicamente gera suporte para que outros fatos possam ser deduzidos e usados como base para a inferência desses processos. Esse tipo de abordagem em sistemas de gerenciamento de emergência gera ao operador um maior conhecimento das milhares de relações possíveis entre os termos citados em um atendimento de emergência, gerando assim uma resposta capaz de subsidiar a tomada de decisão por parte do operador.

Juntamente com o uso da ontologia para o auxílio em processos computacionais visando uma melhor representação de informações, a qualidade de dados a serem consumidos por esses processos é um ponto crítico, pois geralmente os dados provenientes de humanos podem conter imperfeições e não serem concisos, reduzindo assim a efetividade desses sistemas e contribuindo negativamente para o desenvolvimento da SAW de seus operadores.

O gerenciamento da qualidade de dados se dá pela "definição de papeis, responsabilidades, políticas e procedimentos relacionados à aquisição, manutenção, representação e disseminação de dados e informações", (Botega et al., 2017).

Botega et al. (2017) descreve que o entendimento de situações pelos operadores de sistemas de gerenciamento de emergência é fundamental para o processo de tomada de decisão, além de influenciar no comportamento de humanos e sistemas, visto que a falta desse entendimento em domínios em que, manter e proteger a vida humana é o principal objetivo, podem acarretar a erros inevitáveis e irreversíveis.

4. Metodologia

É conhecido que ontologias servem de base à sistemas que visam suportar SAW, processos computacionais como mineração e fusão podem se beneficiar de suas relações e inferir novas informações utilizando seus significados representados (Botega et al., 2017).

O primeiro passo para atingir os objetivos propostos foi a realização de uma entrevista estruturada com o Corpo de Bombeiros Militar do Distrito Federal (CBMDF) e as Brigadas de combate a incêndios do JBB onde foi possível identificar suas principais atividades e procedimentos para o atendimento de uma ocorrência de incêndio.

Após essa etapa, foi realizado um levantamento dos principais termos e entidades presentes nas respostas fornecidas pelos bombeiros, levando em consideração a patente e o tempo de serviço de cada bombeiro. Considerando

questões sobre o “clima”, os termos mais citados foram: estação do ano, temperatura, umidade e velocidade do vendo. Sobre termos relacionados à localização foram mencionados: endereço, tipo da área, tipo de vegetação e principais vias de acesso. À respeito da vítima: foram apontadas como principais informações, a quantidade e o seu estado. Sobre o incêndio propriamente dito é importante saber sobre o seu nível (pequeno, médio ou grande), o seu tipo (subterrâneo, superfície ou de copa) sobre a sua intensidade, dimensão, velocidade de propagação, altura das chamas e cor da fumaça.

Com esse resultado foi possível realizar a construção e manutenção de uma Análise de Tarefas Dirigida por Objetivos, ou GDTA, sendo que parte dele é apresentado na Apêndice A. Este modelo representa além de objetivos a serem alcançados, decisões a serem tomadas, tarefas a serem cumpridas e requisitos informacionais para contemplar essa demanda, o nível de importância de cada informação revelada pela entrevista possibilitou criar uma escala de prioridades informacionais (Endsley, 2003).

No Apêndice A é possível observar que as informações já se encontram separadas de acordo com os níveis de SAW, sendo que o nó descrito como "bombeiro de campo" possui informações do nível 1 de SAW que representa à percepção do humano tais como: reportar a central a real situação da emergência, identificar testemunhas e se possível o incendiário. Já no nível 2 da SAW que representa à compreensão as informações são: avaliar a situação e determinar a natureza da emergência.

O nó descrito como "comandante", as informações relacionadas à percepção (SAW nível 1), como: avaliar a situação e determinar a natureza da emergência, assim que recebe as informações do bombeiro de campo e de denúncias pelo aplicativo DF100Fogo. No nível de compreensão (SAW nível 2), as informações são relacionadas à alocação de recursos, a coordenação das equipes de atendimento e identificar o tipo de emergência, podendo ser ela um incêndio, um acidente ou um atendimento médico. Este nível também possui informações como clima, local e a gravidade da situação à serem consumidas e assimiladas pelo operador humano do sistema que monitora a emergência. Toda e qualquer emergência se trata de uma situação dinâmica, sendo assim, é importante que as informações sejam sempre atualizadas e de mais fácil entendimento possível. No terceiro nível de SAW, denominado projeção da situação, as informações remetem à evolução do incêndio, tais como: o terreno, direção de alastramento, tipos de matas, e etc. Com essas informações completas e coesas, é possível gerar um modelo de representação a ser consumido por sistemas de tomada de decisão.

Após essas etapas foi possível obter o insumo necessário para a construção de uma primeira versão da ontologia utilizando a metodologia 101 de Noy McGuinness (2001), com conceitos, relações e restrições do domínio em que a mesma será aplicada.

O Apêndice B ilustra a primeira versão da ontologia com instâncias que representam o seguinte relato de emergência: "Está acontecendo um incêndio no parque nacional próximo ao anfiteatro, as chamas estão com um metro e meio mais ou menos e a fumaça esta preta, tem um moço que tentou apagar as chamas e ficou com as mãos queimadas".

Junto a esse relato foram obtidas informações sobre a geolocalização do celular que enviou o alerta, a geolocalização do anfiteatro, o tipo de mata presente no local, se há neste local vias de acesso e quais são, também foi gerado informações sobre o clima como, umidade, velocidade do vento, temperatura e estação do ano.

Com essas informações e tendo como base o GDTA, foi possível instanciar a ontologia para essa ocorrência. Quanto às questões referentes à qualidade da informação, neste trabalho foi adotada a Metodologia para Avaliação da Qualidade de Dados e Informações no Contexto de

Consciência Situacional de Emergências - IQESA (Botega, et al. 2017),

Tal metodologia é baseada em três etapas, (1) Elucidação dos requisitos de qualidade para ajudar a definir os critérios (dimensões) a serem avaliados a cada transformação da informação; (2) Modelagem e Aplicação de Funções e Métricas para Quantificar Dimensões de Qualidade e (3) Representação da informação situacional qualificada. Neste contexto as dimensões de qualidade de dados utilizadas são: consistência, precisão, atualidade e completude, as quais receberam uma pontuação de 1 à 10 para identificar o nível de qualidade em cada instância.

5. Considerações Finais

Até o presente momento o GDTA encontra-se em fase de conclusão, sendo um dos próximos passos a reorganização do mesmo respeitando os níveis de SAW e com esse insumo será necessário atualizar a versão da ontologia atual.

A ontologia desenvolvida tem como objetivo auxiliar o sistema DF100Fogo (Saran et. al. 2017), através da representação de informações sobre a situação de emergência, da mesma forma ela irá ilustrar todas as entidades representadas na ocorrência e suas possíveis relações, inferindo assim novos dados sobre a situação, pois, relações não explicitas entre as entidades podem ser obtidas por meio de consultas a ontologia durante a situação de emergência, esse auxilio também se estenderá aos operados dos sistema contribuindo assim com a obtenção e melhora de sua SAW e permitindo uma tomada de decisão mais segura.

Será incorporado ao trabalho o Vocabulário de Qualidade de Dados (The Data Quality Vocabulary – DQV), o mesmo se trata de boas práticas com dados na Web levando em consideração o feedback de consumidores de dados, suas anotações, as políticas e os certificados de Agências de qualidade de dados.

Após essa etapa concluída, antes que a ontologia seja de fato incorporada ao sistema, serão desenvolvidos ensaios práticos com a ontologia utilizando consultas SPARQL, que se trata de uma linguagem de consulta semântica para banco de dados capaz de manipular e recuperar dados armazenados no formato de Framework de Descrição de Recursos (RDF).

Referências

BOTEGA, L.C.; OLIVEIRA, A. C. M.; PEREIRA JUNIOR, V. A.; SARAN, J. F.; VILLAS, L. A.; Araujo, R. B.. Quality-Aware Human-Driven Information Fusion Model. In: **20th International Conference on Information Fusion**, 2017, Xian. 20th International Conference on Information Fusion. 2017.

DQV – **Data on the Web Best Practices: Data Quality Vocabulary**. Disponível em: < https://www.w3.org/TR/vocab-dqv/#intro>. Acesso em 21 de julho de 2017.

ENDSLEY, Mica R. et al. Situation awareness oriented design: From user's cognitive requirements to creating effective supporting technologies. In**: Proceedings of the Human Factors and Ergonomics Society Annual Meeting**. Sage CA: Los Angeles, CA: SAGE Publications, 2003. p. 268-272.

KOKAR, Mieczyslaw M.; MATHEUS, Christopher J.; BACLAWSKI, Kenneth. Ontology-based situation awareness. **Information fusion**, v. 10, n. 1, p. 83-98, 2009.

NOY, Natalya F. et al. Ontology development 101: A guide to creating your first ontology. 2001.

SARAN, Jordan F. et al. Data and information fusion in the context of emergency management: The DF100Fogo project. In: **Information Systems and Technologies (CISTI)**, **2017 12th Iberian Conference on.** IEEE, 2017. p. 1-6.

Apêndice A – GDTA - Análise de Tarefas Dirigidas por Objetivos

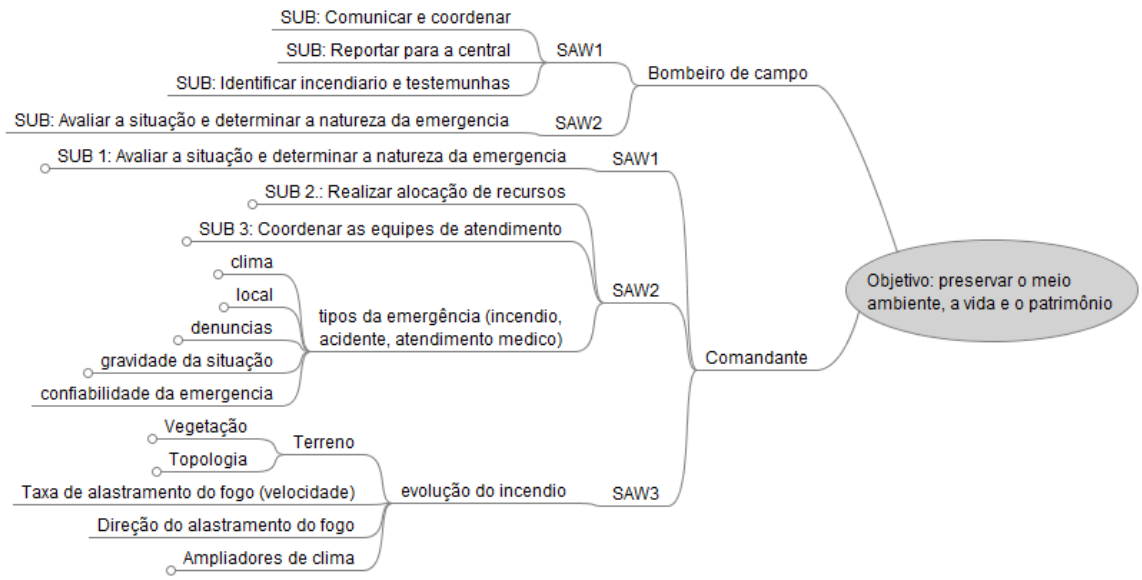


Figure 1: - Informações para o atendimento de uma emergência respeitando os níveis de SAW. Fonte: GIHC (2017).

Apêndice B – Ontologia

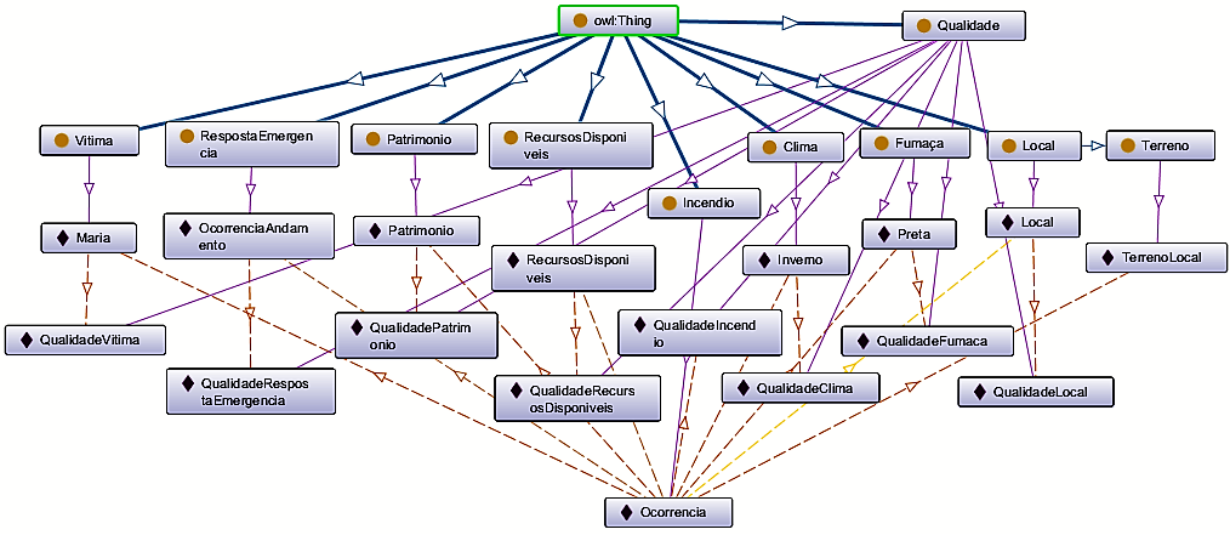


Figure 2: Ontologia para o domínio de incêndio. Fonte: GIHC (2017).

ANÁLISE QUANTITATIVA DE EVENTOS CRIMINAIS UTILIZANDO ABORDAGEM SEMÂNTICA

Quantitative Analysis of Criminal Events Using Semantic Approach

Gustavo Marttos Cáceres Pereira[1], João Henrique Martins[2], Leonardo Castro Botega[2]
(1) Grupo de Interação Humano-Computador, Centro Universitário Eurípides de Marília, Av. Hygino Muzzi, Marília – SP
marttos@univem.edu.br
(2) Stratelli – Inteligência Estratégica, Av. Hygino Muzzi, Marília – SP
joao.henriq.75@gmail.com, botega@univem.edu.br

Resumo:
O gerenciamento de informações de riscos utilizando dados criminais apresenta desafios associados à aquisição de Consciência Situacional, tais como a dinamicidade, heterogeneidade, variedade e o grande volume de dados. Além disso, representar adequadamente as informações no domínio criminal, tais como dados sobre vítimas, criminosos, locais e a própria situação de crime, pode contribuir para processos de análise quantitativa de dados, tornando assim a tomada de decisão ser mais assertiva. Ao mesmo tempo, há ainda a necessidade do desenvolvimento de modelos semânticos que representem a realidade das situações de ambientes de riscos, o que produz ainda mais desafios para a quantificação de entidades e características relevantes. Portanto, este trabalho tem por objetivo apresentar o processo de desenvolvimento de quantificação de dados em ontologias para o domínio de gerenciamento de riscos, visando suportar a extração de dados específicos para ampliar a capacidade de obtenção de Consciência Situacional e permitir que as decisões tomadas, baseadas nestas informações, sejam mais assertivas. Para tal, será empregada a análise de tarefas dirigidas por objetivos, a análise de vocabulários e propriedades no contexto criminal a fim de restringir sua interpretação e a avaliação de pesos e valores semânticos desses vocabulários. Sendo assim, ao final, será possível mensurar as informações necessárias dentro de um sistema de gerenciamento de riscos.
Palavras-chave: Consciência Situacional; Ontologia; Análise Quantitativa; Gerenciamento de Riscos

Abstract:
The management of risk information using criminal data presents challenges associated with the acquisition of Situational Awareness, such as dynamicity, heterogeneity, variety and large data volume. In addition, adequately representing information in the criminal domain, such as data on victims, criminals, places and the actual crime situation, can contribute to quantitative data analysis processes, thus making decision making more assertive. At the same time, there is still a need for the development of semantic models that represent the reality of situations in risk environments, which produces still more challenges for the quantification of entities and relevant characteristics. Therefore, this paper aims to present the process of a development of data quantification in ontologies for the domain of risk management, aiming at support the extraction of specific data to increase the capacity to obtain Situational Awareness and allow the decisions taken to be more assertive. For this, the analysis of tasks directed by objectives, the analysis of vocabularies and properties in the criminal context will be used in order to restrict their interpretation and the evaluation of semantic weights and values of these vocabularies. Therefore, in the end, it will be possible to measure the necessary information within a risk management system.
Keywords: Situational Awareness; Ontology; Quantitative Analysis; Risks Management

1. Introdução

Consciência Situacional (*Situational Awareness* - SAW) é um conceito fundamental para auxiliar a tomada de decisão em ambientes complexos e dinâmicos em uma variedade de domínios, entre eles o de gerenciamento de riscos (BOSSÉ; ROY; WARK, 2007).

SAW é definida como a percepção dos elementos no ambiente dentro de um volume de tempo e espaço, a compreensão do seu significado e a projeção de seu status em um futuro próximo (ENDSLEY, 1998).

Neste contexto, operadores de sistemas de gerenciamento de riscos, como os que lidam com dados criminais, e que estão constantemente sob alta pressão e expostos a uma gama de informações sensíveis, precisam manter seus níveis de SAW elevados para assim sustentar o melhor

retrato de uma situação crítica e tomar a decisão mais assertiva, evitando prejuízos à vida, ao patrimônio e também ao meio ambiente.

Por ser um estado cognitivo do operador, a SAW não garante que este tomará a melhor decisão, entretanto garante melhores subsídios para que possa melhorá-la.

Para suportar a obtenção e manutenção de SAW, foi constatado que o uso de modelos semânticos, mais especificamente ontologias, quando aplicados para suportar sistemas de gerenciamento de riscos, podem contribuir para uma melhor assertividade nas inferências úteis à tomada de decisão (MATHEUS; KOKAR; BACLAWSKI, 2003).

A tendência dos sistemas dinâmicos é utilizar ontologias ou outros modelos semânticos para que os dados sejam representados. Entretanto, esta ação gera novos desafios para quantificar os dados necessários para que assim sejam transformados em informações que apoiem à tomada de decisão dos operadores de sistemas de gerenciamento de riscos.

Fluit *et al.* (2006) apontaram a dificuldade dos operadores em fornecer termos que melhor descrevam as suas necessidades de informação. Esta dificuldade aumenta quando expressões lógicas simples são utilizadas, o que tornou o processo de quantificação mais complicado.

A quantificação é fundamental para a estruturação de um processo de análise crítica sobre os dados.

A partir dos dados oferecidos pelo sistema, deseja-se extrair tipos específicos de informações e quantificá-las para que sejam armazenadas e posteriormente analisadas por meio de métodos estatísticos, obtendo outras informações geradas a partir de dados preditivos dentro de um domínio. Sagion *et al.* (2007) afirmam que informações quantitativas (numéricas) são necessárias em sistemas dinâmicos e inteligentes.

Quando utilizada uma ontologia em um sistema inteligente, passa-se a ter mais qualificação dos dados, em vez de quantificação, tornando desejável a aplicação de métodos que supram essa necessidade.

Este trabalho tem como objetivo apresentar o desenvolvimento de um processo de quantificação de informações de riscos, organizadas e representadas semanticamente em ontologias de domínio, mais especificamente no contexto criminal, com o principal objetivo de contribuir com a aquisição de SAW.

2. Gerenciamento de Riscos e SAW

Endsley (1995) afirma que dentro de ambientes de avaliação de riscos, é grande a necessidade de tomadas de decisões precisas e as tarefas dependem de análises contínuas e que estejam sempre atualizadas.

Um exemplo corriqueiro de SAW é um motorista de carro, o qual precisa estar ciente de todas as ações tomadas, como quando acelerar e frear, quando deve parar no semáforo e principalmente estar atento aos pedestres e carros que estejam ao seu redor, observando as condições do tráfego para que assim ele possa escolher o melhor trajeto, se possível.

A complexidade e a dinâmica de um ambiente de avaliação de riscos aumentam proporcionalmente conforme a quantidade de variáveis existentes nesse ambiente, tornando a aquisição e manutenção de SAW processos mais difíceis.

De acordo com o Modelo de Consciência Situacional proposto por Endsley (1988), existem três níveis os quais os operadores podem atingi-los conforme o seu desenvolvimento dentro do ambiente, que são influenciados direta e indiretamente por fatores internos e externos: (a) nível um: perceber a dinâmica, o estado e os atributos de elementos pertinentes ao ambiente; (b) nível dois: compreender a situação atual, criando relações entre os elementos, desenvolvendo um contexto de acordo com os objetivos e metas esperados, priorizando elementos importantes e o que estas informações combinadas representam; e (c) nível três: projetar estados futuros, tendo ciência do que os elementos são e o que significam perante a situação atual, sendo necessário ter a habilidade de prever quais serão as ações futuras.

É possível qualificar a SAW de um operador por meio de suas habilidades, preconceitos, experiências, metas e objetivos, entretanto até os mais qualificados podem tomar decisões errôneas caso sua SAW não esteja nos níveis adequados.

Em ambientes emergenciais - cujas falhas podem provocar perdas de vidas e danos ao meio ambiente e ao patrimônio público - é necessário adquirir, manter e melhorar a SAW constantemente e para isso é preciso que existam sistemas que a apoiem.

Os operadores precisam agir de forma decisiva dentro de intervalos de tempo curtos com informações muitas vezes incompletas e/ou com muitos dados dos quais é difícil extrair informações relevantes.

Os sistemas inteligentes devem fornecer informações precisas, de forma rápida e em um formato correto aos operadores. Eles devem ser elaborados de modo a auxiliarem na tomada de decisão mais assertiva pelo operador, considerando o tempo de resposta como fator limitador.

Como exemplo de ambiente emergencial, é possível utilizar o domínio militar, constatando o uso de SAW e a aplicação de seus níveis da seguinte forma: (1) percepção do operador sobre o posicionamento e condicionamento físico de sua equipe, armamento disponível e a quantidade de tropas inimigas; (2) compreensão sobre o avanço de tropas inimigas, possibilidade de deslocamento ou combate; e (3) visão de um futuro próximo em relação à tropa do inimigo, se sua atual posição representa uma ameaça ou risco iminente à equipe.

3. Ontologia como Modelo de Apoio a SAW

Ontologia é a forma como a lógica explica o significado pretendido de um vocabulário formal, ou seja, seu compromisso ontológico com uma conceitualização particular do mundo (GUARINO, 1998).

O seu uso possibilita construir relações organizadas entre diversos termos de um domínio específico, tornando possível a contextualização dos dados e auxiliando no processo de interpretação.

Kokar, Matheus e Baclawski (2009) afirmam que a diferença da obtenção e manutenção de SAW entre humanos e computadores é que no primeiro é possível mensurar, enquanto no segundo é preciso definir e implementar este processo.

As ontologias tornam-se de fato computacionais a partir do momento que elas são implementadas, algo além de conceitos, viabilizando o desenvolvimento de aplicações mais robustas e inteligentes.

Uma ontologia é valiosa somente se a maioria da comunidade envolvida em seu uso aceitar seus principais conceitos e estrutura, por isso é necessária uma definição formalizada (KOKAR; MATHEUS; BACLAWSKI, 2009).

Para utilizar SAW, especialmente para tomada de decisões, é preciso reconhecer diversas situações, avaliar o impacto sobre os objetivos, relacionar propriedades a situações particulares e comunicar as descrições destas às demais pessoas. Com isto, existem dois requisitos adicionais em relação às representações situacionais: (1) elas podem ser classificadas por tipos de situação, e (2) elas podem ser tratadas como objetos físicos ou conceituais.

Dois aspectos importantes que não podem deixar de ser considerados são os de atributos e relacionamentos, os quais devem ser associados com valores que possam mudar ao longo do tempo. Devido às várias formas que as situações podem ser abordadas, principalmente aquelas caracterizadas como emergenciais, estas devem exigir robustez e solidez da solução de SAW.

4. Metodologia

Sistemas que buscam apoiar a avaliação e análise de situações devem envolver um esforço para estimular a SAW em operadores, o mais próximo possível da realidade, uma vez que erros de SAW podem comprometer a vida ou o patrimônio. Adquirir SAW ainda é algo desafiador, frente ao grande volume, variedade e dinamicidade de dados.

Para Botega (2016), esta heterogeneidade característica dos ambientes emergenciais demandam que os dados sejam processados e representados de forma semântica, utilizando ontologias, em busca de uma informação situacional significativa.

Dentro do domínio criminal, por meio do processo de quantificação semântica, deve ser possível mensurar quais os locais de

origens mais recorrentes das denúncias, qual o horário de pico de ocorrências, a atualidade das denúncias (horário atual contra o horário da denúncia), quais as características recorrentes de vítimas, criminosos, objetos e locais.

O processo de quantificação semântica envolve o estudo de metodologias de mensuração semântica, entre eles o de Pesos e Valores Semânticos (apresentada a seguir), e a avaliação de vocabulário de integração e associação de relacionamentos semânticos, por meio da compreensão do domínio de gerenciamento de risco.

Para identificar as necessidades dos operadores foi empregada a Análise de Tarefas Dirigida por Objetivos (*Goal-Driven Task Analysis* – GDTA), uma metodologia útil para revelar as informações necessárias para que operadores humanos tomem decisões ao longo do uso de um sistema de avaliação de situações de risco, bem como as tarefas que os mesmos devem realizar. Saber as informações demandadas por cada tarefa é o primeiro passo para o processo de quantificação, pois é necessário ter conhecimento daquilo que se deseja mensurar, representando semanticamente dados que façam sentido. Parte do GDTA é apresentada no Apêndice A, a qual apresenta um de seus objetivos, que é a caracterização das quatro entidades de um crime de roubo (vítimas, criminosos, objetos e locais).

O estudo de vocabulários e propriedades utilizadas no contexto criminal é uma etapa fundamental do trabalho, pois se torna possível a descrição de uma realidade específica, restringindo sua interpretação. Estes serão criados somente após o entendimento do domínio, constituindo a semântica e conceitos. Compreendê-los como um todo proporciona um conhecimento amplo das características e propriedades de suas classes. Kokar, Matheus e Baclawski (2003) afirmam que quanto mais extenso for o vocabulário, melhor poderá ser a descrição dos relacionamentos que acontecem dentro do domínio.

A ontologia em desenvolvimento pode ser vista no Apêndice B, a qual contempla as classes principais que caracterizam um crime de roubo e suas respectivas subclasses e relacionamentos.

Antes de quantificar os resultados por meio da representação do modelo semântico, é necessário que exista um ou mais dados, bem estruturados, cujos significados sejam relevantes.

Por fim, a quantificação em si é baseada na metodologia de Peso e Valor Semântico, proposta por Hepp *et al.* (2006), a qual leva em consideração o grau de especificidade do vocabulário e propriedades. Ela é baseada na fundamentação de que uma propriedade, quando muito utilizada, é geralmente menos específica do que uma propriedade pouco utilizada.

O primeiro passo desta metodologia é dar ao atributo do vocabulário um peso semântico, em seguida o valor semântico é calculado somando os pesos semânticos de todos os atributos. Caso uma classe não possua atributos, o valor semântico é igual zero.

Cada atributo tem realizada sua contagem de entradas entre a relação atributo e classe, portanto, segundo Hepp (2004), ele recebe um peso semântico que é igual ao seu valor recíproco de sua frequência de uso.

5. Resultados Esperados

Até o presente momento, o GDTA já foi elaborado e aplicado junto aos operadores da PMESP (Polícia Militar do Estado de São Paulo). A ontologia de domínio está em fase de desenvolvimento, embasada nas respostas obtidas do questionário do GDTA.

É esperado que com a ontologia concluída apoiando à SAW, seja possível desenvolver o processo de quantificação semântica voltado para o domínio de sua aplicação, no caso o de gerenciamento de riscos, possibilitando mensurar os relacionamentos, similaridades e atributos das classes. Uma vez mensurados, os dados podem se tornar informações valiosas que auxiliem na tomada de decisões de operadores.

4 Considerações Finais

Ao se utilizar um modelo semântico apoiado em SAW em um sistema dinâmico, é possível qualificar os dados, atribuindo um

significado ao contexto em que se encontram, porém perde-se o necessário para a quantificação dos dados: expressões lógicas e numéricas.

Este trabalho está em fase de desenvolvimento, principalmente a ontologia de domínio, a qual está representada no Apêndice B. Concluída, será possível representar semanticamente as informações geradas a partir de dados quantificados, servindo como suporte aos operadores de sistemas de gerenciamento de riscos com o objetivo de contribuir com a melhoria de sua SAW.

Para os próximos passos os RDFs (*Resource Description Framework* – modelo de representação de ontologia) serão gerados e analisados utilizando consultas SPARQL (*SPARQL Protocol and RDF Query Language* – linguagem de consulta semântica de dados), atribuindo os pesos semânticos e recuperando a quantificação no contexto desejado.

Referências

BOSSÉ, É.; ROY, J.; WARK, S. **Concepts, Models, and Tools for Information Fusion**. [S.l.]: Artech House, Incorporated, 2007. (Artech House intelligence and information operations library). ISBN 9781596930810.

BOTEGA, L. C. **Modelo de Fusão Dirigido por Humanos e Ciente de Qualidade de Informação**. 247 p. Tese (Doutorado) - UFSCar - Universidade Federal de São Carlos, 2016.

ENDSLEY, M. R. **Toward a theory of situation awareness in dynamic systems**. Human Factors: The Journal of the Human Factors and Ergonomics Society, SAGE Publications, v. 37, n. 1, p. 32–64, 1995.

ENDSLEY, M. R. **Design and evaluation for situation awareness enhancement**. In: SAGE PUBLICATIONS. Proceedings of the human factors and ergonomics society annual meeting. [S.l.], 1988. v. 32, n. 2, p. 97–101.

FLUIT, Christiaan; SABOU, Marta; VAN HARMELEN, Frank. **Ontology-based information visualization: toward semantic web applications**. In: Visualizing the semantic web. Springer London, 2006. p. 45-58.

GUARINO, N. **Formal ontology and information systems**. In: Proceedings of FOIS.
[S.l.: s.n.], 1998. v. 98, n. 1998, p. 81–97.

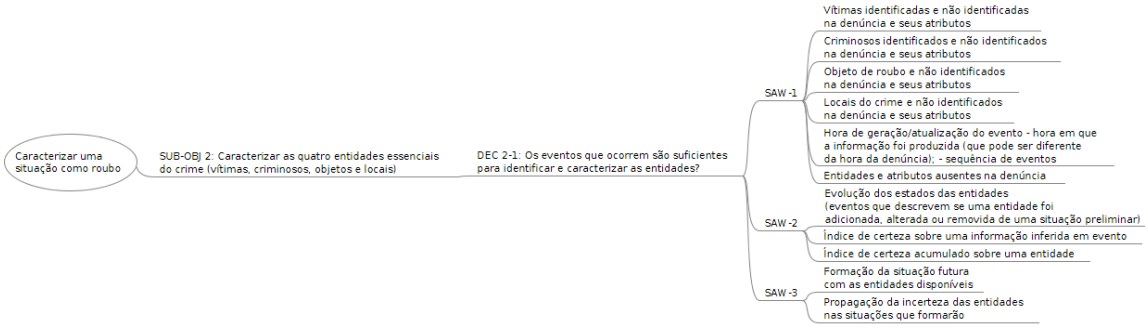
HEPP, Martin. **Measuring the Quality of Descriptive Languages for Products and Services**. In: MKWI (E-Business). 2004. p. 157-168.

HEPP, Martin; LEUKEL, Joerg; SCHMITZ, Volker. **A quantitative analysis of product categorization standards**: content, coverage, and maintenance of eCl@ ss, UNSPSC, eOTD, and the RosettaNet Technical Dictionary. Knowledge and Information Systems, v. 13, n. 1, p. 77-114, 2007.

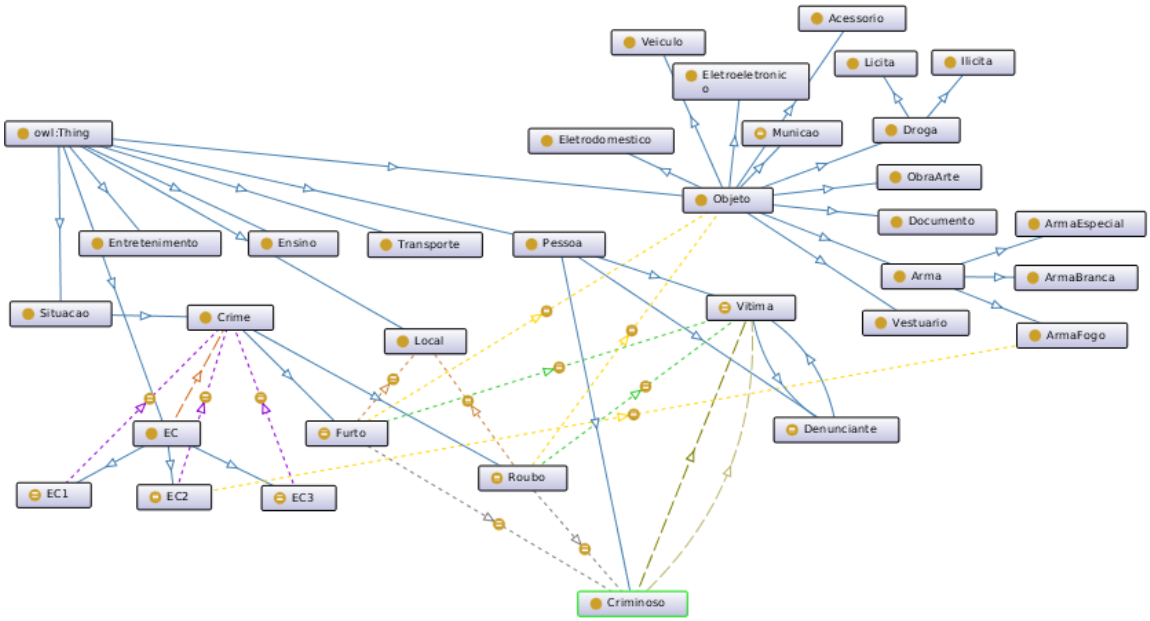
KOKAR, Mieczyslaw M.; MATHEUS, Christopher J.; BACLAWSKI, Kenneth. **Ontology-based situation awareness**. Information fusion, v. 10, n. 1, p. 83-98, 2009.

MATHEUS, C. J.; KOKAR, M. M.; BACLAWSKI, K. **A core ontology for situation awareness**. In: Proceedings of the Sixth International Conference on Information Fusion. [S.l.: s.n.], 2003. v. 1, p. 545–552.

Apêndice A – Análise de Tarefas Dirigida Por Objetivos (*Goal-Driven Task Analysis* – GDTA)



Apêndice B – Ontologia de Domínio Desenvolvida



O PAPEL ESTRATÉGICO DA WEB SEMÂNTICA NO CONTEXTO DO BIG DATA

The Strategic Role of Semantic Web in the Big Data Context

Caio Saraiva Coneglian[1], Rodrigo Dieger[1], José Eduardo Santarém Segundo[2], Miriam Captrez[3]

(1) Universidade Estadual Paulista (UNESP), Av. Hygino Muzzi Filho, 737, Mirante, Marília - SP, 17.525-900, {caio.coneglian, rdieger}@gmail.com,

(2) Universidade de São Paulo (USP), Av. Bandeirantes, 3900 - Vila Monte Alegre, Ribeirão Preto - SP, 14040-900, santarem@usp.br

(3) Western University, 1151 Richmond St, London, ON N6A 3K7, Canadá, mcapretz@uwo.ca

Resumo:
A Web Semântica apresenta um corpus teórico e diversas tecnologias e aplicações que demonstram a sua consistência, inclusive no que tange ao uso de seus conceitos e de suas tecnologias em outros escopos não se limitando unicamente a Web. Neste sentido, os projetos de Big Data podem tirar proveito da aplicação dos princípios e dos desenvolvimentos realizados na área da Web Semântica, para aperfeiçoar os processos de análises de dados, em especial na inserção de características semânticas para contextualização dos dados. Assim, esta pesquisa tem como objetivo analisar e discutir o potencial das tecnologias da Web Semântica como meio de integração e desenvolvimento de aplicações de Big Data. Utilizou-se uma metodologia qualitativa exploratória, onde buscou-se pontos de convergência entre a Web Semântica e Big Data. Foram identificados e discutidos três pontos principais: a aplicação do Linked Data enquanto fonte de dados para o Big Data; o uso de ontologias nas análises de dados; e o uso das tecnologias da Web Semântica para promoção da interoperabilidade em cenários de Big Data. Neste sentido, foi possível identificar que a Web Semântica, em especial no que permeia suas tecnologias e aplicações, pode auxiliar significativamente o desenvolvimento do Big Data, por fornecer um paradigma complementar dos aplicados majoritariamente nas análises de dados.
Palavras-chave: Web Semântica; Big Data; Tecnologias da Web Semântica.

Abstract:
The Semantic Web presents a theoretical corpus and a range of technologies and applications that demonstrate its consistency, including in use of its concepts and its technologies in other scopes than the Web. In this sense, Big Data's projects can take advantage of the application of principles and developments in the area of the Semantic Web, to improve the processes of data analysis, especially in the insertion of semantic characteristics for data contextualization. Thus, this research aims to analyze and discuss the potential of Semantic Web technologies as a means of integrating and developing Big Data applications. An exploratory qualitative methodology was used, where we searched for points of the literature and documentary texts dealt with the convergence between the Semantic Web and the Big Data. Three main points were identified and discussed: the application of Linked Data as a data source for Big Data; the use of ontologies in data analysis; the use of Semantic Web technologies to promote interoperability in Big Data scenarios. Therefore, it was possible to identify that the Semantic Web, especially with regard to its technologies, can help Big Data, since it provides a paradigm different from those applied mainly in data analysis.
Keywords: Semantic Web; Big Data; Semantic Web Technologies.

1 Introdução

Vive-se a era do *Big Data*. O intenso processo de evolução e utilização das tecnologias computacionais e informacionais que se tem vivenciado nos últimos anos, vem acelerando de maneira radical a expansão e integração dos mais variados dispositivos e ambientes informacionais digitais, impactando a forma como estão sendo criados e utilizados os dados e as informações oriundas destes contextos.

A geração e consumo de dados vem se tornando uma parte importante da vida diária de pessoas e das organizações em geral, particularmente com a disponibilidade e uso massificados da tecnologia e aplicações da Internet. Zikopoulos, Eaton e Deroos (2012) definem que a era do *Big Data* é resultado das mudanças que tem ocorrido no mundo, onde por meio dos avanços das tecnologias, foi possível que pessoas e programas se intercomunicassem durante todo o tempo.

Em decorrência deste novo paradigma, observa-se um aumento exponencial no volume, na variedade (fontes, formatos e esquemas distintos) e na velocidade com que dados e informações vem sendo criados e disponibilizados. Estudo publicado pelo

International Data Corp (IDC), prevê que a criação de dados aumentará para cerca de 163 zettabytes (ZB) até 2025, um aumento de dez vezes nos valores de 2016, e também considera que a coleta, gerenciamento e análise de dados sejam a força motriz por trás de quase todas as atividades humanas na próxima década (GANTZ e REINSEL, 2017).

Esse rápido e contínuo crescimento, somado às limitações dos métodos e formas tradicionais de análise e processamento (levando-se em conta suas características), apresentam inúmeros desafios relacionados à maneira como tornar estes dados e informações disponíveis para uso de maneira efetiva. Beyer e Laney (2012) definem *Big Data* como o alto volume, alta velocidade e/ou alta variedade de informações que requerem novas formas de processamento para permitir melhor tomada de decisão, nova descoberta do conhecimento e otimização de processos.

Apesar de provenientes de uma direção diferente ao *Big Data*, os conceitos e as tecnologias da Web Semântica permitem reunir fontes heterogêneas de dados para explorar e fornecer significado a diferentes conjuntos, facilitando a aplicação do processamento semântico.

A partir da interoperabilidade de tecnologias e conceitos desses diferentes campos, permite-se um novo processo de descoberta de conhecimento, agrupando e organizando a informação disponível de maneira eficiente e integrada, permitindo dessa forma que se explore, analise, processe e transforme dados a partir de fontes distintas.

Diante deste cenário, o objetivo deste artigo é analisar e discutir o potencial das tecnologias da Web Semântica como meio de integração e desenvolvimento de aplicações de *Big Data*. Além disso, procura demonstrar os principais desafios da integração de dados relacionados com este tema.

O texto foi organizado com uma introdução, seguido de uma seção tratando dos pressupostos teóricos tanto de *Big Data* quanto de Web Semântica. Em seguida, são apresentados procedimentos metodológicos do trabalho, finalizando com os resultados e discussões e as considerações finais.

2 *Big Data* e Web Semântica: além das fronteiras da Web

Nos últimos anos, podemos observar de maneira significativa o avanço exponencial no número de pesquisas e aplicações que desenvolvem e exploram os conceitos relacionados a *Big Data*. Laney (2001), em uma das primeiras definições sobre este tema, afirma que o Big Data se caracteriza essencialmente a partir de três aspectos: volume, velocidade, variedade. Volume está estritamente relacionado ao tamanho e quantidade de dados. Velocidade refere-se a aspectos da dinâmica de crescimento e processamento dos dados. Variedade à diversidade de origens, formas e formatos dos dados (DEMCHENKO et al., 2013).

A propagação e disseminação de dados oriundos das redes sociais, comunicação entre máquinas, sensores, bem como a análise e aproveitamento de artefatos digitais e bases de dados existentes, ou ainda tecnologias emergentes como a "Internet das Coisas" e o fenômeno dos dados abertos, produzem-se em larga escala e tornam praticamente qualquer coisa como dado ou conteúdo, que precisam ser cada vez mais bem interpretados e examinados. No entanto, a maioria desses dados é ainda inacessível, pois precisamos de tecnologia e ferramentas para encontrar, transformar, analisar e visualizar dados para torná-los consumíveis para a tomada de decisões (BANSAL, 2014).

Neste sentido, questões que permeiam o significado dos dados desempenham um papel fundamental no que se refere ao uso efetivo e ao aproveitamento das informações e do conhecimento extraídos. Para enfrentar esses desafios, as tecnologias e os conceitos de diferentes campos podem ser combinados, permitindo um avançado processo de descoberta de conhecimento.

Quando direcionamos nossa abordagem para o significado dos dados, os conceitos e as tecnologias da Web Semântica se apresentam de maneira proeminente e definem um componente estratégico para a tratativa da variedade de dados no cenário do *Big Data*.

O conceito da Web Semântica foi concebido a partir de 2001, apontando uma Web na qual os computadores poderiam entender o contexto das pessoas, para poder

interpretar o significado da informação (BERNERS-LEE; HENDLER; LASSILA, 2001). As tecnologias da Web Semântica permitem que se criem repositórios de dados, se construam vocabulários e se estabeleçam regras para definição e representação dos dados na Web, mas não se limitando a ela. Além disso, apresenta conceitos e tecnologias para representar conhecimento e suas relações, utilizando uma série padrões e ainda um conjunto de melhores práticas para a publicação de dados estruturados no *Linked Data* (BERNERS-LEE, 2006).

Esses padrões semânticos possuem recursos compatíveis com as necessidades de dados existentes e estrito alinhamento com o *Big Data*. Características que de maneira geral refletem sobre representação do conhecimento, interoperabilidade de dados, e recuperação da informação também definem um importante aspecto neste contexto para resolver questões relacionadas com análise e a variedade de dados.

Acredita-se que o desafio técnico mais importante hoje na gestão de *Big Data* é o aspecto da variedade (heterogeneidade de dados e diversidade das fontes de dados). Para tratar a heterogeneidade, a abordagem semântica é a que melhor se apresenta para resolver estas problemáticas. Para entender, relacionar e interpretar dados, é necessário o significado explícito dos dados, que é dado pelo aproveitamento efetivo das tecnologias e abordagens semânticas.

Ao estudar diversas literaturas sobre tecnologias da Web Semântica e Big Data, identifica-se que estas desempenham um papel importante para converter dados em conhecimento. Em comparação com outras tecnologias, as tecnologias semânticas fornecem conhecimento prévio para o contexto dos dados, interoperabilidade, escalabilidade, integração e aceitos como padrão de expressividade de dados.

3 Procedimentos Metodológicos

Para atingir os objetivos deste trabalho, utilizou-se uma metodologia qualitativa exploratória, onde buscou-se pontos em que a literatura e textos documentais tratavam da convergência entre as tecnologias e os conceitos da Web Semântica e os processos que tangenciam o *Big Data*.

Para realizar a pesquisa, identificou-se primeiramente temáticas de estudos em que há essa relação iminente da aplicação das tecnologias da Web Semântica no cenário do *Big Data*. Posteriormente, foi realizado uma explanação sobre cada um dos pontos identificados, apontando como ocorre o uso das tecnologias da Web Semântica, além de verificar como esta utilização contribui para os processos de *Big Data* como um todo.

4 Resultados e Discussões

A partir dos procedimentos apontados, identifica-se cenários em que a aplicação da Web Semântica pode ocorrer no âmbito do *Big Data*. Passando desde os pontos relativos às próprias fontes de informações, até na inserção de um número maior de argumentos nas análises de dados, a Web Semântica, juntamente com alguns de seus conceitos, tecnologias e aplicações pode trazer semântica e contextualização nos processos que se relacionam ao *Big Data*.

Neste contexto, na sequência busca-se apresentar os principais pontos em que a Web Semântica pode denotar um papel estratégico e de grande relevância principalmente para a tratativa da variedade e a descoberta de novas relações e padrões entre os grandes volumes de dados que se apresentam em um cenário de *Big Data*.

4.1 *Linked Data*: conectando o *Big Data*

O meio como as informações estão estruturadas em cenários de Big Data é significativamente distinto daqueles conjuntos de dados estruturados seguindo os princípios do *Linked Data*. Em suma, a maioria dos dados tratados como *Big Data* são desestruturados ou semi-estruturados, enquanto na perspectiva do *Linked Data,* são integralmente estruturados.

A diferença entre estes dois cenários é acentuada pela existência de metadados que apontem o contexto e o significado que os conjuntos de dados estabelecem dentro do *Linked Data*, e que de modo geral não se refletem no contexto do *Big Data*. Desta forma, os dados de *Linked D*ata tornam-se uma importante fonte de informação, ao fornecer dados estruturados e com semântica formal, tratando de um domínio específico.

No entanto, o *Linked Data* contempla um escopo limitado de conjuntos de dados, que

foi tratado e enriquecido a partir de procedimentos computacionais em ambientes minimamente controlados, tendo assim, função e princípios diferentes do *Big Data*, que irá contemplar dados das mais variadas fontes, sem apresentar um rígido controle sobre a estrutura destes dados. Assim, o *Linked Data* não pode ser utilizado como um substituto das fontes informacionais de grande volume do *Big Data*, mas sim um elemento complementar nos processos de análises de dados.

Há diversas correntes defendidas sobre os métodos utilizados durante os processos de análises de dados, que irão apresentar os pontos que devem ser considerados, bem como as fases aplicadas para a análise. Um destes autores é Bugembe (2016), que divide em seis o que ele chama de fases para obtenção de valor dos dados durante as análises: 1) fonte; 2) captura e armazenamento; 3) processamento e fusão, 4) acesso; 5) análise; e 6) exposição.

O autor, ao discutir essas diversas fontes, vai inserindo fase a fase como deve ser realizada a coleta dos dados, as preocupações quanto a escolha das fontes, o processamento, a análise, entre outros. Desta forma, identifica-se sempre a busca por relacionar informações relevantes e que possam de alguma forma possuir confiabilidade. Neste sentido, o *Linked Data* se mostra como uma fonte auxiliar aos dados, capaz de fornecer aos processos subsequentes uma maior confiabilidade, além de permitir que as relações realizadas nos processos de fusão, ocorram com um número maior de argumentos, permitindo ainda que fontes relacionadas sejam incluídas e utilizadas durante o processo.

Em síntese, o *Linked Data* traria dados estruturados e semanticamente formalizados ao processo de análise, permitindo com que a exploração dos dados brutos (não estruturados e semiestruturados) na busca de extrair *insights* e padrões comportamentais, seja aprimorado ao considerar uma fonte que permita contextualizar e conduzir a realização de inferências com um nível lógico mais profundo nesta integração entre o *Linked Data* e os demais dados. Um instrumento que contribui para o Linked Data e que pode aprimorar nos processos de Big Data são as ontologias, exploradas na sequência.

4.2 Ontologias como estratégia para a análise e organização do conhecimento

As ontologias são instrumentos centrais para a Web Semântica por representarem formalmente um determinado domínio, explicitando axiomas nas relações existentes entre os recursos. Essa característica discutida por Santarem Segundo e Coneglian (2016), demonstra o potencial computacional que as ontologias possuem ao representar um determinado domínio, promovendo a realização de inferências quando se usa as ontologias na descoberta de informações.

Desta forma, o uso de ontologias pode ocorrer em diversas etapas das análises de dados em cenários de *Big Data*, por possibilitar um nível de semântica formal essencial nos processos que visam extrair valor dos dados.

Um possível uso das ontologias neste contexto, se caracteriza pela necessidade de pesquisadores da área de Big Data explorarem o poder das correlações estatísticas ao analisar grandes conjuntos de dados que podem estar relacionados, e assim extrair algum valor destas massas de dados. Mayer-Schönberger e Cukier (2013) afirmam que: “Previsões com base em correlações estão na essência do Big Data”, o que demonstra como as teorias lógicas, matemáticas e estatísticas auxiliam significativamente na tomada de decisão dos gestores ao analisar os dados.

Neste sentido, as ontologias por serem um aparato tecnológico capaz de expressar um domínio com lógicas, e com capacidade representacional que permite a realização de inferências, podem trazer um suporte significativo nestes processos que estão inter-relacionando bases de dados, e assim permitindo a realização de predições.

Pereira Junior et al. (2016, p. 103, tradução nossa) discorre sobre a possibilidade do uso de ontologias para a fusão de informação, afirmando que os processos tradicionais de fusão são baseados unicamente na sintaxe, ao invés do significado dos termos, enquanto a fusão semântica com ontologias “[...] permite gerar informações com qualidade aprimorada e mais fiel ao ambiente real”.

Diante desses pontos, o uso das ontologias na fusão de dados surge como um meio de tornar os resultados desse processo computacional mais aprimorado e eficiente, trazendo aos processos de Big Data a inserção da semântica e do contexto na análise em si. Tal questão se mostra como um contraponto aos métodos de análises que se focam unicamente nas relações estatísticas e matemáticas dos dados, que não deixam de ter valor, mas passam a ser complementadas por uma análise mais profunda do contexto que os dados se encontram.

Uma consequência da adoção de ontologias para a realização das chamadas fusões de informações semânticas, seria a possibilidade de tornar o processo de análise, discutido por Bugembe (2016) mais aprimorado, por ter um instrumento informacional que embasa a realização da fusão e possibilita inferências nesta fase de análise, a partir dos axiomas das propriedades das ontologias. Outro ponto promovido pelas ontologias trata da interoperabilidade, que se mostra como um outro ponto essencial para o *Big Data* e que pode ser aprimorado a partir dos conceitos e das tecnologias da Web Semântica.

4.3 *Big Data* e os desafios da interoperabilidade semântica dos dados

Interoperabilidade de dados pode ser contextualizada a partir da capacidade fornecida aos sistemas para interpretar de maneira automática e precisa o significado dos dados trocados. Para alcançar a interoperabilidade de dados semânticos, os sistemas não precisam apenas trocar seus dados, mas também trocar ou concordar com modelos explícitos desses dados (HARMELEN, 2008).

No contexto de *Big Data*, dados oriundos de fontes não estruturadas e heterogêneas se estabelecem como uma de suas principais características. Alcançar a interoperabilidade semântica nestes casos pode ser considerado um grande problema, visto principalmente a variedade de características e particularidades de cada fonte de dados observadas a partir deste cenário.

As tecnologias e os conceitos da Web Semântica permitem aplicar enriquecimento semântico aos dados por meio do uso de vocabulários específicos, ontologias e padrões de metadados. Além disso, outra vantagem apresentada por este modelo fundamenta-se no fato de ser um padrão estabelecido para que os dados sejam lidos e interpretados a partir de agentes computacionais, promovendo uma autonomia e independência para os sistemas que fazem uso efetivo dos dados concebidos a partir deste modelo, permitindo reduzir o custo e a complexidade da integração de dados.

As soluções atuais de processamento e armazenamento e recuperação de dados heterogêneos e distribuídos no contexto do *Big Data*, oferecem níveis de escalabilidade, robustez, tolerância a falhas e elasticidade sem precedentes. No entanto, não é possível compartilhar o potencial das tecnologias da Web Semântica em grande parte dessas soluções, visto que os valores atribuídos aos dados normalmente não possuem uma anotação semântica explícita. Assim, a possibilidade de combinar dados não estruturados em grande escala com dados estruturados e tecnologias da Web Semântica, expande as oportunidades em Big Data de processar dados de novas formas e combinações.

Victorino et al. (2017) aponta uma proposta de um ecossistema de Big Data para análises de dados abertos governamentais, em que tecnologias da Web Semântica, como ontologias, dão suporte a realização de interoperabilidade e de processos analíticos dos dados abertos. Este trabalho demonstra como as tecnologias da Web Semântica podem contribuir efetivamente, estando integrado com as principais ferramentas de Big Data existentes.

Padrões da Web Semântica e *Linked Data* como o RDF (*Resource Description Framework*), que conforme define a W3C (2004), tem como um dos principais objetivos criar uma rede de informações a partir de dados distribuídos, e o protocolo SPARQL (*Simple Protocol and RDF Query Language*) para recuperação da informação em ambientes semânticos, destacam-se como exemplos concretos na direção de oportunidades e alternativas estratégicas para a problemática da interoperabilidade de dados semânticos na era do *Big Data*.

5 Considerações Finais

A Web Semântica a partir da sua concepção original em 2001, vem evoluindo significativamente, em especial no que tangencia a criação de conceitos e de tecnologias que possam promover os princípios idealizados por seus criadores. Diante dessa evolução, a Web Semântica transcendeu as barreiras da própria Web, fornecendo instrumentos que auxiliam instrumentos computacionais nos mais diversos âmbitos, inclusive em bases de dados privadas e corporativas. Isso se estabeleceu principalmente pela forma como a Web Semântica passou a conceber o tratamento dos dados, contribuindo com instrumentos que favorecem a contextualização em um determinado domínio.

Um cenário que se apresentou como expoente na utilização das contribuições da Web Semântica ao fornecer meios para a realização de inferências e de lógicas e processos para descoberta de conhecimento, foi o Big Data, em especial para a realização de análises de dados que se enquadram neste contexto. Essa união entre os processos do *Big Data* com as tecnologias da Web Semântica pode ser estratégica e fundamental para tornar as análises mais efetivas, considerando um número maior de argumentos, a partir de fontes organizadas e estruturadas, apresentando uma maior contextualização do domínio que está sendo analisado. Diante de tais pontos, esta pesquisa buscou identificar e apresentar algumas intersecções existentes entre os processos de *Big Data* e as tecnologias da Web Semântica, indicando como estas últimas contribuíram para aprimorar em especial as análises de dados realizadas.

A utilização do *Linked Data* como fonte de dados, o uso de ontologias para aprimorar os processos de fusão e análises e o aperfeiçoamento da interoperabilidade no *Big Data*, foram os três pontos que foram discutidos nesta pesquisa, apontando alguns detalhes sobre como se daria a aplicação de algumas tecnologias da Web Semântica para tornar os processos de *Big Data* mais contextualizado semanticamente.

Portanto, esta pesquisa avança na intersecção entre estes dois campos de estudos, discorrendo sobre como a Web Semântica, que apresenta um corpus teórico mais consistente, para tornar o *Big Data* mais eficiente ao inserir uma ótica semântica nos processos analíticos. Enquanto trabalhos futuros, busca-se realizar a implantação de experimentos que comprovam na prática a viabilidade dos pontos discutidos.

Referências

BANSAL, S. K. Towards a Semantic Extract-Transform-Load (ETL) Framework for Big Data Integration. **IEEE**, jun. 2014. Disponível em: < http://bit.ly/2ulZ3a5>.Acesso em: 22 jul.2017.

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The semantic web. **The Semantic Web**, v. 284, n. 5, p. 28–37, maio 2001.

BERNERS-LEE, T., 2006. **Linked Data Principles**. Disponível em < http://bit.ly/1x6N7XI>. Acesso em: 15 jul. 2017

BEYER, M. A., LANEY, D., 2012. **The importance of "Big Data'**: a definition. Stamford, CT: Gartner.

BUGEMBE, M. **Finding Value in Data**: Determining Where Data Science has The Greatest Impact. O'Reilly: Sebastopol, 2016.

DEMCHENKO, Yuri et al. Addressing big data issues in scientific data infrastructure. In: Collaboration Technologies and Systems (CTS), 2013 **International Conference on**. IEEE, 2013. p. 48-55.

GANTZ, J., REINSEL, D., 2017. **Data Age 2025**: The Evolution of Data to Life-Critical. Don't Focus on Big Data; Focus on the Data That's Big. Disponível em: <http://bit.ly/2tPWoU8>. Acesso em: 21 jul. 2017.

HARMELEN, F. Semantic Web Technologies As The Foundation For The Information Infrastructure. In: VAN OOSTEROM, P.; ZLATANOVA, S. (Eds.). . **Creating Spatial Information Infrastructures**. [s.l.] CRC Press, 2008. p. 37–52.

LANEY, D., 3D Data Management: Controlling Data Volume, Velocity and Variety. 2001.

MAYER-SCHÖNBERGER, V; CUKIER, K. **Big data**: A revolution that will transform how we live, work, and think. Boston: Houghton Mifflin Harcourt, 2013.

PEREIRA JUNIOR, V A. et al. Using Semantics to Improve Information Fusion and Increase Situational Awareness. In: Advances in Safety Management and Human Factors. **Anais…** Springer International Publishing, 2016. p. 101-113.

SANTAREM SEGUNDO, J. E.; CONEGLIAN, C. S. Web semântica e ontologias: um estudo sobre construção de axiomas e uso de inferências. **Inf & Inf**, [S.l.], v. 21, n. 2, p. 217–244, dez. 2016. Disponível em: <http://bit.ly/2uLpbgL>. Acesso em: 22 jul. 2017.

VICTORINO, M. C. et al. Uma proposta de ecossistema de big data para a análise de dados abertos governamentais concetados. **Informação & Sociedade**, v. 27, n. 1, 2017.

W3C. Resource Description Framework (RDF). 2004. Disponível em: <https://www.w3.org/RDF/>.Acesso em: 28 ago. 2017.

ZIKOPOULOS, P.; EATON, C.; DEROOS, D. **Understanding BigData**: Analytics for enterprise class hadoop and streaming data. McGraw-Hill, New York. 2012.

Data Visualization of the Brazilian National High School Exam: VisDadosEnem

Visualização de Dados do Exame Nacional Brasileiro do Ensino Médio: VisDadosEnem

César H. C. Santos[1], Maykon C. Freitas[1], Robson R. Lemos[1], Alexandre L. Gonçalves[1]
(1) UFSC, R. Pedro J. Pereira, 150. Araranguá, SC, Brasil, {robson.lemos, a.l.goncalves}@ufsc.br

Resumo:
Hoje em dia, baseado na quantidade de dados disponíveis surgem situações em que a visualização da informação torna-se fundamental para o entendimento e interpretação dos dados. Dentro deste contexto, este estudo tem como objetivo relacionar e explorar, através de técnicas de visualização, dados abertos educacionais em um contexto multidisciplinar. Para tal, foi desenvolvido uma aplicação Web para visualização de dados do exame nacional Brasileiro do ensino médio (ENEM). Neste estudo utilizou-se os microdados do ENEM, a partir do portal Brasileiro de dados abertos. Como resultado das análises comparativas realizadas sobre os dados, por estados, regiões e tipos de escolas, observou-se quais estados obtiveram as maiores e menores médias do ENEM em relação as médias nacionais. Por exemplo, no estado de Santa Catarina os alunos oriundos de escolas particulares obtiveram 12.70% acima da média nacional para este tipo de escola e a menor média de idade dos alunos inscritos correspondendo a 21,6 anos de idade. A aplicação Web de visualização possibilitou uma análise comparativa sobre o grande conjunto de dados e a exploração de possibilidades de visualização que seriam dificilmente possíveis de realizar por meio de representações tradicionais.
Palavras-chave: Visualização da Informação, Ciência dos Dados, Dados Abertos, Educação

Abstract:
Nowadays, based on the amount of available data, there are situations in which the information visualization becomes fundamental for the understanding and interpretation of the data. In this context, the study aims to relate and explore, through visualization techniques, open educational data in a multidisciplinary context. For that, a Web application was developed for data visualization of the Brazilian national high school exam (ENEM). In this study, the ENEM microdata was used, from the Brazilian open data portal. As a result of the comparative analyzes performed on the data, by states, regions and types of schools, it was observed which states obtained the highest and lowest averages of ENEM in relation to national averages. For example, in the state of Santa Catarina, students from private schools obtained 12.70% above the national average for that type of school and the lowest average age of enrolled students corresponding to 21.6 years old. The visualization Web application made possible a comparative analysis on the large dataset and the exploration of visualization possibilities that would be difficult to perform through traditional representations.
Keywords: Information Visualization, Data Science, Open Data, Education

1 Introduction

Based on the technological advances in data storage and retrieval there is a massive increase in the information available to any type of user through a web browser. There is a large volume of information available on the Internet, but the data are irrelevant when there is no meaning. In order to get the information, the data must be interpreted and related so that they are presented within a context and allow the generation of knowledge. In this way, developers have been paying special attention to data visualization. According to Freitas et al. (1995), combining aspects of computer graphics, human-computer interaction and data mining, the information visualization allows the presentation of data in a graphics form. In that way, the user can use their visual perception to better analyze and understand the information.

In general, the data do not have a direct, obvious and natural representation, which contributes to the need of a good visual analysis technique. Such a technique in turn can use visual representations to present the data in a graphics form (LUZZARDI, 2003). In addition, according to Card, Mackinlay and Shneiderman (1999) the use of visual representations of interactive data with the support of the computer allows to increase the knowledge.

The information visualization is ideal for exploratory data analysis. Human eyes are naturally attracted to trends, patterns and exceptions that would be difficult to find using traditional approaches such as tables or text (FEW, 2009). The information visualization corresponds to the visual representation of abstract data to increase knowledge (SHIXIA et al., 2014). For example, representation of temporal statistics on the numbers of healthy children births, stock market trends, and information associated with the use of social networks. In general, interactive visual interfaces facilitate the process of expanding

knowledge and assist in the process of identifying patterns that are usually difficult to perceive.

2 Purpose of the Study

This study aims to relate and explore a large educational open dataset through information visualization techniques. For that, the database chosen was the microdata of the Brazilian national high school exam (ENEM), which has more than 7 million lines.

In order to investigate the relevance of information visualization tools, the Web application entitled VisDadosEnem has been developed so that the following research question can be answered: interactive information visualization tools are relevant in relating and exploring educational open data in a multidisciplinary context?

3 Methodology

In the elaboration of this study an applied and technological research was used. In order to obtain the data, the data portal of the Brazilian Federal Government was used, where open data is available from the most diverse public areas. And, for the Web application project, technologies were adopted that are more adequate to the process of treatment and visualization of educational data of the ENEM.

3.1 VisDadosEnem

Nowadays there is an increase in the data volume from the most different sources, as well as an increase in the open data made available by the Brazilian government (DADOS ABERTOS, 2016) and other governments (MÁCHOVÁ and LNENICKA, 2017). Open data from several areas can be found, such as Education, Health, Technology, Defense, Security, Transportation, Traffic, Social Security, and Labor. With the data available in the administrative, technical and financial sectors of business and government organizations, there is a need for systems that allow to perform data analysis. Further, the analysis results make it possible to assist decision makers in extracting the information needed for decision-making process.

For the Web application project, it was carried out the data modeling and the interface design. For data modeling, the conceptual model was first elaborated based on entity-relationship diagrams. In order to do that, only one table was needed to represent the candidate with their attributes. For that, it was included the enrollment number, the type of school attended, the state that resides, the marks in the knowledge areas, and the candidate's age based on 2014 ENEM microdata (DADOS ABERTOS, 2017). In addition, the logical model for the identification of the primary key, as well as the types and size of each attribute was elaborated. And, finally, for the description of the physical model, the table containing its attributes was created. Regarding the database, we chose MySQL© and for the Web application to be able to query the database, the PHP© language was adopted.

For the interface design, the conceptual design model, the prototype of the conceptual model and the detailed design of the user interface were developed. For that, it was taken into account functionalities associated to the data visualization methods to explore an adequate visual analysis. In this way, technologies for developing Web applications such as HTML©, CSS© and Javascript© for the interface development were adopted. For the data visualization, the D3© (Data Driven Documents) library (D3, 2017) was used, which seeks to facilitate the understanding of the data by combining the interaction with the information visualization techniques. Finally, we used the NVD3© library (NVD3, 2017) that provides ready-made reusable graphics and graphic components for D3©.

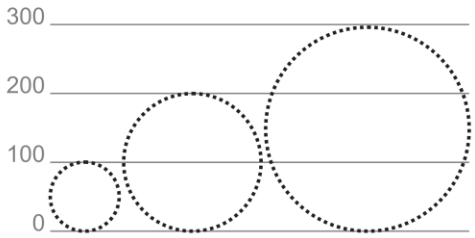
The information visualization techniques can be based on graphics, quantitative, hierarchical information, and number exploration. In addition, information visualization techniques can be classified according to a given set of data (WARD, GRINSTEIN, KEIN, 2015), such as: Multivariate data are those that do not usually have an explicit spatial attribute; Hierarchies and Trees are those that contain hierarchical characteristics and are considered a repository of data where there is a relationship with data subitems; Graphs and Networks are defined by their relationship characteristics based on graphs; And, Text and Documents are considered as being

literal, as a string, and defined by the set of objects as words, sentences, paragraphs, and documents.

The information visualization technique adopted for the visual representation of the educational data was the Multivariate data. The technique was based on a combination of elements where we can make use, for example of points, lines, and regions. We chose to develop a simple visualization with an easy understanding, and based on that, we decided to use bubbles (circles) for the visual representation. That type of visualization technique is known as bubble maps. The circle is a popularly known flat geometric figure, and, according to the variables to be displayed, the following properties have been added: area, color, and border color.

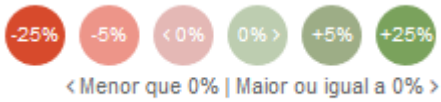
The circle area property is directly associated with the numeric value of the variable. The circle size is proportional to the value associated with the variable, as shown in Figure 1. In the case of VisDadosEnem, it indicates the number of candidates enrolled in the exam.

Figure 1: Circle area property. Prepared by the authors (2017).



The circle color property serves to classify the variable by intervals (Figure 2). The color property can assume two characteristic colors, where one color represents the lowest value and the other, the highest one. The interval between those values is represented by the variation of the intensity of each color. Higher values (positives or negatives) will have a more intense color tone. In the case of VisDadosEnem, it represents the comparison of the percentage of the average grade of a particular school type with a state or with Brazil.

Figure 2: Circle color property. Prepared by the authors (2017).



And, the border color property classifies the type that the variable can assume (Figure 3). Its color can vary according to the existing type, but it cannot be repeated in the same set, due to its unicity. In the case of VisDadosEnem, it represents the type of school (e.g., public, mostly public, private, and mostly private).

Figure 3: Border color property. Prepared by the authors (2017).

VisDadosEnem allows the visualization of the number of enrolments, the average grades by state or region, and the average age of the candidates. It also allows a comparative analysis of the averages obtained for each state or region by each type of school or area of knowledge through interactive visualization techniques. The Web application has the following main functionalities: Visualization by area of knowledge, by data between states, by types of schools, and by region[1].

The visualization by area of knowledge[1] allows visualizing the grades average obtained in all the states in the different types of school by area of knowledge. As shown in Appendix A, the interface allows the user to activate interactively select areas of knowledge for visualization through a check box menu.

The visualization of data between states[1] allows the comparison of grade averages obtained from one state compared to another state or to Brazil.

The visualization by type of schools[11] allows to analyze comparatively the performance of the types of school present in the states in function of the percentage of the grades average with respect to Brazil and the

[1] VisDadosEnem is available at: http://labdata.sites.ufsc.br/visdadosenem

average age of the candidates. The visualization occurs through bubble maps organized in a 2D chart (Appendix B) where the vertical axis represents the percentage of grades average for the different types of school and the horizontal axis represents the average age of the candidates.

And finally, the visualization by region[1] allows to analyze the performance of schools by region. The visualization also occurs through bubble maps organized in a 2D chart where the vertical axis represents the percentage of grades average for the regions of Brazil and the horizontal axis represents the average age of the candidates.

4 Results

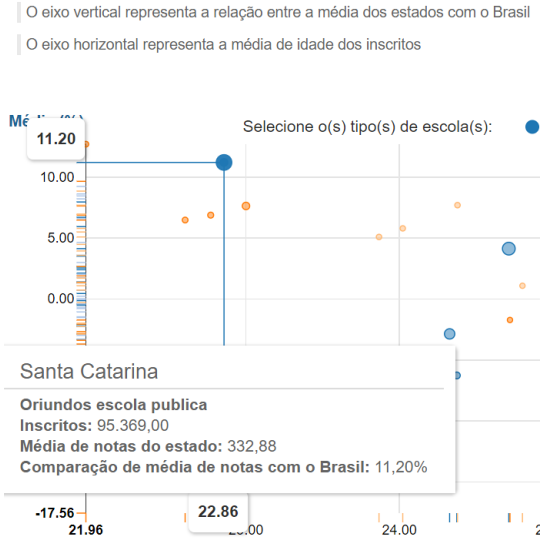
Through the performed analysis with the information visualization techniques, it was possible to relate and explore the information available in VisDadosEnem.

In the visualization feature by area of knowledge (Appendix A), it was possible to observe which states in Brazil, with students coming from public or mostly public schools, obtained an average grade higher than the national average compared to other states, as shown in Figure 4. That is, according to the bubble maps visualization the states that are above average in the results for public schools have circles, with the inner color in the green color variation and, respectively, present the border color in the blue color variation. It means that students are either in public schools (dark blue) or have attended mostly of the time in public schools (light blue).

Figure 4: A section of the visualization by area of knowledge. Prepared by the authors (2017).

In the visualization feature by type of schools (Appendix B), it was possible to observe, that the state of Santa Catarina had the highest grade average regarding students from private schools with 12.70% above the national average with the lowest average age of students enrolled corresponding to 21.6 years old. And, also, for students from public schools with 11.20% above national grade averages with the average age corresponding to 22.8 years old, as shown in Figure 5. On top of that, the lowest grade average regarding students from private schools was in the state of Amazonas with 17.56% below national grade averages with the average age corresponding to 26.8 years old. Also, for students from public schools, the state of Mato Grosso do Sul obtained the lowest grade average with 13.32% below the national average with the average age corresponding to 26.7 years old.

Figure 5: A section of the visualization by type of school. Prepared by the authors (2017).



4 Final Considerations

Although VisDadosEnem presents a set of basic functionalities for information visualization, the use of the visualization features made possible comparative analysis in the ENEM open data in order to relate and explore educational information. That type of exploratory analysis of information would be difficult to perform through tables or traditional graphics. In this way, with this

simple information visualization environment, one can see the relevance of a visualization tool for visual data exploration in a large volume of data.

In addition, the study provided an online tool of public utility, which can be easily accessed and used for different types of visual analysis. It may be of interest to educational bodies, which can analyze results related to different types of schools, states, and regions. In this way, it can contribute to decision making process based on the information identified through the information visualization techniques.

In terms of future work, the educational open data used for the information visualization can be used in different scenarios. For example, it can be used to perform statistical studies, and to evaluate the quality of the public school in comparison to the private school in relation to the average grades obtained. Besides that, it is possible to identify which Brazilian states have a better or worse performance in the ENEM. Finally, as future work it would be interesting to extend the functionality of VisDadosEnem to allow the automatic insertion of updated ENEM data. In this way, as the data become available in the open data portal, it would be possible to perform temporal analysis as well as performance comparison between schools or states over time.

References

CARD, K. S.; MACKINLAY, J. D.; SHNEIDERMAN, B. **Readings in Information Visualization, using vision to think**. Morgan Kaufmann, Cal. USA, 1999.

DADOS ABERTOS. **Portal Brasileiro de Dados Abertos.** Available at: <http://dados.gov.br/>. Accessed: 22 Jul. 2017.

D3. **Data-Driven Documents**. Available at: < http://d3js.org/>. Accessed: 22 jul. 2017.

FEW, S. **Now you see it: simple visualization techniques for quantitative analysis, Analytics.** Press, 2009.

FREITAS, C. M. D. S; WAGNER, F. R. Suporte às tarefas da análise exploratória visual. **Revista de Informática Teórica e Aplicada**, vol. 2, n.1, p. 5-36, jan. 1995.

LUZZARDI, P. R. G. **Critérios de Avaliação de Técnicas de Visualização de Informações Hierárquicas**. Tese de doutorado (Programa de Pós-Graduação em Computação), UFRGS, Porto Alegre, 2003.

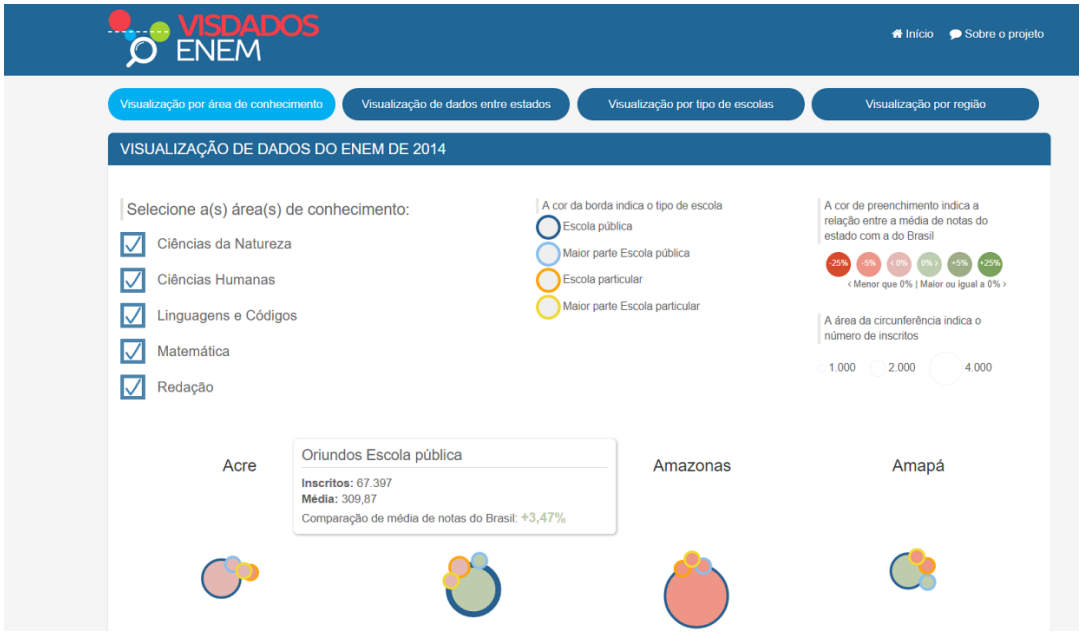
MÁCHOVÁ, R.; LNENICKA, M. Evaluating the Quality of Open Data Portals on the National Level. **Journal of Theoretical and Applied Electronic Commerce Research**, vol. 12, n. 1, p. 21-41, jan. 2017.

NVD3. **Re-usable Charts for D3**. Available at: <http://nvd3.org/>. Accessed: 22 jul. 2017.

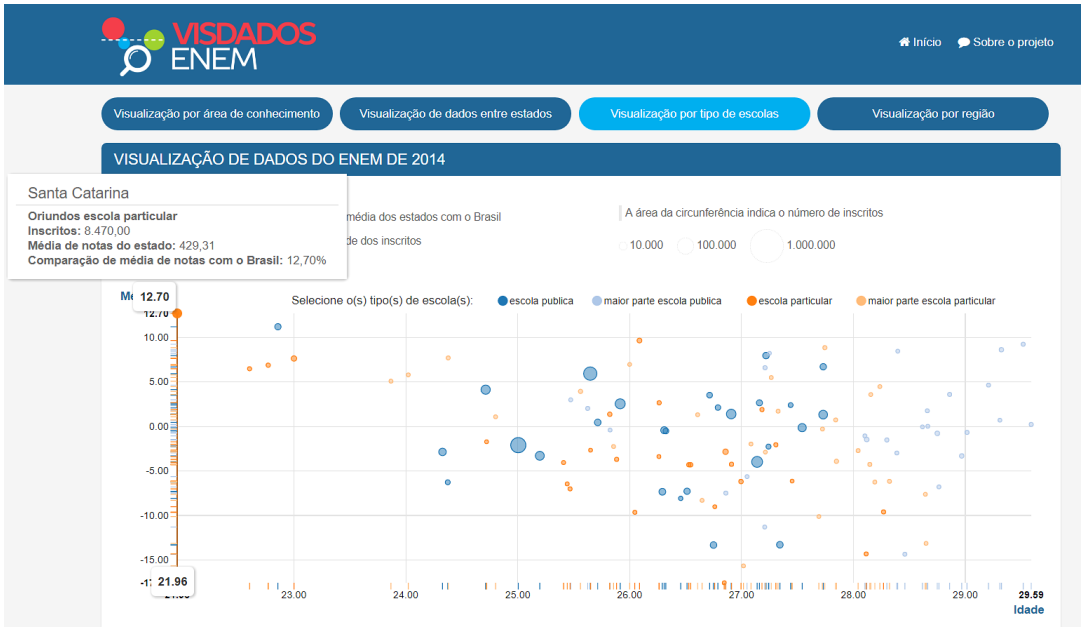
SHIXIA, L.; WEIWEI, C.; YINGCAI, W.; MENGCHEN, L. A survey on information visualization: recent advances and challenges. **The Visual Computer**, vol. 30, n. 12, p. 1373-1393, 2014.

WARD, M.; GRINSTEIN, G.; KEIN, D. **Interactive data visualization: foundations, techniques, and applications**. CRC Press, 2015.

Appendix A - Visualization by area of knowledge. Exploratory analysis of the information through visualization by area of knowledge. The check boxes for each knowledge area can be enabled or disabled. Prepared by the authors (2017).



Appendix B – Visualization by type of schools. Exploratory analysis of the information through visualization by type of school. Each state has a representation for each of the 4 types of schools (i.e., blue circle - public school, blue light - mostly public school, orange - private school, and light orange - mostly private school). The circles for each type of school can be enabled or disabled. Prepared by the authors (2017).



Visualização de Informações de Variação de Incidência Criminal em Sistema Orientado à Obtenção de Consciência Situacional

Information Visualization of the Criminal Incidence Variation in a System Oriented to the Acquisition of Situational Awareness

Lucas Zanco Ladeira[1], Leonardo Castro Botega[12], João Henrique Martins[2], Vagner Pagotti[2]

(1) Grupo de Interação Humano-Computador, Centro Universitário Eurípides de Marília, Marília – São Paulo – Brasil, lznladeira@gmail.com.

(2) Stratelli – Inteligência Estratégica, Marília – São Paulo – Brasil, botega@univem.edu.br.

Resumo:

Consciência Situacional (SAW – Situational Awareness) é um processo cognitivo que diz respeito à percepção, entendimento e projeção dos estados de entidades de interesse em um determinado ambiente crítico. Esse conceito é de grande importância para sistemas de gerenciamento de riscos com base em dados criminais, pois falhas de SAW podem influenciar negativamente as decisões de analistas e comprometer suas ações, principalmente quanto à alocação de recursos de segurança. Para tal, considera-se que as informações disponíveis para a aquisição de SAW devem ser representadas de forma precisa, contribuindo com a diminuição das chances de um operador tomar uma decisão que possa criar riscos à integridade física de pessoas ou patrimônios. Desta maneira, é necessário determinar quais e de que maneiras as informações devem ser apresentadas ao usuário, considerando o volume, complexidade e heterogeneidade dos dados. Para contribuir com a SAW de operadores de sistemas de avaliação de riscos, são propostas neste trabalho visualizações para representar a incidência criminal sob parâmetros customizáveis, possibilitando uma análise quanto à evolução no volume de ocorrências ao longo do tempo. Tais visualizações compreendem sobreposições em mapa, desenvolvidas utilizando a biblioteca D3JS, e diferentes gráficos para representar consultas feitas por analistas. Resultados preliminares são promissores e ilustram a aplicabilidade da abordagem para promover a SAW de analistas criminais.

Palavras-chave: Visualização de informações; consciência situacional; cenário criminal.

***Abstract*:**

Situational Awareness (SAW) is a cognitive process that concerns the perception, understanding and projection of the states of entities of interest in a given critical environment. This concept is of great importance for risk management systems based on criminal data, because SAW failures can negatively influence analysts' decisions and compromise their actions, especially regarding the allocation of security resources. To this end, it is considered that the information available for the acquisition of SAW should be accurately represented, contributing to the reduction of the chances of an operator making a decision that could create risks to the physical integrity of people or assets. In this way, it is necessary to determine what and in what ways the information should be presented to the user, considering the volume, complexity and heterogeneity of the data. In order to contribute to the SAW of operators of risk assessment systems, it is proposed in this work visualizations to represent the criminal incidence under customizable parameters, allowing an analysis as to the evolution in the volume of occurrences over time. Such visualizations comprise map overlays, developed using the D3JS library, and different graphs to represent queries made by analysts. Preliminary results are promising and illustrate the applicability of the approach to promoting the SAW of criminal analysts.

Keywords: *Information visualization; situational awareness; criminal scenario*.

1 Introdução

Consciência situacional refere-se a estar ciente dos estados de entidades de interesse presentes em um cenário e entender o que as informações disponíveis sobre estes representam (ENDSLEY, M. R., 2011). A mesma pode ser dividida em três níveis: nível 1 a percepção dos elementos presentes, nível 2 a compreensão da situação e nível 3 a projeção de estados futuros (ENDSLEY, M. R., 2011).

O primeiro nível caracteriza-se pela percepção das informações presentes, ou seja, no cenário abordado, notar a incidência criminal em uma determinada região, evoluindo de acordo com o tempo. Um outro exemplo seria do painel de um carro que mostra a quantidade de combustível, a quilometragem, a velocidade em um dado momento, entre outros. Considerando que os outros níveis dependem do que é apresentado no primeiro ele é o mais importante.

No segundo nível, é criado o relacionamento entre as informações apresentadas e analisado o cenário como um todo. Portanto, é possível avaliar no exemplo dado a velocidade do carro, a localização dele na via, outros carros próximos para saber se está perigosamente próximo.

Por fim, no terceiro nível é feita a projeção do cenário em um tempo futuro. Considerando o exemplo seria avaliar se o carro irá bater no da frente se continuar na mesma velocidade que está, ou a necessidade de mudar de via caso o caminho necessite.

É possível observar a necessidade de SAW em sistemas críticos ou de gerenciamento de risco, sendo que, uma decisão errada pode ocasionar risco a integridade física de pessoas. Para obter uma boa SAW o usuário deve compreender a representação das informações disponíveis para o completo entendimento do cenário e projetar as mudanças do mesmo.

Contudo, devido ao grande volume de dados disponíveis para a análise criminal, é difícil considerar todos os registros e identificar todos os locais de risco. Além disso, o texto de registros criminais (boletins de ocorrência) é heterogêneo e complexo. Sendo assim, algumas maneiras de os utilizar para agregar eficiência no combate à criminalidade é processar esses dados com fusão de dados ou representá-los de forma a facilitar a identificação de padrões tanto em horários, locais e ações.

Uma maneira de identificar padrões nas ocorrências criminais e estimular a SAW é pela representação gráfica das informações e através da aplicação dos princípios de SAW (ENDSLEY, M. R., 2011). Como exemplo o primeiro princípio que dita a organização das informações em torno dos objetivos. Como podemos observar em (JAKKHUPAN, W.; KLAYPAKSEE, P., 2014) são apresentadas informações sobre o local das ocorrências como também a incidência das mesmas.

Analistas criminais devem interpretar devidamente as informações e buscar o melhor subsídio para sua SAW, portanto, o que será considerado na tomada de decisão assertiva depende do estímulo que este recebe do sistema. Adicionalmente, para a visualização de informações, é de grande importância a qualidade da apresentação de um grande número de informações.

2 Objetivos

O objetivo deste trabalho é apresentar visualizações inovadoras para representar informações criminais distribuídas espacialmente e transformada ao longo do tempo. Tais visualizações visam contribuir com a obtenção de consciência situacional de humanos operadores a partir do estímulo ao entendimento da variação da incidência criminal em cada território, utilizando dados de ocorrências oficialmente registradas.

Com uma melhor SAW, tomada de decisão poderá ser enriquecida com melhores subsídios e contribuir para a análise de dados de regiões críticas.

Essas informações são apresentadas em mapas customizados, orientados à representação da incidência criminal, bem como em gráficos para contribuir com a observação e estimulo de SAW quanto à evolução do número de ocorrências sob múltiplos filtros.

3 Metodologia

Os dados do domínio de gerenciamento de riscos demandam de visualizações adaptáveis, considerando a sua dinamicidade, heterogeneidade, complexidade e seu potencial crítico.

Os dados empregados nesse trabalho são de boletins de ocorrência do Estado de São Paulo, processados sob métodos de inferência utilizando processamento de linguagem natural, mineração e fusão de dados (BOTEGA, L. C., et al, 2017). Neste contexto, são buscados nos dados, termos e significados que remetam a tipos e características de crimes, em sinergia com a consulta desejada por analistas criminais.

Tais tipos e características de crimes, em conjunto com variáveis ambientais, definem a incidência criminal, que contabiliza e generaliza os eventos criminais por região. Desta maneira, podemos dividir essa incidência em diferentes níveis desde o nível seguro até o de extremo perigo. Nesse trabalho foram utilizados 5 níveis para descrever a evolução dos índices de criminalidade.

Além disso, é importante ressaltar que tal incidência criminal é dinâmica, o que demanda a possibilidade de customização e reu-

tilização das técnicas empregadas para regiões (setores censitários, cidades) e/ou informações diferentes, considerando que podem ser consultados diferentes tipos de crime, quantidades, níveis de incidência distintos, entre outros. Um exemplo de consulta é por um crime específico como o estupro, nesse caso a quantidade entre anos distintos será diferente do que considerando a ocorrência total de crimes.

Para a implementação destes requisitos, foi utilizada uma biblioteca *javascript* chamada *Data-Driven Documents* (D3JS). Essa biblioteca disponibiliza funções para a criação de visualizações aplicando efeitos gráficos, tais como pontos e retas, além de objetos na Document Object Model (DOM) de websites. O que favorece a utilização da mesma em comparação com outras como processing e dygraphs, é a vasta lista de exemplos e disponibilização de tutoriais *online*.

Em complemento, é importante mencionar que a participação de especialistas no domínio de gerenciamento de risco foi fundamental para a obtenção e incorporação de requisitos de análise criminal para a sustentação, viabilizada por entrevistas e análise de tarefas dirigida por objetivos (GDTA) (BOTEGA, L. C., 2016).

Finalmente, destaca-se que neste trabalho foram também considerados os princípios de design de interfaces e visualizações orientados a SAW, postulados por (ENDSLEY, M. R., 2011).

4 Visualização de Incidência Criminal para a Obtenção de Consciência Situacional

Na primeira visualização são apresentadas três sobreposições (*overlays*) distintas, no mesmo mapa de São Paulo, representando a evolução na incidência criminal. A primeira delas apresenta uma incidência alta, exemplificando o cenário de roubos, considerando o tom de vermelho preenchendo o estado. As cores aplicadas nas sobreposições seguem os princípios de consciência situacional (ENDSLEY, M. R., 2013). O mesmo pode ser observado na Figura 1.

Figura 1 - Estado de São Paulo com incidência criminal

Considerando um cenário no qual um filtro com dois anos distintos é aplicado. Além disso, a quantidade de crimes no estado aumentou, comparando o primeiro e o segundo anos. Em um primeiro ano a incidência criminal registrou 60 crimes e em um segundo ano registrou 75. Sendo assim, para apresentar essa diferença é proposto preencher o estado com uma cor que representa a quantidade de ocorrências no último ano e as bordas com a cor do ano anterior.

O tamanho da borda representa o quanto aumentou a incidência sendo que quanto maior a borda mais a incidência difere entre os anos escolhidos no filtro. Isso pode ser observado na Figura 2.

Figura 2 - Aumento da incidência criminal

Um outro cenário é a aplicação de um filtro com dois anos distintos, onde a incidência criminal diminuiu relacionando o primeiro ano com o segundo. O primeiro ano teve 60 incidências e o segundo 40. Para apresentar a borda possui a cor do ano anterior, onde a cor de preenchimento do estado é referente a incidência criminal do último ano. Além disso, a área de preenchimento do estado é menor o que indica que a quantidade de ocorrências

de um ano para o outro diminuiu, podemos observar na Figura 3.

Figura 3 - Diminuição da incidência criminal

Considerando as visualizações apresentadas anteriormente podemos identificar que caso a quantidade criminal tenha diminuído 70% quase o estado todo seria preenchido com a cor branca. Para que isso não ocorra é necessário normalizar a incidência, onde o peso da mesma tenha um limite superior mantendo a visualização aplicável.

Um outro ponto a ser considerado é que, para estados diferentes, devemos apresentar a mesma visualização caso a diferença na incidência criminal durante os anos aplicados no filtro seja a mesma. Sendo assim, a fórmula de normalização utilizada foi uma modificação da publicada por (YUSUF, L. M., et al, 2010), onde ocorre a variação entre [0, 1] dos dados. Na mesma x representa uma diferença na incidência criminal a ser aplicada a normalização, min é a menor diferença encontrada, e por fim, max é a maior diferença. A Equação 1 foi empregada para a normalização.

$$n = \frac{\mathrm{x}-\mathrm{min}}{\mathrm{max}-min} \qquad (1)$$

Após a normalização dos dados, com testes feitos nas sobreposições, foi possível observar a utilização de no máximo 35 pixels de preenchimento branco sem deformar o preenchimento vermelho. Dessa maneira, é utilizada a dispersão dos valores já normalizados sendo de 0 até 35 pixels pela multiplicação do resultado por 35.

É necessário citar que o mesmo vale para o aumento na incidência criminal não deixando que a cor do ano anterior preencha um espaço maior do que a cor do último ano. Dessa maneira, essa visualização apresenta a diferença na quantidade de ocorrências comparando dois anos, sendo que, o último ano aplicado no filtro tem maior foco.

Um outro cenário a ser considerado é da utilização de múltiplos filtros temporais onde devem ser apresentadas a evolução criminal entre vários anos distintos. É suposta a aplicação de filtros dos anos de 2012 até 2016, sendo que, pela quantidade de informações diferentes a visualização no mapa poderia confundir o usuário. Além disso, ocorreu apenas crescimento na quantidade de ocorrências entre os anos apresentados. Sendo assim, foi feito um gráfico, utilizando como base um exemplo encontrado no site da D3JS, chamado *sequences sunburst*, sendo que, o mesmo foi customizado.

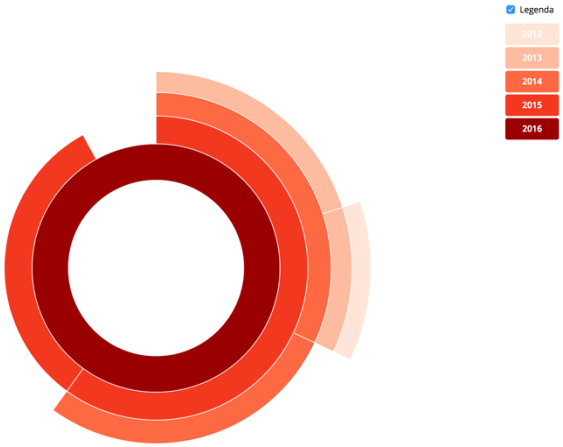


Figura 4 - Incidência criminal em anos distintos

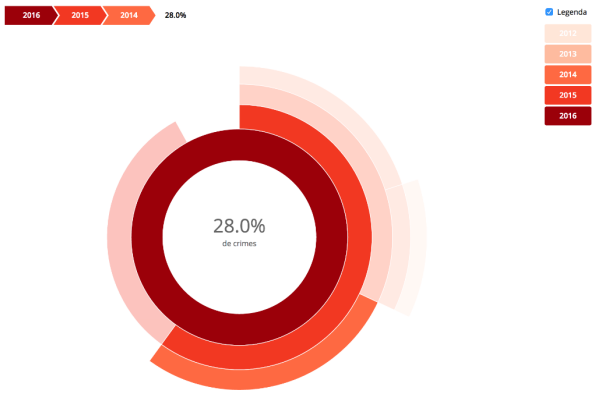


Figura 5 - Comparativo de porcentagem de aumento da incidência criminal

É possível observar o resultado na Figura 4 onde cada cor representa um ano distinto. Ao selecionar uma seção é apresentado o crescimento na incidência criminal relacionado com o ano anterior. No caso do exemplo do ano de 2013 para 2014 houve um aumento

de 28% de crimes no estado de São Paulo. Isso pode ser observado na Figura 5.

Um ponto a ser considerado é o de que os dados, para serem apresentados nesse gráfico, devem estar ordenados de forma crescente tornando possível o correto cálculo da diferença entre os anos. Sendo assim, um usuário da aplicação observaria os anos em ordens diferentes dependendo do que ocorreu tornando não satisfatório para o entendimento das informações dispostas.

Pode-se afirmar que é necessária uma visualização mais clara das informações e, para anos iguais no filtro, estejam dispostas da mesma maneira no gráfico, onde a diferença entre os anos esteja descomplicada. Dessa maneira é proposto um gráfico em barras que ao focar com o ponteiro do mouse em uma barra referente a um ano é apresentada a incidência criminal do mesmo. É possível observar que o gráfico da Figura 6 apresenta as informações de maneira clara, onde o usuário não precisa de muito tempo para entender o que está disposto.

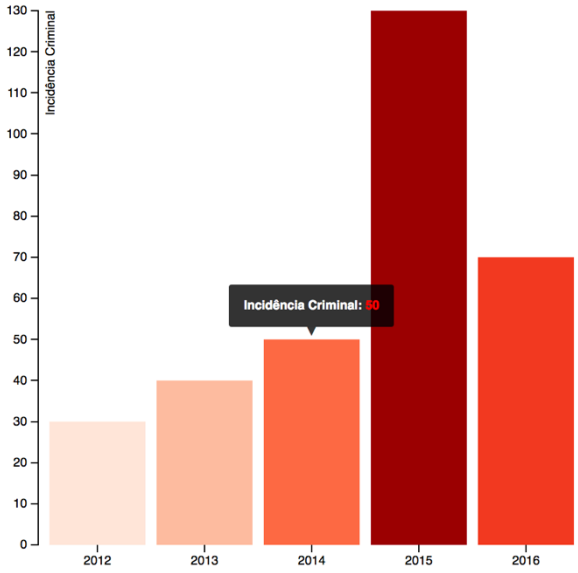


Figura 6 - Gráfico em barras comparativo da incidência criminal em anos distintos

Para que sejam utilizadas diferentes visualizações de acordo com a quantidade de filtros temporais aplicados é proposta a visualização em uma sobreposição no mapa para comparar dois anos apenas. Sendo que, caso sejam aplicados mais de dois anos nos filtros a sobreposição pode apresentar a diferença na incidência criminal do primeiro e último ano, e ao selecionar o estado o gráfico em barras é apresentado com todas as informações.

4 Considerações Finais

O objetivo deste trabalho foi apresentar visualizações que representam informações criminais e seu nível de incidência distribuído por regiões, visando contribuir com a obtenção de consciência situacional.

Os dados de incidência criminal representam dados criminais obtidos de anos distintos pela aplicação de dois ou mais filtros temporais. Essas visualizações são apenas do estado de São Paulo, sendo que é possível aplica-las a mesma para todos os estados brasileiros, municípios ou setores censitários. No caso da aplicação em setores censitários é possível avaliar a migração do crime considerando que setores com características e localizações próximas geralmente estão sujeitos aos mesmos crimes.

Também foram apresentados conceitos sobre consciência situacional, que é de grande importância para a tomada de decisão de analistas criminais em sistemas de gerenciamento de risco.

Como trabalhos futuros é necessário analisar quais componentes de SAW as visualizações propostas possuem maior impacto, como também a aplicação da metodologia SART (*Situation Awareness Rating Technique*) em especialistas do domínio.

Referências

ENDSLEY, M. R. Designing for Situation Awareness: An Approach to User-Centered Design. Segunda Edição, 2nd ed. Boca Raton, FL, USA, 2011.

JAKKHUPAN, W.; KLAYPAKSEE, P. A Web-Based Criminal Record System Using Mobile Device: A Case Study of Hat Yai Municipality. Em 2014 IEEE Asia Pacific Conference on Wireless and Mobile, ago. 2014. p. 243–246.

“Eu fui roubado”. Disponível em: <http://eufuiroubado.com/>. Acessado em: 05 de jul. 2017.

“Onde fui roubado”. Disponível em: <http://www.ondefuiroubado.com.br/>. Acessado em: 05 de jul. 2017.

"Wikicrimes". Disponível em: <http://wikicrimes.org/>. Acessado em: 05 de jul. 2017.

BOTEGA, L. C.; OLIVEIRA, A. C. M.; PEREIRA JUNIOR, V. A.; SARAN, J. F.; VILLAS, L. A.; ARAUJO, R. B. Quality-Aware Human-Driven Information Fusion Model. 20th International Conference on Information Fusion, 2017.

BOTEGA, L. C. Modelo de Fusão Dirigido por Humanos e Ciente de Qualidade de Informação. 2016.

ENDSLEY, M. R. The Oxford Handbook of Cognitive Engineering, 2013. p. 88-108.

YUSUF, L. M.; OTHMAN M. S.; SALIMG J. Web Classification Using Extraction and Machine Learning Techniques. International Symposium on Information Technology, vol. 2, jun. 2010. p. 765–770.

WIDAT'2017 – Sessão Técnica III

Dia: 05/09/2017 – Horário: 16h30 às 18h30

ASPECTOS DE EXPERIÊNCIA DE USUÁRIO NO PORTAL WIKICI

Aspects of User Experience on WikiCI *Website*

Victor Ubiracy Borba[1], Elaine Parra Affonso[2], Ricardo Cesar Gonçalves Sant'Ana[3]

(1) UNESP, Av. Hygino Muzzi Filho,737, Mirante, Marília-SP, borba.victor.borba@gmail.com
(2) UNESP, Av. Hygino Muzzi Filho,737, Mirante, Marília-SP, elaineaffonso@marilia.unesp.br
(3) UNESP, Av. Hygino Muzzi Filho,737, Mirante, Marília-SP, ricardosantana@marilia.unesp.br

Resumo:
Os usuários têm produzido cada vez mais conteúdo dentro de ambientes colaborativos, com isso, é imprescindível que sua experiência dentro desses ambientes seja a mais satisfatória possível. Para tal, a Arquitetura da Informação (AI) junto a Ciência da Informação (CI) tem a finalidade de proporcionar o equilíbrio das necessidades do usuário, emergindo no contexto do sentimento do usuário, suas alegrias e frustrações, questões que podem ser interpretadas pela Experiência de Usuário (*User Experience* – UX). Este trabalho tem como objetivo analisar aspectos da AI, especificamente em relação à aderência do portal WikiCI aos princípios determinados pela UX. Como metodologia foi adotada uma análise qualitativa, de caráter exploratório, onde utilizou-se de pesquisa bibliográfica para explanar os conceitos da UX em ambientes digitais, e com isso foi realizada a avaliação dos elementos da WikiCI segundo as facetas da UX. Por meio da análise realizada neste trabalho, observou-se que a aderência da WikiCI aos princípios da UX mostrou-se adequada, pois o website apresenta elementos que podem ser observados nas facetas da UX. Conclui-se que, para os usuários terem uma experiência totalmente satisfatória, é relevante que desenvolvedores de websites atentem-se às questões da AI, e principalmente aos princípios determinados pela UX.

Palavras-chave: Ciência da Informação; Arquitetura da Informação; Wiki; Experiência de Usuário; UX.

Abstract:
Users have been producing more and more content inside of collaborative environments, so it is indispensable that their experience inside these environments is as satisfactory as possible. For this, Information Architecture (AI) together with Information Science (CI) has the purpose to providing the balance of user needs, emerging in the context of user's feeling, their joys and frustrations, issues that can be interpreted by the User Experience (UX). This work has a objective to analyze aspects of AI, specifically regarding the adherence of the WikiCI website to the principles determined by UX. As methodology was adopted a qualitative analysis, of exploratory character, where it used bibliographical research to explain the concepts of UX in digital environments, and it was carried out rating of the elements of the WikiCI according to the facets of UX. Through the analysis conducted in this work, it was observed that the adherence of the WikiCI to the principles of UX proved to be adequate, because the website presents elements that can be observed in the facets of UX. it is relevant that website developers attend to AI questions, and especially to the principles determined by UX.

Keywords: Information Science; Information Architecture; Wiki; User Experience; UX.

1 Introdução

O ambiente Web se tornou uma das mais ricas fontes de informação, o qual disponibiliza conteúdos de diversas naturezas, tais como textos, áudios, vídeos, imagens. Nesse meio de compartilhamento de informações, o usuário está com um papel cada vez mais relevante na construção de conteúdo digital, principalmente com o surgimento de sites de conteúdo colaborativo.

No âmbito de ambientes que proporcionam conteúdo colaborativo, a plataforma mais utilizada é a *wiki,* um ambiente que permite aos usuários acesso rápido a informação por meio do acesso a textos interligados por *links* de hipertextos. Para Tonke (2005), uma *wiki* é um ambiente de trabalho coletivo, na qual os usuários podem inserir, alterar, excluir e acessar informações por meio de uma plataforma digital. Uma das *wikis* que se destaca é

Wikipédia[1], enciclopédia livre e multilíngue, que vem ganhando destaque devido a grande quantidade e qualidade de conteúdos, e que recentemente conta com uma enciclopédia semântica, a *DBpedia*[2].

Baseado em ambientes que proporcionam conteúdo colaborativo e com o intuito de contribuir com meio acadêmico na recuperação de conceitos relacionados à CI, alunos e docentes do Programa de Pós Graduação em Ciência da Informação (PPGCI) da Unesp de Marília-SP, criaram a WikiCI, um *website* que concentra os conceitos publicados pelo programa na área da CI, juntamente com as referências do conteúdo publicado.

A WikiCI é um ambiente em formato de *wiki* que foi construído utilizando a plataforma MediaWiki[3], uma plataforma gratuita, de código aberto, mundialmente utilizada para construir sistemas de informação, tais como enciclopédias (ex.: Wikipédia), documentação de outros sistemas e portfólios, que possam ser construídos a partir de texto, imagens, vídeos, áudios, *links*, hipertexto, etc.

Este ambiente foi construído com o objetivo de apresentar conceitos relacionados à CI, bem como divulgar os conceitos publicados pelos pesquisadores e alunos do PPGCI, possibilitando assim, de maneira simples e direta, a recuperação de conteúdos e definições publicados pela CI para a comunidade.

No entanto, por se tratar de um ambiente que visa atingir a comunidade, é necessário que as questões de organização, representação e uso da informação sejam analisadas neste ambiente e, com isso, a Ciência da Informação (CI) pode contribuir, pois, segundo Le Coadic (1996), a CI é a ciência que estuda as propriedades gerais da informação (natureza, gênese e efeitos) em meio aos processos de geração, distribuição, organização, representação, processamento, comunicação e uso da informação. Saracevic (1995, p. 4) ressalta que a CI possui uma interdisciplinaridade que foi introduzida pelas diferentes experiências daqueles que procuram soluções para problemas e, tais interdisciplinaridades, podem ser mais fortemente percebidas com sua aproximação a Biblioteconomia, Ciência da Computação, a Ciência Cognitiva e a Comunicação.

Assim, a CI pode ter importante papel na construção de uma base teórica e na definição de caminhos para que as novas tecnologias contribuam para o atendimento das necessidades informacionais, já que cabe a esta ciência o papel de investigar o comportamento da informação, seu fluxo e os meios para o seu acesso (BORKO, 1968; CAPURRO, 2003).

Em relação aos aspectos da interação de indivíduos em ambientes digitais, Rosenfeld, Morville e Arango (2015) ressaltam que Arquitetura da Informação (AI) tem a finalidade de proporcionar o equilíbrio das necessidades do usuário, emergindo nesse contexto o sentimento do usuário, suas alegrias e frustrações, essas questões podem ser melhores estudadas dentro de uma das vertentes da AI, denominada de Experiência de Usuário, termo em inglês *User Experience* (UX).

O termo Arquitetura da Informação surge com a preocupação de melhor organizar e disponibilizar informações para determinado público, e com exponencial aumento da criação de websites, este termo torna-se ainda mais relevante. Para Camargo e Vidotti (2011, p. 24), a arquitetura da informação é:

> […] uma área do conhecimento que oferece uma base teórica para tratar aspectos informacionais, estruturais, navegacionais, funcionais e visuais de ambientes informacionais digitais por meio de um conjunto de procedimentos metodológicos a fim de auxiliar no desenvolvimento e no aumento da usabilidade de tais ambientes e de seus conteúdos.

Para uma experiência agradável do usuário em um meio digital, é imprescindível que sua navegação não seja interrompida em momento algum, ou seja, no momento em que o usuário interage com a interface, ele não pode sofrer frustrações. De acordo com a ISO 9241-210, definida pelo padrão

1 Wikipédia, enciclopédia livre e colaborativa. Disponível em: <https://pt.wikipedia.org/>

2 Dbpedia, enciclopédia semântica com dados estruturados baseados na Wikipédia." Disponível em: <http://wiki.dbpedia.org/>

3 MediaWiki software livre e de código aberto. Disponível em: <https://www.mediawiki.org/>

internacional de ergonomia da interação do sistema humano, a UX são “percepções e respostas de uma pessoa, resultantes do uso ou uso antecipado de um produto, sistema ou serviço”. De acordo com a definição da ISO, a experiência do usuário inclui todas as emoções, crenças, preferências, percepções, respostas físicas e psicológicas dos usuários, comportamentos e realizações que ocorrem antes, durante e após o uso.

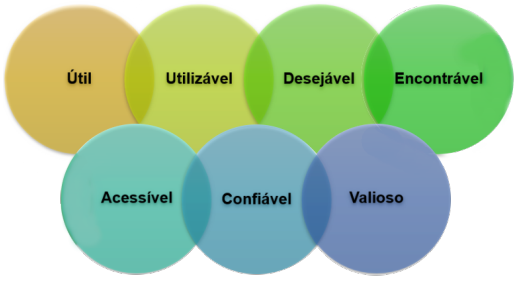
Neste sentido, os estudos sobre UX e usabilidade podem contribuir no desenvolvimento de sistemas de informação, e torna-se indispensável o auxílio de um profissional da arquitetura da informação.

Portanto, pode-se dizer que a experiência do usuário pode ser positiva ou negativa, dependendo de seu sentimento quando o mesmo interage com o ambiente informacional, sendo ele digital ou analógico. Com isso, o planejamento de um ambiente informacional não está relacionado única e exclusivamente com a gestão da informação, e sim também com aspectos que dizem respeito ao design e arquitetura da informação.

Para entender melhor o termo UX, Ferreira et al. (2016) desenvolveram um quadro (Apêndice A) que apresenta definições de autores que trabalham com UX, bem como suas reflexões acerca destas definições em ambientes digitais.

Por meio do quadro apresentado no Apêndice A, conclui-se que a UX pode atender diversos tipos de usuários, tornando-se necessário compreender não somente o sentimento, mas sim a cultura, os hábitos e opiniões dos usuários. Nesse contexto, Morville (2004), destaca alguns requisitos que os serviços devem atender para proporcionar experiências positivas ao usuário, assim, identifica sete facetas ilustradas na Figura 1.

Figura 1 - Facetas da UX



Fonte: MORVILLE (2004, tradução nossa)

Morville (2004) define as facetas como:

- **Útil:** útil ou utilidade, significa que o desenvolvimento de produtos ou sistemas, deve-se prezar pela utilidade, aplicando o conhecimento dos profissionais que desenvolvem o produto ou sistema para soluções inovadoras cada vez mais úteis;
- **Utilizável:** utilizável ou usabilidade, não se refere apenas as facilidade de uso, a usabilidade foca em uma boa interação entre o humano e o computador;
- **Desejável:** despertar o desejo do usuário, buscar pela eficiência, o design deve ser atenuado pela apreciação, poder e valor da imagem, identidade, marca e outros elementos de design emocional;
- **Encontrável:** refere-se à encontrabilidade, criar sites de fácil navegação, com objetos claramente localizáveis, para que usuários possam encontrar o que precisam;
- **Acessível:** refere-se à acessibilidade, os sites devem ser acessíveis e navegáveis por pessoas com e sem deficiência;
- **Confiável:** refere-se à credibilidade, se os usuários confiam e acreditam no ambiente informacional e no conteúdo nele presente.
- **Valioso:** devem fornecer valor aos patrocinadores, para organizações sem fins lucrativos, a experiência do usuário deve colaborar com a ampliação da missão e para as organizações com fins lucrativos, deve proporcionar satisfação ao cliente.

As facetas permitem que o produto ou serviço seja analisado, a fim de identificar pontos críticos, e possíveis melhorias para aprimorar a experiência do usuário no ambiente informacional, para que seu sentimento seja satisfatório e positivo.

Este trabalho utilizará das facetas da UX para analisar a WikiCI , foco de estudo deste trabalho, com o propósito de identificar os elementos da AI que possam ser melhorados

para trazer maior satisfação aos usuários desta *wiki*.

2 Objetivos

Este trabalho tem como objetivo analisar aspectos da Arquitetura da Informação, especificamente em relação à aderência do portal WikiCI aos princípios determinados pela Experiência do Usuário[4].

3 Procedimentos Metodológicos

A metodologia deste trabalho baseou-se em: uma análise qualitativa de caráter exploratório, onde utilizou-se de pesquisa bibliográfica para explanar os conceitos da UX em ambientes digitais, além da avaliação dos elementos da WikiCI segundo as facetas da UX propostas por Morville (2004), e verificação das questões de acessibilidade na faceta "Acessível" por meio da ferramenta *Examinator*.[5].

4 Resultados

A análise da WikiCI com a finalidade de identificar elementos que contemplassem as facetas de UX resultou no quadro apresentado no Apêndice B.

Em relação à faceta Acessível, observa-se que, a WikiCI atende quase todos os requisitos de acessibilidade, de acordo com os padrões analisados e notas alcançadas por meio da ferramenta *Examinator* e, além disso, a WikiCI é responsiva, o que a torna acessível em qualquer dispositivo. Porém, existem alguns pontos que a ferramenta automática não analisa, tais como, múltiplos idiomas, disponibilização de recursos como LIBRAS[6], que quando alcançados permitem uma maior acessibilidade para o usuário. Dentre os problemas de acessibilidade encontrados em *wiki*, *links* com a descrição da legenda igual a descrição do texto é a falha mais comum, isso não é uma falha de navegação mas é um erro grave de acessibilidade, e com certeza causará um sentimento negativo ao usuário, pois a legenda não explica exatamente o que é aquele *link*.

A faceta Confiável é um aspecto relevante em sistemas de informação, pois expressa o quanto aquele sistema é confiável. No caso do Wikipédia, existem muitas criticas em relação à credibilidade de seus conteúdos, uma vez que, qualquer usuário pode alterar o conteúdo das páginas. No entanto, essa deficiência vem mudando com os anos, pois, diversos critérios são analisados para garantir a integridade das informações fornecidas, como citações, validação por revisores de conteúdo, além, de ser possível colocar a referência do conteúdo, o que amplia a credibilidade do ambiente digital. No caso da WikiCI, ela é composta inteiramente de textos confiáveis, tendo em vista que seu conteúdo são citações de trabalhos publicados, o que provê integridade para o conteúdo que é disponibilizado para o usuário.

Sobre a faceta Desejável, a WikiCI permite que o usuário insira nas páginas diversos elementos, tais como imagem, vídeo, áudio, texto, além disso, possui recurso de hipertexto, que permite uma navegação rápida e fácil para usuários. Outro recurso que reflete tranquilidade ao usuário é a ferramenta de busca, pois caso ele não encontre o que deseja de forma rápida, pode optar por utilizar a ferramenta de busca. Assim como qualquer outro sistema, a WikiCI tem sua logomarca, que com suas peças de quebra-cabeça, remete a ideia de algo que pode-se construir de modo colaborativo.

Na faceta Encontrável, o recurso do sistema de busca se torna indispensável, pois permite localizar conteúdos de forma ágil. A WikiCI permite que navegação seja realizada por meio de *links*, que possibilitam uma navegação rápida e simples entre as páginas da *wiki*, assim, o usuário pode navegar da página principal para uma página de conteúdo ou do próprio conteúdo para outras páginas de conteúdo. Além disso, possui *links* externos que direcionam para as fontes das citações, e para a documentação da MediaWiki.

A faceta Útil determina o conteúdo e a apresentação, a WikiCI possui apresentação na primeira página, onde esclarece seus objetivos e disponibiliza uma lista de

4 Experiência do Usuário (UX) de acordo com o termo em inglês *User Experience*

5 *Examinator* - ferramenta de verificação da acessibilidade. Disponível em: <http://examinator.ws/>

6 LIBRAS – Linguagem Brasileira de Sinais

conteúdos para serem acessados, logo, permite o acesso rápido e simples aos assuntos pertinentes à CI.

A faceta Utilizável refere-se à usabilidade, com isso, foi identificado alguns aspectos relevantes da WikiCI, como: o design claro e objetivo, favorecendo a leitura e compreensão do conteúdo; a padronização das páginas de conteúdo, possuindo título, citação e referência. Esta *wiki* ainda conta com páginas de ajuda, com acesso rápido a documentação do *MediaWiki* e, ainda foram criadas páginas de ajuda com imagens e instruções para auxiliar a construção e edição de conteúdo da *wiki* e criação de usuários.

Por fim, em relação à faceta Valioso, a WikiCI visa apresentar conceitos publicados pelos pesquisadores e alunos do PPGCI, facilitando, assim, a recuperação de conteúdos publicados pela Ciência da Informação para a comunidade. Ressalta-se que, a WikiCI não possui fins lucrativos, sendo o cadastro de novos usuários gratuito, portanto, os mantenedores deste *website* não visão lucro, e não tem interesses em propagandas de empresas ou patrocinadores.

Estes resultados demonstram que por meio da utilização das facetas da UX é possível a construção de ambientes informacionais que atendam as necessidades do usuário de forma satisfatória.

5 Considerações Finais

Este trabalho realizou uma análise na WikiCI para verificar aspectos relacionados a experiência do usuário no ambiente, assim, considerou os princípios da AI e os elementos propostos pelas facetas da UX. Por meio da análise realizada neste trabalho, observou-se que aderência da WikiCI aos princípios da UX mostrou-se satisfatória, pois o *website* apresenta elementos que podem ser observados nas facetas da UX.

Sua estrutura de hipertexto se mostra adequada para atender diferentes tipos de sistemas, sendo possível disponibilizar conteúdo informacional digital com a navegação eficiente, o que torna este tipo de ferramenta atrativa e viável para disponibilização de informações, garantindo ainda uma interface amigável ao usuário, que é o principal beneficiário desta ferramenta.

No entanto, com o auxílio da ferramenta *Examinator* notou-se que existem algumas questões de acessibilidade que podem ser aprimoradas, por exemplo, textos que estão justificados ou *links* com a descrição da legenda igual a descrição do texto, sendo que o correto seria descrever na legenda algo explicativo, que facilite o entendimento daquele link.

Conclui-se que, para que o usuário tenha uma experiência totalmente satisfatória, é relevante que desenvolvedores de *websites* atentem-se às questões da AI, e principalmente aos princípios determinados pela UX. Ressalta-se também, a importância do profissional da informação, que pode contribuir para tornar esses ambientes digitais mais úteis e valiosos para usuários, assegurando que a informação esteja sempre disponível e acessível.

Referências

BORKO, Harold. Information science: what is it?. **Journal of the Association for Information Science and Technology**, v. 19, n. 1, p. 3-5, 1968. Disponível em: <http://www.josesales.com.br/arquivos/BORKO%20Harold%20-%20Ci%C3%AAncia%20da%20informa%C3%A7%C3%A3o.pdf >. Acesso em: 10 de jul. de 2016.

CAMARGO, LS de A. VIDOTTI, SABG Arquitetura da Informação: uma abordagem prática para o tratamento de conteúdo e interface em ambientes informacionais digitais. **Rio de Janeiro: LTC**, 2011.

CAPURRO, Rafael. Epistemologia e Ciência da Informação. IN: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 5., 2003. Belo Horizonte. **Anais Eletrônicos... Belo Horizonte: ENANCIB**, 2003. Disponível em: <http://docslide.com.br/documents/capurro-r-epistemologia-e-ciencia-da-informacao-2003.html >. Acesso em: 10 de jul. de 2016.

FERREIRA, AMJF da C, MARTINEZ, SMR, CONEGLIAN, CS, VIDOTTI, SABG, SANTARÉM SEGUNDO, JE. Experiência de Usuário: uma análise do ambiente Wikipédia.

Seminário em Ciência da Informação. **Londrina: SECIN,** 2016. Disponível em: <http://www.uel.br/eventos/cinf/index.php/secin2016/secin2016/paper/view/351/172>. Acesso em Agosto de 2016.

FDIS, IsO. 9241-210: 2009. Ergonomics of human system interaction-Part 210: Human-centered design for interactive systems (formerly known as 13407). **International Organization for Standardization (ISO). Switzerland**, 2009. Disponível em: <https://www.iso.org/standard/52075.html>. Acesso em: 1 de ago. de 2017.

LE COADIC, Yves François. A Ciência da Informação. tradução de Maria Yêda FS de Filgueiras Gomes. Brasília: Briquet de Lemos, 1996. **Perspectivas em Ciência da Informação**, v. 1, n. 2, 1996.

MORVILLE, Peter. User experience design. **Ann Arbor: Semantic Studios LLC**, 2004. Disponível em: <http://semanticstudios.com/user_experience_design/>. Acesso em: 15 de ago. de 2016.

ROSENFELD, L., MORVILLE, P., ARANGO, J. Information Architecture: For the Web and Beyond. **O'Reilly Media, Inc."**, 2015.

SARACEVIC, Tefko. Interdisciplinary nature of information science. **Ciência da informação**, v. 24, n. 1, p. 36-41, 1995. Disponível em:<http://www.brapci.ufpr.br/brapci/_repositorio/2010/03/pdf_dd085d2c4b_0008887.pdf>. Acesso em: 12 de jul. de 2016.

TONKE, E. Making the Case for a Wiki. Ariadne. 42. **Online Journal). Retrieved November**, v. 18, p. 2006, 2005. Disponível em: <www.ariadne.ac.uk/issue42/tonkin>. Acesso em: 18 de ago. de 2016.

Apêndice A - Definições e reflexões de UX para ambientes digitais

Definição UX	Reflexões sobre UX e websites
Experiência do Usuário abrange aspectos da interação do usuário com uma empresa, seus serviços e seus produtos de forma clara e proporcionando sentimentos positivos na utilização dos mesmos. A Experiência do Usuário verdadeira vai além de oferecer aos clientes recursos ou o que eles dizem que querem (NORMAN, 2008).	Os websites devem proporcionar satisfação ao usuário sem que ele perceba ou dizer o que e como deseja.
A UX tem uma abordagem holística e multidisciplinar para o design de interfaces para produtos digitais. Dependendo do produto pode integrar design de interação, design industrial, arquitetura de informação, design de interface visual, design institucional e design centrado no usuário, assegurando a coerência e consistência em todas as dimensões do projeto. A Experiência do Usuário define a forma, comportamento e conteúdo de um produto. (GRABRIEL-PETIT, 2005).	Num website a Experiência do Usuário pode definir de que maneira acontece a interação com o ambiente e a forma da recuperação do conteúdo.
Experiência do Usuário é o modo como um produto funciona no mundo real, como ele funciona na prática, como a pessoa entra em contato com ele e tem que trabalhar com ele. As experiências das pessoas em relação a um mesmo produto são diferentes. (GARRET, 2002)	Os usuários acessam um ambiente digital com autonomia. As formas de interação com o sistema são múltiplas, em que cada pessoa pode interagir da forma que lhe convir.
Uma boa experiência pode ser definida pelo quanto um produto é usável, se ele é funcional (ele faz o que as pessoas esperam que ele faça), eficiente (quão rápido é possível atingir o objetivo sem cometer erros) e desejável (qual a resposta emocional para o produto) (KUNIAVSKY,2003)	Um website funcional poderá satisfazer o usuário, atender a sua necessidade; e responder de maneira eficiente e eficaz.
Descreve a Experiência do Usuário, características de produtos e sistemas, em sete facetas: Útil: diz respeito a grau de utilidade; Utilizável: facilidade de uso; Desejável: envolve os elementos emocionais do usuário; Encontrável: facilidade de localizar o que precisam; Acessível: qualquer usuário pode acessar; Confiável: credibilidade do usuário em relação ao design; Valioso: valor para os patrocinadores. Estas facetas ajudam no desenvolvimento de Web Sites, sempre balanceando o tripé usuário, contexto e conteúdo. (MORVILLE, 2004)	A Experiência do Usuário acontece quando é possível englobar vários elementos voltados à satisfação do usuário, balanceando em um mesmo ambiente o contexto, conteúdo e usuário.

Fonte: FERREIRA et al. (2016, p.328)

Apêndice B – Avaliação dos Elementos da WikiCI segundo as facetas da UX

Faceta	Elemento	Comentário
Acessível	Acessibilidade - Página Inicial	Por meio da ferramenta *Examinator,* foi alcançada nota 8.9 na página inicial, com 12 testes, na qual 10 foram classificados positivamente e 2 negativamente, sendo um dos testes negativos o texto estar justificado e não alinhado para a esquerda ou direita, o texto justificado pode dificultar a leitura para alguns usuários, e o outro teste negativo seria hipertextos com a legenda igual ao texto.
	Acessibilidade - Páginas de Conteúdo	Ainda utilizando a ferramenta *Examinator*, as páginas de conteúdo alcançaram notas entre 9 e 9.7, sendo que em todas as páginas de conteúdo houveram 10 testes classificados positivamente e 1 negativamente.
Confiável	Controle de Conteúdo	Os próprios usuários da *wiki* podem controlar o conteúdo dos demais usuários, abrindo discussões ou editando as páginas.
	Comunidade	A comunidade pode entrar em contato para participar da *wiki* e também para informar sobre conteúdos indevidos ou incorretos.
	Histórico de Alterações	Permite que o usuário visualize o histórico de alterações da página.
	Fontes de informação	Sendo esta uma *wiki* que contém conceitos e definições produzidos pelo PPGCI, todas as citações possuem referência, portanto, caracterizando como 100% autêntico.
Desejável	Sistema de Busca	Possui sistema de busca para facilitar a encontrabilidade do conteúdo.
	Hipertextos	Hipertextos podem ser criados em qualquer conteúdo, permitindo uma navegação rápida entre as páginas da *wiki*. Fora dos menus, toda a navegação é feita por meio de hipertextos.
	Logomarca	A logomarca da WikiCI utiliza de apelo visual peças de quebra-cabeça, que transmitem a ideia de construção e colaboração
	Elementos multimídia	O ambiente *wiki* permite que sejam inseridas imagens, vídeos, áudios, textos, *links*, etc. Porém até o momento a WikiCI possui somente texto, hipertexto e imagens.
Encontrável	Sistema de Busca	Permite localizar de forma rápida conteúdos dentro da WikiCI.
	Links	A navegação é feita por *links* por meio de hipertexto, sendo o fluxo construído da página inicial para as páginas de conteúdos e de uma página de conteúdo para outras páginas de conteúdo.
	Links Externos	Possui *links* externos para fontes de citação e documentação da *wiki* no MediaWiki.

(continua)

(continuação)

Faceta	Elemento	Comentário
Útil	Conteúdo	Conteúdo de fácil acesso, claro e objetivo. Permite obter de maneira rápida e simples conceitos sobre assuntos pertinentes à CI.
	Apresentação	A apresentação é feita na página inicial e também em uma página chamada “Sobre”[7] que mostra informações sobre a WikiCI.
Utilizável	Padrão de Páginas	Todas as páginas de conteúdo seguem um padrão, possuindo título, citação e referência.
	Documentação	Por utilizar a plataforma MediaWiki, utiliza da documentação padrão fornecida por esta ferramenta
	Design	Possui design claro e objetivo, favorecendo a leitura e compreensão do conteúdo.
	Criação e edição de Conteúdo	Foram criadas páginas de ajuda para auxiliar a construção e edição de páginas.
Valioso	Página Inicial	Esclarece de forma sucinta e clara quais são os objetivos da WikiCI.
	Conteúdo	O conteúdo é apresentado ao usuário de forma clara e objetiva,por meio de citação sobre determinado conceito, e no final da página é evidenciada a referência para citação.

Fonte: Elaborado pelos autores**.**

7Sobre, página com informações sobre a WikiCI. Disponível em: <http://codaf.tupa.unesp.br:8081/index.php/WikiCI:Sobre>

PLANO DE GERENCIAMENTO DE DADOS NO CONTEXTO DOS REPOSITÓRIOS DE DADOS DE UNIVERSIDADES

Data Management Plan in the context of university Data Repositories

Elizabete Cristina de Souza de Aguiar Monteiro[1], Ricardo Cesar Gonçalves Sant'ana [(2)]

(1) UNESP, Av. Hygino Muzzi Filho,737, Mirante, Marília-SP, beteaguia@yahoo.com.br
(2) UNESP, Av. Hygino Muzzi Filho,737, Mirante, Marília-SP, ricardosantana@marilia.unesp.br

Resumo:
Plano de Gerenciamento de Dados é o documento formal o qual descreve o conjunto de informações e instruções relacionado à gestão de dados científicos no seu ciclo de vida abordando critérios de coleta, organização, descrição, gerenciamento, disponibilização, acesso e curadoria dos dados tanto pelo pesquisador quanto pelos repositórios de dados. A elaboração do Plano de Gerenciamento de dados auxilia os pesquisadores e os profissionais atuantes nos repositórios. O objetivo deste estudo foi investigar quantos e quais repositórios de dados das 100 melhores universidades do mundo disponibilizam Planos de Gerenciamento de Dados e identificar aspectos relacionados a possíveis benefícios gerados pela adoção destes PGDs. A metodologia utilizada teve como base a pesquisa bibliográfica para a estruturação da fundamentação teórica concomitante à metodologia quantitativa e qualitativa. Foi utilizado o método exploratório para a coleta de dados para fazer o levantamento dos repositórios de dados das 100 melhores universidades do mundo através sítio *webometrics.* Os resultados mostram que, dentre as universidades alencadas, apenas 36 disponibilizam Planos de Gerenciamento de dados e que as instruções do PGD variam dependendo das características dos repositórios e dos conjuntos de dados neles depositados.
Palavras-chave: Plano de Gerenciamento de Dados; Gestão de dados; Dados científicos; Repositório de dados.

Abstract:
Data Management Plan is the formal document which describes the set of information and instructions related to the management of scientific data in its life cycle addressing criteria of collection, organization, description, management, availability, access and curation of the data both by the researcher and data repositories. The elaboration of the Data Management Plan helps the researchers and professionals working in the repositories. The purpose of this study was to investigate how many and which data repositories of the 100 best universities in the world provide Data Management Plans and identify aspects related to the possible benefits generated by the adoption of these PGDs. The methodology used was based on the bibliographical research for the structuring of the theoretical foundation concomitant to the quantitative and qualitative methodology. Was used the exploratory method to collect data to survey the data repositories of the 100 best universities in the world through website webometrics. The results show that, among the enlisted universities, only 36 provide Data Management Plans and that the instructions of the PGD vary depending on the characteristics of the repositories and the datasets deposited in them.
Keywords: Data Management Plan; Data management; Scientific data; Data repository.

1 Introdução

As instituições acadêmicas e científicas passam a ter cada vez mais a responsabilidade no gerenciamento de dados científicos coletados ou produzidos em grande quantidade, velocidade e variedade por pesquisadores nas diversas áreas do conhecimento. A gestão de dados requer, por parte de seus detentores, planejamento e ações concretas que tragam eficiência não só para coleta e armazenamento como também e, principalmente, para fase de

recuperação desses dados ampliando sua visibilidade e potencial uso (SANT'ANA, 2016).

Na ambiência da investigação científica, esse processo que agrega valor configura-se como importante fator para a ampliação do potencial de impacto dos resultados das pesquisas e da própria instituição (PRYOR, 2012)[1].

Desse contexto emerge a necessidade de políticas para o gerenciamento dos dados envolvidos nas pesquisas. Agências de fomento como a *National Science Foundation* (NSF), *National Institutes of Health* (NIH), *National Oceanographic Data Center* (NODC) e Nasa dos Estados Unidos, Horizon2020 da Europa, AHRC, BBSrc, *Cancer Research* UK, EPSRC, ESRC, MRC, NERC, STFC, *WELLCOMETrust* no Reino Unido estão incentivando, orientando ou mesmo tornando obrigatório a elaboração de Plano de Gerenciamento de Dados (PGD) para os projetos que terão o financiamento de suas pesquisas por essas agências (CORRÊA COUTO, 2016).

Um objetivo de destaque na gestão de dados científicos é assegurar que os mesmos possam ser compreendidos e interpretados por outros pesquisadores ao longo do tempo. Para isso é essencial uma descrição clara e detalhada dos dados, anotações adicionais e informações que contextualizam os dados e possibilitem que transmitam informação e conhecimento no tempo e no espaço (SAYÃO; SALES, 2015).

Os procedimentos descritos para a gestão de dados são documentados em um Plano de Gerenciamento de Dados (PGD), documento direcionado àqueles que estão envolvidos de alguma forma com a gestão desses dados (MONTEIRO, 2017). O PGD auxilia tanto os pesquisadores que coletam e manipulam conjuntos de dados quanto àqueles profissionais que atuam nos repositórios de dados científicos. Couto Corrêa (2016) complementa destacando que o PGD fornece diretrizes para todo o ciclo de vida dos dados.

O Plano de Gerenciamentos de Dados é um plano que descreve diferentes atividades e processos associados ao ciclo de vida de dados e envolve

> [...] a concepção e criação de dados, armazenamento, segurança, preservação, recuperação, partilha e reutilização, todos tendo em conta as capacidades técnicas, considerações éticas, questões legais e estruturas de governança. (COX; PINFIELD, 2014, tradução nossa).

Os procedimentos adotados na execução de um PGD definem e estabelecem métodos de execução das atividades e detalham os procedimentos que serão realizados. O planejamento é um processo cíclico, dinâmico e interativo, em que as fases não precisam ser lineares, pois há uma dinâmica no processo (ALMEIDA, 2005).

Conjuntos de dados de um determinado grupo de pesquisadores podem conter diferentes formatos, tipos e descrições, tornando-os altamente heterogêneos e, à medida que o tamanho dos conjuntos de dados aumenta, o seu gerenciamento tende a se tornar árduo (LEE et al., 2009).

Para contextualizar a coleta de dados, este trabalho utilizou o Ciclo de Vida dos Dados (CVD) (SANT'ANA, 2016), modelo composto por quatro fases: Coleta, Armazenamento, Recuperação e Descarte, sobre as

[1] conceito de Pryor retirado de informações referente ao livro PRYOR, G. (Ed.) **Managing research data**. United Kingdon: Facet Publishing, 2012. Disponível em: <http://www.dcc.ac.uk/news/book-managing-research-data>. Acesso em: 27 set. 2016.

quais perpassam por seis fatores como: Preservação, Disseminação, Direitos Autorais, Qualidade, Integração e Privacidade (Apêndice A).

A fase da coleta caracteriza o processo de obtenção dos dados. No contexto da coleta de dados científicos, participam diversos atores entre eles: pesquisador 1, detentor de dados (profissional responsável pelos dados no repositório), sociedade (pesquisadores 2, 3 e n que farão coleta nos repositórios).

A fase Coleta aparece tanto no momento do depósito dos dados no repositório pelo Pesquisador 1 (CVD - Repositório), quanto no momento da coleta dos pesquisadores 2, 3 e n (CVD - Pesquisador) quando coleta seus dados para sua pesquisa no repositório, conforme demonstrado no Apêndice B.

Desafios no âmbito da Ciência da Computação e da Ciência da Informação, tais como àqueles que ocorrem em todas as fases do CVD, Preservação, Disseminação, Direitos Autorais, Qualidade, Integração e Privacidade, permanecem em aberto o que torna difícil descobrir, compartilhar ou reutilizar dados pois:

1) dados valiosos podem ter sido descartados;
2) tecnologias da informação tendem a ter processo de obsolescência altamente acelerado;
3) formatos incompatíveis podem tornar os dados difíceis ou impossíveis de integrar;
4) o fluxo de dados entre domínios pode ser impedido por metadados incompletos, imprecisos e/ou mal descritos;
5) muitos cientistas relutam em compartilhar dados devido à falta de recompensa, às questões de propriedade intelectual e documentação apropriada (LEE et al., 2009).

Geralmente a instituição que implementa repositório de dados tem em seu sítio um documento ou informações relacionadas ao plano de gestão de dados para orientar os que depositarão conjuntos de dados na elaboração de PGDs. Cada fase e fator do Ciclo de Vida dos Dados devem ser considerados na elaboração do PGD.

Os Repositórios de dados científicos são ambientes digitais implementados nas universidades com infraestrutura para dar suporte aos pesquisadores no gerenciamento e na disponibilização de dados científicos facilitando a outros pesquisadores reutilizá-los (MONTEIRO, 2017).

Repositórios de dados científicos contribuem no gerenciamento de grandes quantidades de dados. Os repositórios de dados são mantidos por conjuntos de ações que viabilizam o armazenamento de dados visando à otimização da coleta pelos pesquisadores, o que amplia as potencialidades de reuso destes dados (MONTEIRO, 2017).

2 Objetivos

Os objetivos deste estudo foram investigar Repositórios de Dados das 100 melhores universidades do mundo para verificar quantos e quais disponibilizam em seus sítios Plano de Gerenciamento de Dados e identificar aspectos relacionados a possíveis benefícios gerados pela adoção destes PGDs.

3 Procedimentos Metodológicos

A metodologia utilizada teve como base a pesquisa bibliográfica concomitante à metodologia quantitativa e qualitativa. Foi utilizada a coleta de dados para o levantamento dos repositórios de dados das 100 melhores universidades do mundo. A coleta de dados se iniciou com a busca das melhores universidades do mundo por meio do ranking *webometrics.info.* definindo o escopo com as 100 melhores ranqueadas. A localização dos repositórios de dados nas universidades foi realizada nos meses de julho a setembro de 2016. Em seguida foi

realizada a pesquisa exploratória para o levantamento das páginas oficiais das universidades identificadas para localização dos repositórios de dados. Não foram analisados repositórios com acesso restrito ou com link quebrado. O processo de recuperação dos dados foi realizado por meio de coleta dos Planos de Gerenciamento de Dados dos repositórios de dados encontrados.

4 Resultados

A análise incluiu a identificação dos repositórios de dados das universidades e a identificação dos Planos de Gerenciamento de Dados.

O Apêndice C ilustra os caminhos que foram seguidos para a coleta nos repositórios de dados. Os resultados apontaram que 55 universidades dispõem de repositórios de dados. Dessas, 36 têm PGD, os quais forma analisados.

Os repositórios das universidades que têm PGD são: *Harvard University, Massachusetts Institute of Technology, Stanford University, University of California Berkeley*, *University of Michigan*, *University of Washington,* University of Wisconsin Madison*, University of Pennsylvania, University of Oxford, Yale University*, *University of Cambridge, Michigan State University*, *University of Texas Austin*, *University of California San Diego, Pennsylvania State University*, *University of Illinois Urbana Champaign*, *University of North Carolina Chapel Hill*, *Princeton University*, *University College London*, *University of British Columbia*, *University of Maryland Baltimore*, *Purdue University*, *California Institute of Technology Caltech, University of Virginia*, *University of California Irvine, University of Arizona*, *University of Edinburgh, Washington University Saint Louis*, *Simon Fraser University*, *Virginia Polytechnic Institute and State University*, *Tufts University, Ruprecht Karls Universität Heidelberg*, *University of Copenhagen*, *University of Amsterdam*, *Universiteit Utrecht*, *University of California Los Angeles UCLA.*

Os repositórios de dados documentam as instruções e normativas nos PGDs no qual mencionam vários aspectos do Ciclo de Vida dos Dados. As instruções inclusas no PGD variam dependendo das características dos repositórios e dos conjuntos de dados neles depositados.

Percebeu-se que cada repositório elaborou seu PGD de acordo com as necessidades e particularidades de sua comunidade e do tipo de conteúdo abordado nos conjuntos de dados.

Para auxiliar os pesquisadores na elaboração de Planos de Gerenciamento de Dados incluindo os requisitos necessários para tal, identificou-se o DMPtool[2] que é uma ferramenta da Universidade da Califórnia e que fornece orientações sobre instituições financiadoras específicas que exigem PGD e um guia para a elaboração do documento.

Os PGDs tem sua propriedade intelectual vinculada a quem os criou. O pesquisador que elabora o PGD no DMPtool pode optar em compartilhar seu PGD publicamente contribuindo com outros pesquisadores (DMPtoll, 2017).

Os usuários do DMPTool podem visualizar amostras de planos, requisitos das agências financiadoras e exibir as alterações mais recentes feitas em seus planos uma vez que permite ao usuário criar um documento editável para apresentar a uma agência de financiamento. Pode, ainda acomodar versões diferentes à medida que os requisitos de financiamento mudam (DMPtool, 2017).

O uso de PGD pelos pesquisadores e por repositórios de dados podem proporcionar benefícios a todos os envolvidos, pois orienta sobre vários aspectos conforme descrito nos vários repositórios pesquisados:

[2] https://dmptool.org/

- Fornece opções de acesso flexíveis: os dados ficam acessíveis a todos, ou com acesso restrito mediante solicitação, dependendo das opções do pesquisador;
- Os dados recebem um URL permanente sob a forma de um identificador de objeto digital (DOI) para que o pesquisador possa conectar seus dados para suas publicações;
- Acesso a longo prazo: identificadores persistentes e DOIs tornam mais fáceis aos pesquisadores localizarem e citarem os dados;
- Exemplos de como citar os conjuntos de dados fator que indica aspectos relacionados aos direitos autorais;
- Indicação de quais licenças estão atribuídas aos conjuntos de dados recomendando como os dados podem ser utilizados;
- Análise quantitativa: mostra com que frequência os dados são vistos e feito *download*;
- Maximiza a reutilização: pode ter consulta com o pesquisador para garantir que os dados estejam em um formato e estrutura que melhor facilite o acesso a longo prazo, descoberta e reutilização;
- Com o uso de padrão de metadados, os dados podem ser indexados pelo *Google*, o que aumenta a possibilidade de outros pesquisadores encontrarem os dados;
- O servidor possui serviço de *backups* e manutenção regular para evitar a perda de dados.
- Nos sítios contém indicações de ferramentas que auxiliam os autores a montarem seus PGDs;
- Indicações de quais licenças estão atribuídas aos conjuntos de dados, o que determina como os dados são licenciados e as formas de utilização.
- Pode auxiliar o pesquisador que vai depositar os dados atender aos requisitos das agências de fomento que se aplicam aos seus conjuntos de dados;
- Orienta sobre aspectos de privacidade dos dados, os quais devem ser anonimizados para que, quando compartilhados não ameaçarem a privacidade dos sujeitos referenciados.

Os procedimentos adotados na execução de um PGD são instrumentos que definem e estabelecem métodos de execução das atividades e detalham a forma exata pela qual os procedimentos serão realizados.

A preparação de um PGD envolve atividades em diferentes graus de formalidade, extensão, periodicidade, metas e objetivos. O desenvolvimento de seu conteúdo envolve as particularidades de cada área abrangida pelo repositório.

As instruções inclusas no PGD variam dependendo dos objetivos e das características dos repositórios e dos conjuntos de dados neles depositados. Essas instruções devem ser claras para não gerar dúvidas e inseguranças.

4 Considerações Finais

Plano de Gerenciamento de Dados é um documento que contribui para o desenvolvimento e gerenciamento dos dados nos repositórios com instruções aos pesquisadores no gerenciamento dos dados desde a coleta e aos profissionais que neles trabalham orientando no gerenciamento por meio de diretrizes para coleta, armazenamento, recuperação e descarte, contribuindo ainda, no atendimento de requisitos relacionados a privacidade, qualidade, integração, disseminação, direitos autorais e preservação dos conjuntos de dados.

Os resultados da análise demonstram que das 100 melhores universidades do mundo, apenas 36 delas disponibilizam PDGs. Quando

considerado que foram analisadas as melhores universidades do mundo, esse fator comprova que ainda se tem um longo caminho para a conscientização, por parte dos envolvidos na gestão dos dados, da importância do PGD para gerenciamento de dados científicos.

A gestão de dados é imprescindível para o bom andamento da pesquisa, porém, dentre as 55 universidades com Repositório de Dados, em 19 delas não foram localizados PGDs. As universidades implementaram o repositório de dados, no entanto, a gestão de dados ainda não está explicitamente evidenciada.

As vertentes apresentadas corroboram que a área da Ciência da Informação, por meio do seu arcabouço teórico e prático pode contribuir com a implementação de repositórios de dados ampliando a atenção dada à gestão de dados por meio do estudo e fomento da utilização de PGDs.

Referências

ALMEIDA, M. C. B. **Planejamento de bibliotecas e serviços de informação**. Brasília, DF: Brinquet de Lemos, 2005.

COX, A. M. PINFIELD, S. Research data management and libraries: current activities and future priorities. **Journal of Librarianship and Information Science**, London, v. 46, n. 4, p. 299-316, 2014. Disponível em: <http://lis.sagepub.com/content/46/4/299.full.pdf+html>. Acesso em: 27 set. 2016.

COUTO CORRÊA, F. Gestíon de datosde investigación. Barcelona: Editorial UOC, 2016. Disponível em: <http://bit.ly/2uwefAX>. Acesso em: 2 jul.2017.

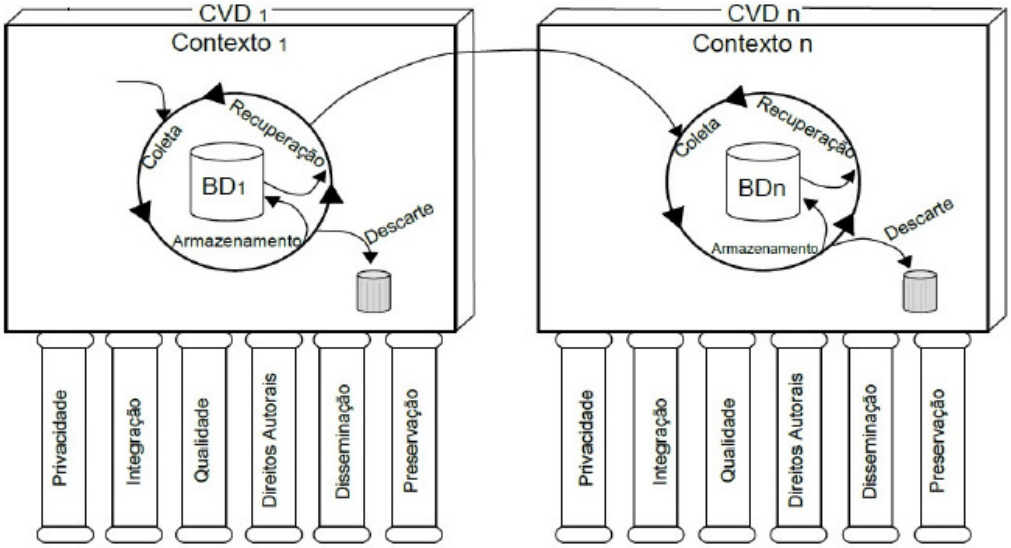
LEE, J. W. et al. DataNet: an emerging cyberinfrastructure for sharing, reusing and preserving digital data for scientific discovery and learning. **AIChe Journal,** New York, v. 55, n. 11, p. 2757-2764, Nov. 2009. Disponível em: <http://onlinelibrary.wiley.com/doi/10.1002/aic.12085/epdf >. Acesso em: 05 jan. 2017.

MONTEIRO, E. C. S. A. **Direitos autorais nos repositórios de dados científicos: análise sobre os planos de gerenciamento dos dados.** 2017. 115 f. Dissertação (Mestrado em Ciência da Informação) – Faculdade de filosofia e Ciências, Universidade Estadual Paulista, Marília, 2017. Disponível em: <http://hdl.handle.net/11449/149748>. Acesso em: 30 abr. 2017.

SANT'ANA, R. C. G. Ciclo de vida dos dados: uma perspectiva a partir da ciência da informação. **Informação e informação**, Londrina, v. 21, n. 2, p. 116-142, maio/ago. 2016. Disponível em: <http://www.uel.br/revistas/uel/index.php/informacao/article/view/27940/20124 >. Acesso em: 20 out. 2016.

SAYÃO, L. F.; SALES. L. F. **Guia de gestão de dados de pesquisa para bibliotecários de pesquisadores**. Rio de Janeiro: CNEN, 2015. Disponível em: <http://carpedien.ien.gov.br:8080/bitstream/ien/1624/1/GUIA_DE_DADOS_DE_PESQUISA.pdf>. Acesso em: 5 out. 2016.

Apêndice A – Ciclo de Vida dos Dados para a Ciência da Informação (CVD-CI)

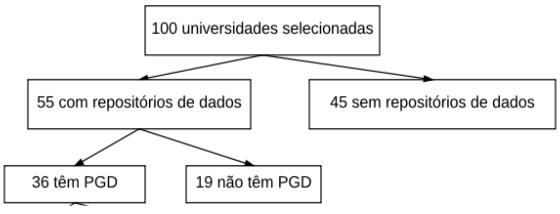


Fonte: SANT'ANA, 2016

Apêndice B – Ciclo de Vida dos Dados no Repositório

Fonte: MONTEIRO, 2017

Apêndice C – Direcionamento das análises



Fonte: Dados da pesquisa, 2017

REPOSITÓRIO DE DADOS CIENTÍFICOS: aspectos sobre privacidade de dados

Scientific Data Repositories: aspects about data Privacy

Elizabete Cristina de Souza de Aguiar Monteiro[1], Elaine Parra Affonso[2], Victor Ubiracy Borba[3]; Ricardo César Gonçalves Sant'Ana[4]

(1) UNESP, Av. Hygino Muzzi Filho,737, Mirante, Marília-SP, beteaguia@yahoo.com.br
(2) UNESP, Av. Hygino Muzzi Filho,737, Mirante, Marília-SP, elainepff@gmail.com
(3) UNESP, Av. Hygino Muzzi Filho,737, Mirante, Marília-SP, borba.victor.borba@gmail.com
(4) UNESP, Av. Hygino Muzzi Filho,737, Mirante, Marília-SP, ricardosantana@marilia.unesp.br

Resumo:
Repositórios de dados científicos são ambientes digitais implementados nas universidades para auxiliar pesquisadores no gerenciamento, disponibilização e acesso a dados científicos, contribuindo com a sua reutilização. Aspectos sobre privacidade de dados dos sujeitos referenciados nas pesquisas devem estar presentes no Plano de Gerenciamento de Dados (PGD), tanto de pesquisadores quanto nos disponibilizados pelos repositórios. O objetivo deste trabalho foi analisar repositórios de dados de universidades para identificar aspectos de privacidade. Para tanto, foi verificado se há menção sobre aspectos de privacidade nos PGDs, e evidenciada as medidas de privacidade propostas. A metodologia utilizada foi quantitativa e qualitativa com o método exploratório, analisando os PGDs das universidades. Os resultados demonstraram que a maioria das universidades com repositórios, mencionam em seus PGDs medidas para proporcionar privacidade de dados, tais como: consentimento informado, aderência às normas da *Health Insurance Portability and Accountability Act* (HIPAA) e supressão de identificadores pessoais. Embora haja menção sobre a necessidade de proteger dados pessoais e evitar ameaças à privacidade dos sujeitos referenciados nas pesquisas, técnicas para anonimização nem sempre estão detalhadas nos PGDs, podendo deixar dúvidas sobre como realizar tais procedimentos, visto que, essas técnicas são fundamentais para preservar a identidade dos participantes das pesquisas e garantir aspectos éticos.

Palavras-chave: Repositório de dados; Privacidade de dados; Anonimização de dados.

Abstract:
Scientific data repositories are digital environments implemented in universities to help researchers in management, availability and access to scientific data, contributing to their reuse. Aspects about data privacy of the subjects referenced in the research should be present in the Data Management Plan (PGD), both from researchers and available by from repositories. The purpose of this work was to analyze university data repositories to identify aspects of privacy. For this, it has been verified if there exists a mention of privacy aspects in PGDs, and the proposed privacy measures are highlighted. The methodology used was quantitative and qualitative with the exploratory method, analyzing the PGDs of the universities. The results shows that most universities with repositories, mention in their PGDs measures to provide data privacy, such as: informed consent, adherence to standards of Health Insurance HIPAA and Accountability Act (HIPAA) and suppression of personal identifiers. Although exists a mention about the need to protect personal data and avoid threats to the privacy of the person referenced in research, techniques for anonymization aren't always detailed in PGDs, and may leave doubts about how to do such procedures, since these techniques are fundamental to preserve the research participants identity and to ensure ethical aspects.

Keywords: Data repository; Data privacy; Data anonymization.

1 Introdução

Os Repositórios de dados científicos são ambientes implementados nas universidades com infraestrutura para dar suporte aos pesquisadores no gerenciamento e na disponibilização de dados científicos e, dessa forma, ampliar o acesso para que outros pesquisadores possam reutilizá-los (MONTEIRO, 2017).

Rodrigues et al. (2010, p. 22-23, grifo nosso) contextualizam repositório de dados como uma extensão de repositórios

> [...] "repositório" designa um sistema informático em que existe uma plataforma de armazenamento de objectos representados em ficheiros, capaz de incorporar novos objectos à medida que são produzidos ou submetidos. O repositório oferece serviços que são dirigidos a quem deposita, a quem pesquisa e aos administradores do sistema. Nos **repositórios de dados** pode ir-se muito além desta visão de repositório de objectos, uma vez que cada conjunto de dados tem características próprias e por isso pode requerer um tratamento diferenciado.

A ambiência de repositórios de dados viabiliza armazenar, representar, gerenciar, disseminar, disponibilizar, e preservar dados neles depositados. Reunir conjuntos de dados nesses repositórios propicia compartilhamento, acesso e reuso de dados entre pesquisadores. As atividades inerentes ao gerenciamento de dados são documentadas no Plano de Gerenciamento de Dados (PGD).

O PGD é um documento ou conjuntos de instruções que orientam àqueles que estão envolvidos com a gestão de dados científicos (MONTEIRO, 2017). Tanto o pesquisador quanto o repositório dispõe de PGD. O pesquisador elabora seu PGD no início de sua pesquisa para gerenciar seus dados. Do mesmo modo, os repositórios de dados disponibilizam PGDs para orientar pesquisadores que vão depositar seus dados e profissionais atuantes no repositório.

Questões de privacidade são fatores preponderantes a serem registrados no PGD, pois, cada vez mais é exigido pelas instituições de pesquisa e agências de fomento que o próprio pesquisador formalize no PGD seu compromisso sobre questões éticas e de privacidade, como os procedimentos para garantir proteção dos dados pessoais, principalmente em relação ao compartilhamento de dados sensíveis.

Assim, estratégias como anonimização de dados e técnicas de criptografia devem ser adotadas pelos profissionais que detém esses dados.

Sayão e Sales (2016, p. 70) ao falarem sobre curadoria digital e dados de pesquisa, ressaltam que "[...] existe uma preocupação forte com questões éticas, de privacidade e de propriedade intelectual [...]".

Tendo em vista a necessidade de proteger dados pessoais quando esses são resultados de pesquisas científicas, torna-se relevante descrever questões e atores envolvidos na fase de coleta de dados, considerando tanto o Ciclo de vida dos dados do pesquisador, quanto o Ciclo de vida de dados do repositório, incluindo as estratégias e verificação dos tipos de dados envolvidos no processo de proteção de dados pessoais.

Em relação aos tipos de dados envolvidos nas questões de proteção da privacidade, esses podem ser classificados em: identificadores, semi-identificadores, atributos sensíveis, e atributos não sensíveis (CIRIANI et al., 2009; DE CAPITANI DI VIMERCATI et al., 2012).

Dados denominados identificadores caracterizam-se por identificar unicamente os indivíduos no conjunto de dados (ex.: CPF, nome, número da Identidade, número de Matrícula) (CIRIANI et al., 2009), e são os

primeiros as serem evidenciados e protegidos quando a finalidade é garantir a privacidade dos indivíduos referenciados nos conjuntos de dados que serão disponibilizados (SAMARATI; SWEENEY, 1998).

Atributos semi-identificadores são aqueles que caracterizam-se por conterem valores que, quando correlacionados e/ou combinados com dados externos, podem proporcionar a identificação do indivíduo e, desta forma, vincular o indivíduo a seus dados confidenciais. Podem ser considerados dados semi-identificadores: data de nascimento, CEP, cargo, função, dados de localização, entre outros (CIRIANI et al., 2009).

Para Sweeney (2002) a divulgação de atributos semi-identificadores deve ser realizada de forma cautelosa, pois, por meio deles, é possível a identificação do sujeito no conjunto de dados.

Os atributos sensíveis são aqueles que representam os dados confidenciais (ex.: doenças, salário, exames médicos, lançamentos de cartão de crédito) que quando expostos podem colocar o indivíduo em situações constrangedoras e, quando não causam ameaças, são denominados atributos não sensíveis (DE CAPITANI DI VIMERCATI et al., 2012).

Durante a investigação científica, o pesquisador coleta dados que podem abranger dados identificadores (nome, CPF), semi-identificadores (data de nascimento, endereço, CEP) e sensíveis (doenças, religião, salário).

No processo da descrição dos procedimentos e diretrizes da gestão dos dados no PGD, o pesquisador pode, se oportuno, solicitar consentimento dos participantes para compartilhamento e uso a longo prazo de dados confidenciais. Logo, é adequado definir e descrever no PGD qual nível de confidencialidade será mantido (MONTEIRO, 2017). Esse consentimento também ajudará o gestor do repositório na disponibilização dos dados.

Além do consentimento do usuário, técnicas de anonimização são relevantes para que o conjunto de dados possa ser compartilhado sem que ocorram ameaças a privacidade dos participantes das pesquisas. Affonso, Oliveira e Sant'Ana (2017) ressaltam que, por meio de técnicas de anonimização, tais como, supressão, generalização, adição de ruídos ou troca de dados (swapping), é possível obter um conjunto de dados anonimizados, que quando disponibilizado, permite acesso aos dados do sujeito, mantendo protegida a sua identidade e minimizando ameaças a privacidade.

Portanto, o conjunto de dados coletados pelo pesquisador, poderá ter dados que contextualizam informações vinculadas a um indivíduo, como dados provenientes dos seus atos, consumo, manifestações e opiniões. Desta forma, esse conjunto de dados, quando compartilhado sem devidas precauções, pode ameaçar a privacidade dos sujeitos referenciados nesses conjuntos de dados.

Para contextualizar a coleta de dados, este trabalho utilizou o Ciclo de Vida dos Dados (CVD) (SANT'ANA, 2016), considerando o fator privacidade. O CVD é um modelo composto por quatro fases: Coleta, Armazenamento, Recuperação e Descarte, sobre as quais permeiam seis fatores: Preservação, Disseminação, Direitos Autorais, Qualidade, Integração e Privacidade (Apêndice A).

A fase da coleta configura o processo de obtenção dos dados. Nessa fase têm-se as atividades

> [...] vinculadas a definição inicial dos dados a serem utilizados, seja na elaboração do planejamento de como serão obtidos, filtrados e organizados, identificando-se a estrutura, formato e meios de descrição que será utilizado. (SANT'ANA, 2013, p. 18).

No contexto da coleta de dados científicos, participam os atores: sujeito alvo (participante da pesquisa), pesquisador, detentor de dados (profissional responsável pelos dados no repositório), comitê de ética, e sociedade (pesquisadores que farão coleta nos repositórios). A fase de coleta acontece tanto no momento da coleta do pesquisador (CVD Pesquisador) quando coleta seus dados para sua pesquisa, quanto no momento do depósito desses dados no repositório (CVD - Repositório). Quando o pesquisador deposita os dados nos repositórios para serem disponibilizados à sociedade, os dados precisam ser anonimizados para que não ocorra ameaça à privacidade dos sujeitos referenciados no conjunto de dados (Apêndice B).

Ressalta-se que no processo de coleta de dados (Apêndice B), devem-se levar em consideração as mesmas estratégias de anonimização de dados para futura disponibilização, tanto em relação ao pesquisador, quanto em relação ao repositório. Sendo assim, o detentor de dados do repositório deve garantir que os dados que serão disponibilizados estejam sob medidas de privacidade, e nos PGDs devem estar explícitas tais medidas.

2 Objetivos

O objetivo deste trabalho foi analisar os repositórios de dados das universidades para identificar aspectos de privacidade de dados na fase de coleta do repositório. Para tanto, buscou-se especificamente: Verificar se há menção sobre aspectos de privacidade nos PGDs; e evidenciar as medidas de privacidade propostas no PGD dos repositórios identificados.

3 Procedimentos Metodológicos

Utilizou-se a metodologia quantitativa e qualitativa. Foi realizada coleta de dados para levantamento dos repositórios de dados das 100 melhores universidades do mundo por meio do ranking *webometrics.info,* definindo o escopo com as 100 melhores ranqueadas. A identificação dos repositórios de dados nas universidades foi realizada nos meses de julho a setembro de 2016.

Em seguida foi realizada pesquisa exploratória para levantamento das páginas oficiais das universidades, para localização dos repositórios de dados. Não foram analisados repositórios com acesso restrito ou com link quebrado. O processo de recuperação dos dados foi realizado por meio de coleta dos PGDs dos repositórios de dados encontrados, verificando menção às questões de privacidade de dados.

4 Resultados

A análise incluiu a identificação dos repositórios de dados das universidades e a identificação dos PGDs, baseando-se na fase de Coleta com o fator Privacidade do CVD do repositório.

As análises demonstraram que: 55 universidades dispõem de repositórios de dados. Dessas, 36 têm PGD. Das universidades que possuem PGD, 78% mencionam aspectos de privacidade nos seus PGDs.

Em relação aos aspectos de privacidade mencionados nos PGDs dos repositórios analisados, elencam-se as seguintes medidas para garantir a proteção de dados dos participantes de pesquisas científicas:

- Consentimento informado;
- Alinhamento da coleta de dados realizada pelo pesquisador de acordo com a política de dados da *Health Insurance Portability and Accountability Act* (HIPAA)[1];
- Aderência a *Family Education Rights and Privacy Act* (FERPA)[2];

[1] Regras para garantir a privacidade de dados pessoais de saúde e seu acesso por profissionais de saúde e outros.

[2] Leis que protegem os dados dos estudantes e seu acesso pelos pais, escolas e outros.

- Supressão de dados identificadores e dados sensíveis;
- Generalização de dados semi-identificadores com a finalidade de minimizar a correlação de dados, pois, por meio dessa técnica, é possível tornar os dados menos específicos, aumentando a quantidade de dados similares no conjunto de dados;
- Uso de técnicas de perturbação para esconder/mascarar dados sensíveis;
- Criptografia de dados;
- Uso de *checklist* para o pesquisador verificar se realizou anonimização de dados e consentimento informado;
- Instrução para que o pesquisador não deposite dados sensíveis no repositório;
- Anonimização de dados seguindo o protocolo do *Institutional Review Board* (IRB)[3];
- Disponibilização de termos de uso e código de conduta sobre uso de dados pessoais e segurança de dados;
- Instruções para que pesquisadores identifiquem dados identificadores, semi-identificadores e sensíveis;
- Armazenamento separado para dados sensíveis;

Embora 78% dos repositórios analisados apresentam em seus PGDs menções às questões de privacidade, observa-se que essas, muitas vezes, não são detalhadas, e não estão explícitas como é realizada as medidas de proteção da privacidade, ou como o pesquisador deverá proceder para realizar anonimização dos dados. Ressalta-se ainda que, três repositórios apenas citam a necessidade de proteção de dados pessoais, no entanto, não apresentam políticas ou medidas para proteção de dados pessoais.

4 Considerações Finais

A precaução dos repositórios relacionada às questões da privacidade dos dados, principalmente dados que envolvem humanos, é evidente na maioria deles. As diferentes medidas indicadas em cada repositório são evidenciadas nos PGDs como uma forma de assessorar os pesquisadores na liberação de seus conjuntos de dados envolvendo dados sensíveis, ponderando os diversos aspectos relacionados a manter a privacidade dos envolvidos na pesquisa e assegurando questões éticas.

Os profissionais que atuam nos repositórios de dados devem estar cientes dos vários aspectos descritos nos PGDs para garantir a privacidade dos dados arquivados, considerando as diretrizes elencadas.

Os pesquisadores devem distinguir as diferentes medidas e técnicas necessárias para proteger a privacidade dos indivíduos e deverão ter cautela no momento da disponibilização de dados sensíveis e dados que podem ser correlacionados com outras bases de dados, tal como, os dados semi-identificadores.

As técnicas utilizadas para anonimização dos dados e medidas para proteção de dados pessoais preservam a identidade dos indivíduos participantes da pesquisa, asseguram ao pesquisador os aspectos éticos e direcionam os profissionais dos repositórios na gestão dos dados que ficam disponíveis para sociedade.

Ainda que as questões de privacidade estejam mencionadas nos repositórios, observou-se que em muitos PGDs as medidas para proteger dados pessoais se apresentam vagas, sem muitos detalhes de como proceder para atingir a anonimização de dados

[3] IRB é um órgão administrativo estabelecido para proteger os direitos, o bem-estar e aspectos sobre privacidade dos sujeitos humanos participantes de pesquisas.

antes de depositá-los no repositório, o que pode ocasionar problemas éticos e de exposição dos participantes das pesquisas. Esse cenário revela a importância de estudos dos fatores envolvidos no compartilhamento de dados de pesquisas e as medidas para proteção da privacidade.

Referências

AFFONSO, E. P.; DE OLIVEIRA, S. C; SANT'ANA, R. C. G. Análise do equilíbrio entre privacidade e utilidade no acesso a dados. **Informação & Sociedade**, v. 27, n. 1, 2017. Disponível em: <http://www.ies.ufpb.br/ojs/index.php/ies/article/view/29422>. Acesso em: 17 jun. de 2017.

CIRIANI, V. et al. Theory of privacy and anonymity. **Algorithms and theory of computation handbook**, 2009.

DE CAPITANI DI VIMERCATI, S. et al. Data privacy: definitions and techniques. **International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems**, v. 20, n. 06, p. 793-817, 2012. Disponível em: <https://www.semanticscholar.org/paper/Data-Privacy-Definitions-and-Techniques-Vimercati-Foresti/7c6abddbd791dddd281c5764dbe859c55ba2e019/pdf>. Acesso em: 10 de jun. de 2016.

MONTEIRO, E. C. S. A. **Direitos autorais nos repositórios de dados científicos: análise sobre os planos de gerenciamento dos dados.** 2017. 115 f. Dissertação (Mestrado em Ciência da Informação) – Faculdade de filosofia e Ciências, Universidade Estadual Paulista, Marília, 2017. Disponível em: <http://hdl.handle.net/11449/149748>. Acesso em: 30 abr. 2017.

RODRIGUES, E. et al. **Os repositórios de dados científicos: estado da arte**. 2010. Disponível em: <http://projeto.rcaap.pt/index.php?option=com_remository&Itemid=2&func=startdown&id=271&lang=pt>. Acesso em: 5 jun. 2016.

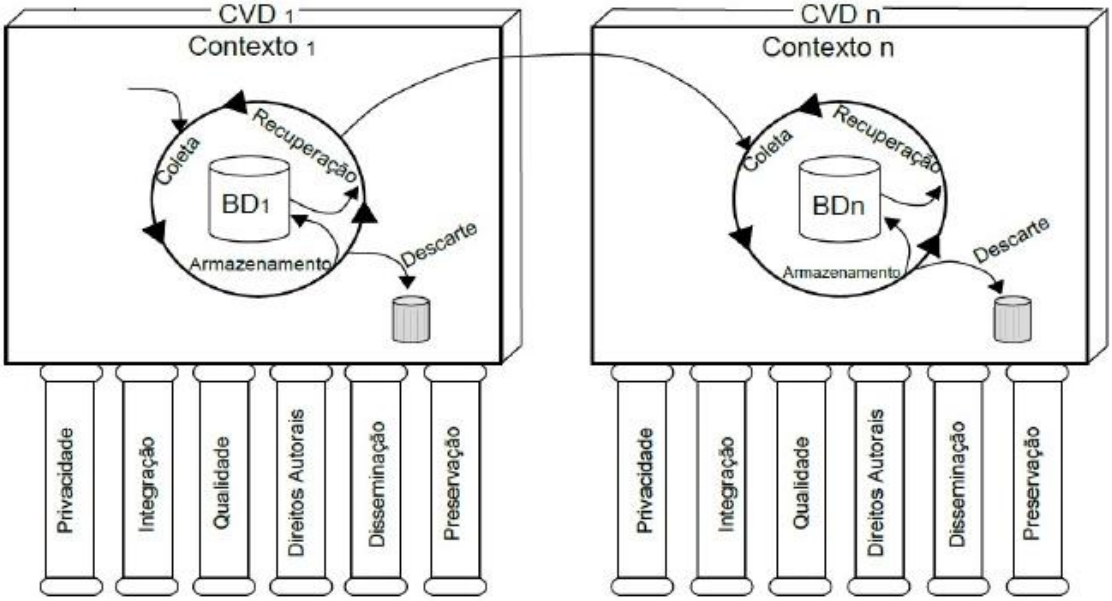
SAMARATI, P.; SWEENEY, L. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. **Technical report, SRI International**, 1998. Disponível em: <https://epic.org/privacy/reidentification/Samarati_Sweeney_paper.pdf>. Acesso em: Maio de 2017.

SANT'ANA, R. C. G. Ciclo de vida dos dados e o papel da Ciência da Informação. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO (ENANCIB), 14., Florianópolis. **Anais eletrônicos...** Rio de Janeiro: ANCIB, 2013.Disponível em: <http://enancib2013.ufsc.br/index.php/enancib2013/XIVenancib/paper/viewFile/284/319>. Acesso em: 14 jul. 2016.

SANT'ANA, R. C. G. Ciclo de vida dos dados: uma perspectiva a partir da ciência da informação. **Informação e informação**, Londrina, v. 21, n. 2, p. 116-142, maio/ago. 2016. Disponível em: <http://www.uel.br/revistas/uel/index.php/informacao/article/view/27940/20124 >. Acesso em: 20 out. 2016.

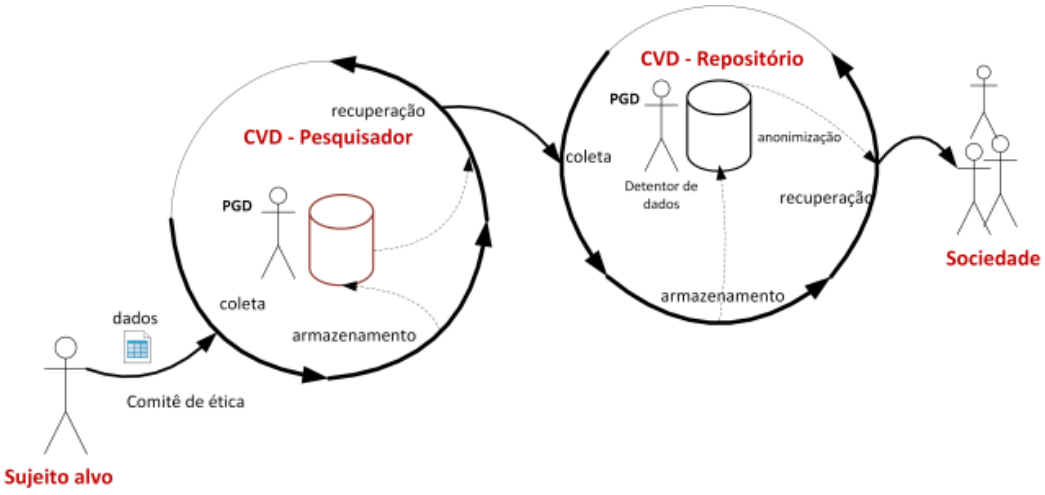
SWEENEY, L. k-anonymity: A model for protecting privacy. **International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems**, v. 10, n. 05,p. 557-570, 2002. Disponível em: <http://www.worldscientific.com/doi/abs/10.1142/S0218488502001648>. Acesso em: 14 jun. 2017.

Apêndice A – Ciclo de Vida dos Dados para a Ciência da Informação (CVD-CI)



Fonte: SANT'ANA, 2016

Apêndice B – Processo de Coleta de dados - Pesquisador e Repositório



Fonte: (Dados da Pesquisa, 2017).

REQUISITOS DE SEGURANÇA PARA PROVEDORES DE SERVIÇOS EM NUVEM DE ACORDO COM A NORMA ISO 27017

Safety Requirements for Cloud Service Providers in Accordance with Standard ISO 27017

Gislaine Parra Freund [(1)], Priscila Basto Fagundes [(2)], Douglas Dyllon Jeronimo de Macedo [(2)]

(1) Dígitro Tecnologia S.A, Florianópolis, SC - Brasil,
gislaineparraf@gmail.com
(2) Universidade Federal de Santa Catarina, Florianópolis, SC - Brasil,
priscila.bfagundes@gmail.com, douglas.macedo@ufsc.br

Resumo:
Com a ocorrência do fenômeno *big data*, surge a necessidade de tecnologia e infraestrutura adequada para suportar esse novo cenário. Neste contexto, os serviços em nuvem atendem essa demanda, porém requerem controles de segurança específicos devido a forma em que os recursos computacionais são tecnicamente concebidos, utilizados e gerenciados. O presente artigo trata-se de um estudo da norma ISO/IEC 27017:2016 com o objetivo de apresentar de maneira direta e objetiva, os requisitos de segurança referentes aos temas: definição de papéis e responsabilidades, controle de acesso e armazenamento dos dados, destinados aos provedores dos serviços em nuvem, conforme as recomendações desta norma. Trata-se de uma pesquisa bibliográfica, de abordagem qualitativa e de caráter exploratório. Para selecionar os temas da norma a serem abordados neste estudo, estes foram analisados considerando a relevância, independente da situação e do propósito de uso dos serviços em nuvem. Foi possível concluir que o tratamento da ambiguidade das responsabilidades e papéis, e a definição das responsabilidades compartilhadas permeiam outros requisitos apresentados e precisam ser abordados com atenção, e que o quadro apresentado neste estudo possibilita um entendimento rápido para aplicação dos requisitos, porém requer adicionalmente avaliações técnicas para operacionalizá-las.
Palavras-chave: *Big data*; Serviços em nuvem; Segurança da Informação; ISO 27017.

Abstract:
With the occurrence of the big data phenomenon, the need for technology and adequate infrastructure to support this new scenario. In this context, cloud services meet this demand, but require specific security controls because of the way in which computing resources are technically designed, used, and managed. The present article deals with a study of ISO / IEC 27017: 2016 with the objective of presenting in a direct and objective way, the safety requirements related to the themes: definition of roles and responsibilities, access control and data storage, For cloud service providers, in accordance with the recommendations of this standard. This is a bibliographical research, with a qualitative and exploratory approach. In order to select the themes of the standard to be addressed in this study, these were analyzed considering the relevance, regardless of the situation and the purpose of using the cloud services. It was possible to conclude that the treatment of the ambiguity of responsibilities and roles, and the definition of shared responsibilities permeate other requirements presented and need to be approached with attention, and that the table presented in this study allows a fast understanding for application of the requirements, Techniques to operationalize them.
Keywords: Big data; Cloud Services, Information Security; ISO 27017.

1 Introdução

Com a ascensão da tecnologia, dados e informações se tornaram ativos de alto valor para as organizações. De acordo com Mayer e CUKIER (2013), o uso massivo de dispositivos tecnológicos contribui para a geração desenfreada de dados, fenômeno referido por ele como a “avalanche de informação".

O volume de dados é de fato um fator relevante e cresce exponencialmente de forma que, o que era visto como futuro muito distante há uma década, já é uma realidade. Conforme exibido por Taurion (2016), a geração de zetabytes diários, deixa de ser uma escala imaginária e futurista e passa a ser uma escala real.

Mayer e CUKIER (2013) relacionam o termo *big data* com a necessidade de aprimoramento da tecnologia para atender a demanda de processamento, armazenamento e análise desse grande volume de dados e informações.

Davenport (2014) também defende que os ambientes precisam estar adequados para as soluções *big data*, de maneira a

armazenar os grandes volumes de dados sendo estes estruturados ou não estruturados, de diferentes tipos e formatos gerados a partir de fluxos intensos e contínuos. Com o intuito de fornecer, dentre outros recursos, soluções para o armazenamento e processamento de grandes volumes de dados, surge a computação em nuvem (CHEN; MAO e LIU, 2014). Algumas das características desta tecnologia é a capacidade de processamento e armazenamento, virtualização de recursos, largura de banda disponível, queda nos custos de *hardware,* etc. (SILVA, 2014)

Observa-se que soluções *big data* necessitam de alta capacidade de processamento e armazenamento para que seja possível a transformação de grandes volumes de dados em resultados que agreguem valor, e os ambientes em nuvem podem oferecer a infra-estrutura necessária para isso.

Porém, os serviços em nuvem demandam um conjunto de requisitos de segurança que precisam ser observados e tratados para não comprometer dados de usuários e de provedores deste tipo serviço.

Para atender essa necessidade, as diretrizes apresentadas na norma ISO/IEC 27002:2013 foram complementadas e em 2016 foi disponibilizada pela ABNT a ISO/IEC 27017 – Código de práticas para controles de segurança da informação para serviços em nuvem. A ISO/IEC 27002:2013 – Código de prática para controles de segurança da informação, fornece diretrizes para práticas de gestão e normas gerais de segurança da informação para organizações de qualquer natureza, tipo e tamanho. Já a ISO/IEC 27017:2016 fornece diretrizes que apóiam a implementação de controles de segurança para clientes e provedores de serviços em nuvem. Seu objetivo é fornecer controles específicos para serviços em nuvem para mitigar riscos inerentes as características técnicas e operacionais oriundas desse tipo de serviço.

O foco deste artigo é elencar e apresentar de maneira direta e objetiva, os requisitos de segurança destinados aos seguintes temas: papéis e responsabilidades pela segurança da informação, acesso e armazenamento dos dados nos serviços em nuvem, conforme as recomendações da norma ISO/IEC 27017:2016. Serão contemplados apenas os itens relacionados aos temas supracitados por serem compreendidos como requisitos essenciais e aqueles que apresentam recomendações específicas para atender as particularidades dos serviços em nuvem destinados ao provedor.

Pretende-se com esse estudo auxiliar profissionais da área da segurança da informação na aplicação destes controles da norma ISO/IEC 27017:2016.

2 Referencial Teórico

O conceito *big data*, assim como os cenários e projetos a ele relacionados, tornou as informações ainda mais importantes para as organizações por assumirem um papel estratégico de apoio na tomada de decisão.

Erl, Khattak e Buhler, (2016) consideram que *big data* tem a capacidade de mudar a natureza de uma empresa e que em algumas delas, a base de suas atividades são os insights que somente *big data* pode entregar.

Vianna, Dutra e Frazzon, (2016) resumem *big data* como a explosão de dados de forma incontrolável e a necessidade de transformar esses dados em informações relevantes para direcionar os negócios. Contudo, o *big data* requer infraestrutura e tecnologias apropriadas para processar e armazenar essa grande massa de dados. Neste contexto, os serviços oferecidos em nuvem atendem essa demanda e possibilitam o armazenamento e processamento de grandes volumes de dados com algumas facilidades de uso.

O *National Institute of Standards and Technology* (NIST) [Mell, Grace, 2011], definiu a computação em nuvem como um modelo de serviço que possibilita o uso de recursos computacionais compartilhados de forma acessíveis, convenientes e provisionados com esforço mínimo de gerenciamento ou interação do provedor. Classifica os serviços em nuvem em três modelos, são eles: Software as a Service (SaaS), Plataform as a Service (PaaS) e Infraestructure as a Service (IaaS). E apresenta quatro padrões de implantação para esses serviços:

- Nuvem privada: provisionada para uso exclusivo de uma organização.
- Nuvem comunitária: provisionada para uso exclusivo de uma comunidade específica que compartilham as mesmas preocupações.
- Nuvem publica: provisionada para uso aberto pelo publico em geral.
- Nuvem híbrida: composta por duas ou mais infraestruturas de nuvens distintas.

Dentre as possibilidades e facilidades oferecidas pelos serviços em nuvem existe a preocupação com aspectos relacionados com a segurança.

A segurança da informação é um tema importante que vem sendo discutido pela maioria das empresas de diferentes segmentos com o intuito de reduzir riscos. Segundo a norma ISO/IEC 27002:2013, segurança da informação é alcançada com a implementação de um conjunto adequado de controles de forma coordenada e coerentes com os riscos associados em uma visão holística da organização.

A computação em nuvem possui fontes de riscos de segurança próprios, derivadas de suas características, que diferem da computação tradicional, tais como: escalabilidade e elasticidade dos sistemas, compartilhamento de recursos, provisionamento de serviços sob diversas jurisdições e visibilidade limitada sobre a implementação de controles de segurança, entre outros (ISO/IEC 27017,2016).

O *Gartner Group* [Brodkin, 2008], destaca sete quesitos de segurança que precisam ser observados na utilização de serviços em nuvem, são eles: acesso privilegiado do provedor aos dados do cliente, cumprimento das regulamentações de segurança por parte dos provedores do serviço, jurisdições específicas quanto aos locais em que os dados serão armazenados, segurança no processo de segregação dos dados e uso de criptografia, recuperação dos dados em caso de incidente em tempo hábil, investigação das ações realizadas durante a prestação dos serviços em nuvem e disponibilidade dos dados mesmo na ocorrência de alterações na estrutura organizacional e estatutária do provedor.

No sentido de apoiar clientes e provedores de serviços em nuvem na implantação de controles de segurança, uma extensão da norma ISO/IEC 27002:2013 denominada de ISO/IEC 27017:2016 foi disponibilizada pela ABNT em meados de 2016. A ISO/IEC 27002:2013 é uma referência que apresenta os controles para a implementação de segurança da informação comumente aceitos, aplicáveis em organizações de qualquer porte e segmento (ISO/IEC 27002, 2013). Já a ISO/IEC 27017:2016 foi projetada utilizando a mesma estrutura de tópicos existentes na ISO/IEC 27002:2013 sendo que alguns deles foram complementados com orientações de segurança específicas para a utilização e o provimento de serviços em nuvem e alguns permaneceram iguais por aplicarem as mesmas orientações gerais de segurança apresentadas na ISO/IEC 27002:2013.

Para facilitar o entendimento da norma ISO/IEC 27017:2016 à interessados pela segurança em serviços em nuvem, esse artigo apresenta parte de seu conteúdo de forma resumida e objetiva.

3 Procedimentos Metodológicos

Para este estudo foram realizadas pesquisas entre os dias 03 e 10/07/2017, no Google Acadêmico e na base de dados *Web Science* para identificar as publicações acadêmicas que contemplam os temas segurança nos serviços em nuvem, vinculada com a norma ISO/IEC 27017:2016. Observou-se que os trabalhos que versam sobre a norma, a referenciam como o padrão que apresenta controles adicionais aos recomendados na ISO/IEC 27002, específicos para a segurança dos serviços em nuvem porém, não foram identificados trabalhos que abordam seus requisitos.

Para selecionar os temas a serem abordados neste estudo, os 13 controles da norma foram analisados com a premissa de identificar aqueles que são essenciais, independente do propósito do uso de serviços em nuvem e dos riscos associados a ele. O resultado desta análise apontou que em todas as situações de uso de serviços em nuvem, as responsabilidades e papéis pela segurança precisam ser definidos, assim como requisitos de segurança no armazenamento e no controle de acesso aos dados. Diante disso, foi realizado o estudo da norma e identificados nos controles, os requisitos relacionados com estes temas e foram apresentados em um quadro. Os requisitos abordados, limitou-se aos aplicáveis ao provedor do serviço em nuvem. Sendo assim, conforme Gerhardt e Silveira (2009), este estudo tem a abordagem qualitativa, visto que não se preocupa com representatividade numérica e trata-se de uma pesquisa básica pois seu objetivo é gerar conhecimentos novos, úteis para o avanço da Ciência, sem aplicação prática prevista. Do ponto de vista dos objetivos, é de caráter exploratório, visto que tem o propósito de promover maior familiaridade com os temas, para torná-los mais explícito ou construir hipóteses (GIL, 1991). Quanto aos procedimentos, é uma pesquisa

bibliográfica, que na definição de Fonseca (2002) “é feita a partir do levantamento de referências teóricas já analisadas, e publicadas por meios escritos e eletrônicos, como livros, artigos científicos, páginas de web sites”.

4 Resultados

A partir do estudo realizado na norma ISO/IEC 27017:2016 foram extraídos os requisitos referentes a organização da segurança quanto aos papéis e responsabilidades, segurança no armazenamento e controle de acesso aos dados, destinados ao provedor do serviço.

Vale ressaltar que a adoção dos requisitos de segurança apresentados na norma ISO/IEC 27017:2016 não elimina a necessidade de adotar os controles preconizados pela ISO /IEC 27002:2013, os quais não foram abordados neste artigo. É recomendada pela própria norma, que ambas sejam consultadas, pois muitos controles, diretrizes e requisitos se aplicam tanto para computação geral quanto em nuvem.

Organização da Segurança da Informação

Definir papéis e responsabilidades para garantir a segurança de dados é primordial em qualquer cenário e mais fácil de ser praticado quando estas atribuições estão concentradas em uma única instituição. Na utilização de serviços em nuvem, os quais envolvem outras partes no processo, de diferentes instituições, essa tarefa se torna mais complexa e precisa ser avaliada com atenção.

A ambigüidade dos papéis e as responsabilidades pela segurança dos dados nestes ambientes é um fator preocupante tanto do ponto de vista do provedor quanto do cliente desses serviços. A recomendação da norma ISO/IEC 27017:2016, é que os provedores de serviços em nuvem acordem e documentem os papéis e responsabilidades pela segurança da informação com seus clientes, prestadores de serviços e fornecedores para evitar a ambiguidade nessas definições e consequências drásticas para ambas as partes. Recomenda definir com detalhes, dentre outros itens, de quem é a responsabilidade pela propriedade dos dados, controle de acesso e manutenção da infraestrutura. Complementa com a orientação de definir e documentar as responsabilidades também pela manutenção e pelas operações desses dados, evitando assim, que práticas vitais tais como backup, recuperação, entre outras, deixem de ser realizadas pela falta de definição sobre a quem compete estas atribuições.

No uso de serviços em nuvem alguns papéis e responsabilidades de segurança são compartilhados, ou seja, são divididos entre os funcionários do cliente e do provedor. Estas atribuições devem ser identificadas, atribuídas às partes, documentadas, comunicadas e implementadas conforme acordado. E o provedor do serviço em nuvem, no papel de custodiante, deve considerar a criticidade dos dados e aplicações de seus clientes na alocação dos papéis e responsabilidades a seus funcionários, além de comunicar a eles sobre os requisitos de segurança envolvidos entre as partes para que sejam cumpridos e gerenciados como parte do serviço provido em nuvem.

Cabe ao provedor do serviço de nuvem, ainda como custodiante dos dados, informar a seus clientes sobre os países e as localizações geográficas que os mesmos podem ser armazenados para que as entidades regulatórias e as jurisdições possam ser mapeadas pelo cliente.

Armazenamento e Controle de Acesso

Serviços em nuvem utilizam ambientes virtuais compartilhados para acesso e armazenamento de dados e necessitam de proteção adicional para evitar acessos não autorizados aos dados, pelos outros clientes do serviço em nuvem que compartilham o mesmo ambiente. Para isso, a norma recomenda que o provedor do serviço em nuvem implemente segregação lógica dos dados do cliente. Ressaltando que, no caso dos serviços que envolvem multilocatários, o provedor deve garantir a segregação e isolamento apropriado para cada locatário. A norma complementa que ao armazenar dados de clientes em áreas de armazenamento compartilhado fisicamente com a tabela de metadados, a segregação dos dados de outros clientes pode ser implementada com a adoção de controle de acesso na tabela de metadados.

O provedor do serviço deve possibilitar que o cliente gerencie os direitos de acesso aos serviços e de seus usuários, fornecendo funções e especificações para registro, restrição e cancelamento desses acessos. A norma orienta ainda que o provedor apoie o uso de ferramentas para gestão de

identidade e gestão de acesso, mesmo que fornecida por terceiros, para facilitar o uso de múltiplos serviços de nuvem com login único e para possibilitar a integração e administração de identidade do cliente com o serviço em nuvem.

Os direitos de acesso privilegiados, os quais permitem acessos aos recursos administrativos do serviço em nuvem, também devem ser controlados. O provedor deve fornecer técnicas adequadas para autenticação dos administradores do serviço em nuvem, tanto para seus funcionários como para os funcionários do cliente, coerentes com os riscos associados a esses acessos. Além disso, o provedor deve disponibilizar ao cliente, informações sobre os procedimentos adotados para gerenciar e armazenar os dados referentes as autenticações realizadas no serviço, tais como: login, senhas, dados biométricos, etc.

Recursos de criptografia podem ser adotados pelos provedores dos serviços em nuvem na proteção dos dados processados e armazenados, cabendo a ele informar o cliente em quais circunstâncias a criptografia é utilizada e sobre quaisquer recursos que possa ser oferecido por ele para auxiliar o cliente na aplicação de proteções criptográficas próprias. Para este item vale uma ressalva referente a existência de algumas jurisdições que requerem a utilização de criptografia para determinados tipos de dados.

O Quadro 01 apresenta os requisitos abordados nessa seção de maneira sumarizada.

Quadro 01: Resumo dos Requisitos de Segurança conforme ISO/IEC 27017:2016 por tema

Temas	Requisitos de Segurança conforme ISO/IEC 27017:2016
Referente ao tema: Responsabilidades e Papéis pela Segurança da Informação, é recomendado ao provedor do serviço em nuvem:	- Evitar ambigüidade – acordar e documentar responsabilidade e papéis com seus clientes, prestadores de serviços e fornecedores. - Definir responsabilidade e papéis quanto a: propriedade dos dados; controle de acesso; manutenção da infraestrutura; manutenção e operações vitais dos dados (backup, recuperação, etc.). - Identificar, documentar, implementar, atribuir e comunicar as responsabilidades que são compartilhadas; - Comunicar a seus funcionários os requisitos de segurança acordados com os clientes para que sejam cumpridos e gerenciados. - Informar os países e as localizações geográficas de armazenamento dos dados.
Referente ao tema: armazenamento e controle de acesso, é recomendado ao provedor do serviço em nuvem:	- Adotar proteção adicional nos acessos em ambientes virtuais compartilhados. - Implementar segregação lógica dos dados. - Possibilitar ao cliente gerenciar os direitos de acesso aos serviços e dados. - Apoiar o uso de ferramentas de gestão de identidade. - Fornecer técnicas adequadas para controle dos acessos privilegiados. - Disponibilizar informações sobre os procedimentos adotados para armazenamento e gerência dos dados de autenticação. - Adotar criptografia e informar ao cliente em quais circunstâncias o recurso é utilizado. - Informar sobre os recursos oferecidos que auxilie o cliente a utilizar proteção criptográfica própria.

Fonte: Desenvolvido pelos autores

Todos os requisitos abordados são recomendação sobre “o que” deve ser tratado pelo provedor de serviços em nuvem. Não apresenta as recomendações tecnológicas de “como fazer” as implementações. Essas definições devem ser tratadas conforme cada situação e considerar a intenção de uso dos serviços em nuvem.

4 Considerações Finais

Observa-se que, no universo dos temas estudados, o tratamento da ambigüidade das responsabilidades e papéis e a definição das responsabilidades compartilhadas permeiam todos os demais temas. Salvo as recomendações que são diretas, ou seja, que indicam a ação exata a ser realizada, como por exemplo, implementar segregação lógica dos dados, na implantação dos demais requisitos é necessário atentar-se para evitar a ambiguidade das responsabilidades e definir com clareza e riqueza de detalhes

todas as atribuições que serão compartilhadas.

Os requisitos identificados na norma foram apresentados em um quadro que sumariza os conteúdos e possibilita um entendimento rápido sobre eles de forma a auxiliar e agilizar sua aplicação, porém requer avaliações técnicas para operacionalizar as recomendações apresentadas. Além disso, o propósito de cada cenário de uso de serviços em nuvem e os riscos de segurança associados a cada um deles deve ser considerado para implementar os requisitos na medida certa.

Para obter uma solução completa de segurança, os outros controles da norma ISO/IEC 27017:2016 devem ser analisados e adotados, pois este trabalho limitou-se a apenas três dos treze controles apresentados na norma. Além disso, a norma ISO/IEC 27002:2013 deve ser consultada assim como materiais adicionais que forem pertinentes ao cenário de aplicação.

Como sugestão de trabalhos futuros indica-se que o estudo seja complementado com os demais temas da norma ISO/IEC 27017:2016. Opções técnicas para a implementação das recomendações também podem ser apresentadas e os temas da ISO/IEC 27002:2013 pode ser interpretados e adaptados para os cenários de computação em nuvem.

Referências

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS NBR ISO/IEC 2002:2013: **Tecnologia da Informação – Técnicas de segurança – Código de prática para controle de segurança da informação**. Rio de Janeiro, 2013.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS NBR ISO/IEC 2017:2016: **Tecnologia da Informação – Técnicas de segurança – Código de prática para controle de segurança da informação com base na ABNT NBR ISO/IEC 27002 para serviços em nuvem**. Rio de Janeiro, 2016.

BRODKIN, Jon. **Gartner: Seven cloud-computing *security risks***
***Cloud computing is picking up traction with businesses, but before you jump into the cloud, you should know the unique security risks it entails*,** 2008.

Disponível em: http://www.infoworld.com/article/2652198/security/gartner--seven-cloud-computing-security-risks.html. Acesso em 03/07/2017.

CHEN Min, MAO Shiwen, LIU Yunhao. ***Big Data: A Survey***. New York, Springer Science+Business Media, 2014.

Disponível em:

https://link.springer.com/article/10.1007%2Fs11036-013-0489-0. Acesso em 05/07/2017.

DEVENPORT, Thomas H. ***Big Data* @ Work: Dispelling the Mhyts, Uncovering the Opportunities**. Boston, Massachusetts: Harvard Business School Publishing Corporation, 2014.

ERL, Thomas & KHATTAK, Wajid & BUHLER, Paul. ***Big Data* Fundamentals Concepts, Drivers & Techniques**. U.S: Arcitura Education Inc, 2016.

FONSECA, J. J. S. **Metodologia da pesquisa científica**. Fortaleza: UEC, 2002. Apostila.

GERHARDT, Tatiana E.; SILVEIRA, Denise T. **Métodos de pesquisa.** Porto Alegre: Editora da UFRGS, 2009.

GIL, ANTONIO CARLOS. **Como elaborar projetos de pesquisa**. 3. ed. São Paulo: Atlas, c1991.

MAYER-SCHÖNBERGER, Viktor; CUKIER, Kenneth. ***Big data*: como extrair volume, variedade, velocidade e valor da avalanche de informação cotidiana**. Rio de Janeiro: Elsevier, 2013.

MELL, Peter. & GRACE Timothy. **The NIST Definition of Cloud Computing.** NIST Special Publication 800-145, 2011. Disponível em: http://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-145.pdf. Acesso em: 05/07/2017 .

SILVA, Roberto Carlos Gomes da. **Migração e segurança em plataformas cloud computing**. Dissertação de Mestrado. 2014. Disponível em:

https://repositorio.ucp.pt/bitstream/10400.14/16110/1/Disserta%C3%A7%C3%A3o-Migra%C3%A7%C3%A3o%20e%20seguran%C3%A7a%20em%20plataformas%20cloud%20computing%20-%20Roberto%20Silva.pdf. Acesso em **17/07/2017.**

TAURION, Cezar. 2016. **Volume, variedade, velocidade, veracidade e valor: Os cinco Vs do *Big Data***. Disponível em: http://computerworld.com.br/volume-variedade-velocidade-veracidade-e-valor-os-cinco-vs-do-big-data. Acesso em 17/07/2017.

VIANNA, Willian Barbosa & DUTRA Moisés Lima & FRAZZON, Enzo Morosini. ***Big Data* e a Gestão da Informação: Modelagem do Contexto Decicional apoiado pela Sistemografia,** 2016. Disponível em: http://www.uel.br/revistas/uel/index.php/informacao/article/view/23327/18993. Acesso em 03/07/2017.

A ARQUITETURA DA INFORMAÇÃO E A SINTAXE DAS LINGUAGENS IMAGÉTICAS NO WEBSITE GUIA GAY FLORIPA[1]

The Information Architecture and the imagetic languages in the Gay Floripa Guide website

Jean Fernandes Brito[1], Rafaela Carolina da Silva[2], Márcio Matias[3]
(1) Universidade Federal de Santa Catarina, Florianópolis, jjeanfernandes@gmail.com
(2) Universidade Estadual Paulista Júlio de Mesquita Filho, Marília/SP, rafaelacarolinasilva@gmail.com.
(3) Universidade Federal de Santa Catarina, Florianópolis, matias97@gmail.com

Resumo:
O mundo após o fim da II Guerra Mundial abarca a inserção de meios automáticos na produção e no oferecimento de serviços, deixando de lado a linearidade, e permitindo a mistura de diferentes modalidades de linguagens. Logo, a proposta da pesquisa é analisar a Arquitetura da Informação e os elementos da linguagem imagética no *website* Guia Gay Floripa. A natureza da pesquisa é qualitativa, descritiva e exploratória. Para a análise do objeto de estudo e o alcance dos resultados esperados, foram analisados os elementos da Arquitetura da Informação presentes na plataforma. Notou-se que o website apresenta um sistema de rotulagem e de navegação capaz de estruturar as mensagens a serem disseminadas, com o uso adequado da convergência de linguagens e da Arquitetura da Informação em frases informativas e imagens que despertam a curiosidade do uso e a facilidade no acesso à informação. Conclui-se que a utilização de ambientes informacionais digitais para a disponibilização de informações turísticas para o público LGBT torna-se relevante, tendo em vista seu caráter informativo e a maior aproximação da instituição com seus usuários a partir desse ambiente colaborativo.
Palavras-chave: Arquitetura da Informação; Linguagens imagéticas; *Website* Guia Gay Floripa; Disponibilização da informação; Uso inteligente da informação.

Abstract:
The world after the end of II World War includes the insertion of automatic means in the production and in the offering of services, leaving aside the linearity, and allowing the miscellany of different modalities of language. Therefore, the research proposal is to analyze the Information Architecture and the elements of the imagetic language in the Gay Floripa Guide website. The nature of the research is qualitative, descriptive and exploratory. For the analysis of the object of study and to reach of the expected results, this research analyzed the elements of Information Architecture present in the platform. It was noted that the creators of the site know how to structure the messages to be disseminated, with the appropriate use of convergence of languages and Information Architecture in informative phrases and images that arouse the curiosity of the use and the ease in access to information. It was concluded that the use of digital information environments for the provision of tourist information becomes relevant in view of its informative nature, and the greater approximation of the institution with its users from this collaborative environment.

[1] Esse artigo é parte integrante da dissertação de mestrado do primeiro autor, do Programa de Pós Graduação em Ciência da Informação - PGCIN da Universidade Federal de Santa Catarina - UFSC e possui como temas gerais a Arquitetura da Informação e Experiência do Usuário em websites de turismo LGBT

Keywords: Information Architecture; Imagetic languages; Website Guia Gay Floripa; Provision of information; Intelligent use of information.

1 Introdução

O mundo após o fim da II Guerra Mundial abarca a inserção de meios automáticos na produção e no oferecimento de serviços, deixando de lado a linearidade, e permitindo a mistura de diferentes modalidades de "[...] linguagem e pensamento – textos, imagens, sons, ruídos e vozes em ambientes multimidiáticos – a digitalização também permite a organização reticular dos fluxos informacionais" (SANTAELLA, 2001, p. 393) no tratamento, armazenamento e disseminação da informação.

Baseando-se na informação acima, esta pesquisa tomou como base a comunidade das Lésbicas, Gays, Bissexuais e Transgêneros (LGBT), uma população ativa e que se utiliza das tecnologias digitais como forma de ampliação e manifestação da informação, não se caracterizando como minorias sociais, mas como um movimento aberto.

O movimento Gays, Lésbicas e Simpatizantes (GLS), como antigamente era designado, transformou-se, nos últimos anos, em um dos movimentos sociais mais expressivos do país (VIANA, 2006). Segundo a autora, alguns traços dessa expressão são desenhados pela presença de suas "rotinas" de ações, de seus interesses, de seus aliados e da sua representação em diversos espaços da sociedade, levando em conta o uso das tecnologias em seus diversos contextos: Turismo, Relacionamentos e Fóruns de discussão.

Nesse sentido, o objeto de estudo deste artigo é o *website* Guia Gay Floripa[2] (Apêndice A), um ambiente informacional digital que disponibiliza informações turísticas para o público gay na cidade de Florianópolis/Brasil. A mesma rede desse ambiente se concentra em outros websites, se expandido às cidades de São Paulo/Brasil, Salvador/Brasil, Belo Horizonte/Brasil e Brasília/Brasil.

A comunidade LGBT se apropria das informações do *website* na medida em que se cria um espaço de visibilidade, divulgação e ampliação do turismo com o uso das Tecnologias de informação e Comunicação (TIC). Logo, pensar em ambientes digitais para o público LGBT é ampliar a participação dessa comunidade em sociedade, otimizando o acesso à informação.

Oliveira e Vidotti (2016) apresentam a ideia de Arquitetura da Informação como uma forma de enxergar e analisar os sites da *web* e as intranets como sendo 'um todo'. Trata-se de um ambiente de informação digital que justapõe, articula e integra as partes de organização, de rotulagem, de busca, de navegação e de representação da informação, produzindo um ambiente em que o usuário pode acessar, usar e se apropriar de informações de natureza digital.

Sob essa perspectiva, Bembem, Oliveira e Santos (2015) destacam que a Ciência da Informação (CI) vive o tempo do conhecimento interativo, que visa significativas mudanças nas formas de elaborar e de acessar o conhecimento, utilizando tecnologias digitais como

[2] Disponível em: http://www.guiagayfloripa.com.br/2/home.htm Acesso em 28.ago.2017

suporte. As informações na ambiente *web* são compartilhadas em tempo real e, em termos de espaço, possibilitam a troca e a dinamização de informações, independentemente da distância existente entre os usuários, que ocorre, basicamente, em uma mesma velocidade.

Nesse contexto, os ambientes informacionais digitais, como as bibliotecas digitais, os repositórios digitais, os portais de notícias, dentre outros, influenciam no processo de transmissão da informação e na adequação desses ambientes ao usuário quanto à usabilidade, Arquitetura da Informação, *design*, experiência do usuário etc.

Assim, percebe-se, com o decorrer da história, que o desejo de criar é uma constante à espécie humana, permeando, na atualidade, a passagem dos meios analógicos às TIC, que englobam os meios digitais.

Na utilização de ambientes digitais colaborativos, o modo de construção do conhecimento se dá, por um lado, por meio da interação entre o homem e a máquina, o que permite uma maior facilidade na articulação das linguagens convergentes do ambiente através da hipertextualidade e da multimodalidade na escolha do conteúdo a ser pesquisado. A interação homem-máquina ocorre quando um ou mais homens, articulando seus conhecimentos na criação de conteúdos digitais, trabalham as plataformas digitais desde sua criação até o compartilhamento de informações em rede.

Por outro lado, nos ambientes digitais também existe a inteligência artificial, que permite a construção de conhecimentos de forma similar à humana, contudo, exibida por mecanismos ou softwares. Essas duas formas de compartilhamento de informação se complementam na construção de conhecimentos em ambientes digitais.

Partindo da ideia de Dondis (2003), no que se diz respeito ao potencial da imagem no processo de absorção das informações, esta pesquisa tem como pressuposto o trabalho do bibliotecário, acrescentando aos seus conhecimentos o uso de elementos imagéticos, o que pode contribuir para a alfabetização digital. Dessa maneira, acredita-se que os elementos imagéticos devam se relacionar com o tipo de público e com a maneira como o conteúdo é visualizado.

Assim, este estudo busca compreender como a utilização das linguagens imagéticas e da Arquitetura da Informação podem contribuir para o acesso fácil e intuitivo às informações contidas em ambientes digitais. Após as considerações apresentadas, chegou-se à seguinte indagação: “Como o *website* Guia Gay Floripa está estruturado em termos de Arquitetura da Informação e de linguagens imagéticas?”.

Sendo assim, essa pesquisa está sustentada pelo eixo temático “Aplicações em tecnologias da informação”, pertecente ao I Workshop de Informação, Dados e Tecnologia (WIDaT), na medida em que estabelece uma tessitura teórica entre Arquitetura da Informação e linguagens imagéticas para, otimização o acesso à informações turísticas ao público gay.

É necessário, portanto, que os ambientes informacionais digitais estejam adequados às necessidades, às competências e aos comportamentos informacionais dos usuários, para que esses venham a construir conhecimento a partir das informações encontradas na Internet.

Segundo Arango, Morville e Rosenfeld (2015), a Arquitetura da Informação é constituída por: sistema de organização, sistema de navegação, sistema de rotulagem, sistema de busca e sistemas de representação, observados por meio

de metadados, vocabulários controlados e tesauros.

A Arquitetura da Informação reúne uma gama de aspectos da Ciência da Informação, da Biblioteconomia e da Ciência da Computação, que têm sido frequentemente divulgados nos assuntos: estudo de usuários, cognição de usuários, políticas de informação, projeto de ferramenta de busca, projeto de interface, metadados e classificação (CAMARGO; VIDOTTI, 2011, p. 25).

A linguagem imagética pode ser conceituada como uma reunião de informações transmitidas e compreendidas direta e imediatamente. Ao contrário da linguagem verbal, a linguagem imagética pressupõe não somente a inteligência humana, mas também uma inspiração não cerebral, decorrente do sistema da visão, do olhar humano. (DONDIS, 2003).

Dessa maneira, a interdisciplinaridade entre a Arquitetura da Informação e as linguagens imagéticas, no âmbito da Ciência da Informação, mais especificamente na análise de websites, está no fato de as linguagens imagéticas proporcionarem subsídios para a estruturação dessas plataformas no que diz respeito ao sistema de rotulagem. Nessa perspectiva, as linguagens imagéticas trabalham a estrutura de rotulagem do *website*, tornando-o atrativo aos olhos de quem utilizará a plataforma.

2 Objetivos

O objetivo geral desta pesquisa é analisar a Arquitetura da Informação e os elementos da linguagem imagética no *website* Guia Gay Floripa, de modo a otimizar o acesso a informações turísticas pela comunidade LGBT. De modo mais específico, caracterizar os instrumentos que permitem a navegação do usuário de modo mais intuitivo nessa plataforma.

3 Procedimentos Metodológicos

A natureza da pesquisa é qualitativa, do tipo descritiva, exploratória e analítica, cujo sujeito é o *website* Guia Gay Floripa. Para a análise do objeto de pesquisa e o alcance dos resultados esperados, foram pesquisados os elementos da Arquitetura da Informação estudados por Arango, Morville e Rosenfeld (2015), assim como estudos da sintaxe da linguagem imagética.

A tipologia exploratória consiste em aprofundar e sintetizar aspectos técnicos e conceituais, obtidos por meio de um referencial bibliográfico e vinculados ao objeto de estudo. A tipologia descritiva procura conhecer e entender as diversas relações que ocorrem no contexto social, político, econômico e nos demais aspectos que envolvem a sociedade.

Para tanto, observou-se a integração do *website* Guia Gay Floripa no contexto da sociedade na qual ele está inserido, a fim de refletir e de trazer resultados para sustentar a problemática do estudo. O método de Estudo de Caso auxiliou no desenvolvimento da pesquisa no sentido de levantar as características da plataforma que permitem a otimização do seu uso pelos usuários.

4 Resultados

A convergência de linguagens, advinda das possibilidades de atuação em diferentes suportes de trabalho, permitiu o contato com as linguagens virtuais, ou seja, com os conceitos que simulam, por meio da computação, os produtos e serviços do mundo real. Dessa maneira, o uso de tecnologias digitais é intensificado.

O elemento mais importante de uma imagem é a tonalidade de sua cor, que varia conforme a presença de luz incidente sobre a composição (DONDIS, 2003). Dessa forma, quanto maior for a quantidade luz presente em uma cor, mais clara ela é, e quanto menor for essa quantidade de luz, mais escura ela é; a ausência

total de luz determina a cor preta; portanto, é a natureza tonal quem determina o que os olhos humanos vêem.

O que a luz revela, juntamente com a percepção do homem, identifica os demais elementos visuais. São eles:

- Ponto: menor unidade da comunicação imagética, que varia sua quantidade de acordo com a complexidade da imagem;
- Linha: proximidade entre vários pontos, chegando ao momento onde não é possível identificá-los individualmente – é a linha quem dá os primeiros passos à ideia de movimento da imagem;
- Cor: as cores mais básicas, denominadas de primárias, são o azul, o vermelho e o verde que, quando combinadas, compõem as cores secundárias e terciárias; as cores frias são aquelas que instigam a serenidade, a refrescância e a paz, diferentemente das quentes, que remetem ao agito e ao calor;
- A junção de cores pode ser cromática (diferentes tonalidades de cores afins), acromática (onde não é possível distinguir as cores) e saturadas (cores relativamente puras, próximas da matiz do cinza);
- Forma: a linha envolve uma forma e é a partir de três formas básicas – círculo, quadrado e triângulo – que são formadas as demais formas;
- Direção: é o lado para onde a linha, a forma, a cor ou a imagem em si se dirige. Existem três direções básicas na linguagem imagética: a horizontal e a vertical – para o quadrado; a diagonal - para o triângulo; e a curva – para o círculo;
- Textura: é o elemento que dá à visão a sensação do tato;
- Escala: a capacidade que os elementos visuais possuem de definirem-se uns aos outros, relação entre o grande e o pequeno;
- Dimensão: uso de pontos de fuga e estratégias para que as imagens demonstrem ter mais de uma dimensão, como é o caso das imagens em três dimensões (3D) - usadas para, mesmo que implicitamente, instigar uma maior relação com os objetos originais;
- Movimento: técnica que ilude os olhos humanos, dando a sensação de que a imagem está se mexendo (SILVA, 2014).

Nessa perspectiva, as técnicas de tratamento das imagens no *website* Guia Gay Floripa variam de acordo com os níveis de equilíbrio e de nivelamento aos olhos humanos. O intuito é impactar, portanto, trabalha-se a tensão, o contraste de cores, as texturas, a profundidade e a distribuição de conteúdos na plataforma.

O *website* é estruturado de forma a tornar a informação disponibilizada cativante, direcionada pelo uso da arte, mas mantendo a complexidade dos assuntos e, ao mesmo tempo, proporcionando diferentes meios de acesso à informação àqueles que não teriam oportunidade de se relacionar com esses conteúdos, se não por meio da plataforma.

No que diz respeito à análise da Arquitetura da Informação, verificou-se a existência dos seguintes elementos: esquemas exatos alfabéticos, cronológicos e geográficos; esquemas ambíguos, distribuídos por tópicos e direcionados a públicos específicos, pertencentes ao sistema de organização; elementos de navegação global e contextual; navegação por links contextuais, cabeçalhos, rótulos e termos de indexação; assim como preferência pela procura de itens conhecidos, dentro do sistema de busca.

Sendo assim, percebeu-se que a intenção do *website* é resgatar informações do dia a dia do mundo LGBT, de modo a informar à sociedade,

em seus diferentes gêneros e idades, a importância dessa comunidade no desenvolvimento político, econômico e social, bem como a necessidade de inclusão desse público nas dinâmicas sociais. Além disso, observou-se o crescimento do acesso ao *website* por públicos diferenciados, o que corrobora para a inserção da comunidade LGBT nas atividades do Estado, mostrando as dimensões abrangentes que esse público atingiu nos últimos anos.

4 Considerações Finais

A utilização de ambientes informacionais digitais para a disponibilização de informações turísticas torna-se relevante tendo em vista seu caráter informativo e a maior aproximação da instituição com seus usuários, a partir desse ambiente. O *website* é um ambiente colaborativo, logo, notou-se a importância dada aos elementos da Arquitetura da Informação, assim como a preocupação com o *design* de interação, com a experiência do usuário, com a usabilidade, com a acessibilidade, entre outras áreas ligadas à Ciência da Informação.

Nessa perspectiva, o usuário consegue entender o significado geral da informação disponibilizada pelo *website*. Ao interpretar tal informação, o usuário especifica o conteúdo da mensagem recebida, passando a criar novas informações e a transmiti-las novamente – princípio da comunicação humana.

Por conseguinte, o que faz com que uma imagem se torne atrativa ou não aos olhos de quem as vê é o jeito, a maneira como a mesma está sendo vista e demonstrada, sendo disponibilizada por meio de suportes de informação analógicos ou digitais.

Referências

ARANGO, J.; ROSENFELD, L.; MORVILLE, P. **Information architecture**: for the web and beyond. Canadá: O'Reilly Media, 2015.

BEMBEM, A. H. C.; OLIVEIRA, H. P. C. de; SANTOS, P. L. V. A. da C. O paradigma social e o tempo do conhecimento interativo: perspectivas e desafios para uma Arquitetura da Informação Pervasiva. **Perspectivas em Ciência da Informação**, v. 20, n. 4, p. 181-196, out./dez. Disponível em: <http://www.scielo.br/pdf/pci/v20n4/1413-9936-pci-20-04-00181.pdf>. Acesso em: 28 ago.2017

CAMARGO, L. S. A. **Arquitetura da Informação para biblioteca digital personalizável**. 2004. Dissertação. (Mestrado em Ciência da Informação) - Universidade Estadual Paulista. 2004. Disponível em: <https://www.marilia.unesp.br/Home/Pos-Graduacao/CienciadaInformacao/Dissertacoes/camargo_lsa_me_mar.pdf>. Acesso em: 28 ago.2017.

DONDIS, D. A. **Sintaxe da linguagem visual**. Tradução de Jefferson Luiz Camargo. 2. ed. São Paulo: Martins Fontes, 2003.

GUIA GAY FLORIPA. Disponível em: <http://www.guiagayfloripa.com.br/2/home.htm>. Acesso em: 18 jul. 2017.

OLIVEIRA, H. P. C. de. **Arquitetura da Informação pervasiva:** contribuições conceituais. 2014. 202 f. Tese (Doutorado em Ciência da Informação) – Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2014.Disponível em: <http://www.marilia.unesp.br/Home/Pos-Graduacao/CienciadaInformacao/Dissertacoes/oliveira_hpc_do_mar.pdf>. Acesso em 28 ago. 2017.

SANTAELLA, L. **Matrizes da**

linguagem visual e pensamento: sonora visual verbal. 3. ed. São Paulo: Iluminuras, 2001.

SILVA, R. C. da. **O uso da informação imagética no processo da inclusão digital**: uma perspectiva para atuações bibliotecárias. 2014. Trabalho de Conclusão de Curso. (Graduação em Biblioteconomia). Universidade Estadual Paulista. 2014.

VIANA, Marislene Rocha. Lutas sociais e redes de movimentos no final do século XX. In: **Revista Serviço Social e Sociedade**. Nº64. São Paulo: Cortez, 2000.

VIDOTTI, S. A. B. G; CUSIN, C. A.; CORRADI, J. A. M. Acessibilidade digital sob o prisma da Arquitetura da Informação. In: GUIMARÃES, J. A. C.; FUJITA, M. S. L. **Ensino e pesquisa em biblioteconomia no Brasil:** a emergência de um novo olhar. São Paulo: Cultura Acadêmica, 2008.

VECHIATO, F. L.; DOMINGUES, V. J.; REBELO, A. M. S.; FERNAL, A. Aplicação da Arquitetura da Informação, da usabilidade e da acessibilidade em websites de arquivos. In: CONGRESSO NACIONAL DE ARQUIVOLOGIA (CNA), 5., 2012, Salvador. **Anais**... Salvador, 2012. Disponível em: <http://www.enara.org.br/cna2012/anais/AnaisVCNA2012.pdf>. Acesso em: 10 mar. 2017.

Apêndice A – Home website Guia Gay Floripa

RECUPERAÇÃO DA INFORMAÇÃO POR TÉCNICA WEBOMÉTRICA: Análise das menções web dos partidos políticos com representação no Senado Federal

Information retrieval by webometric technique: Analysis of web mentions of the political parties with representation in Federal Senate

Eduardo Silveira[1], Márcio Matias[2]

(1) Doutorando no Programa de Pós-Graduação em Ciência da Informação da Universidade Federal de Santa Catarina, Florianópolis, edusilveira1985@gmail.com.

(2) Docente no Programa de Pós-Graduação em Ciência da Informação da Universidade Federal de Santa Catarina, Florianópolis, matias97@gmail.com.

Resumo:

Apresenta a aplicação de técnica webométrica entre as conexões na web dos partidos políticos com representação no Senado Federal com o intuito de entender como se relaciona cada partido político na rede. Para tanto, a metodologia utilizada teve como abordagem o método quantitativo com a mensuração de cinco indicadores, no qual foram analisados o tamanho de site de cada partido político e a busca intercalada de dois partidos políticos de forma simples, por operador Booleano e por website com menção web. Entre os resultados verificou que todos os partidos políticos em estudo possuem conexões na web, que a busca com operador Booleano filtra mais páginas que uma busca simples. Em relação ao tamanho de site o PSDB é o que tem o maior número de páginas, assim como o que mais emite menções web dos demais partidos políticos. Ainda em relação aos resultados o coeficiente de Correlação Linear de Pearson apontou uma classificação muito forte para as variáveis Tamanho de site confrontada com Menções web recebidas e para Busca simples confrontada com Busca Booleana. Conclui-se que estudos entre recuperação da informação e webometria possuem correlação, assim como as conexões dos partidos políticos.

Palavras-chave: Recuperação da Informação; Webometria; Menção web; Correlação Linear; Partidos Políticos.

Abstract:

It presents the application of webometric technique between the connections into the web of the political parties with representation in the Federal Senate with aims to understand how each political party relates on the network. To that end, the methodology used was the quantitative method with the measurement of five indicators, in which the size of each political party's site and the search of two political parties were analyzed in a simple way, by Boolean operator and by website with Web mention. Among the results it was verified that all the political parties in study have connections in the web, that the search with Boolean operator filters more pages than a simple search. In relation to the size of the site, the PSDB is the one with the largest number of pages, as well as the one that most emits web mentions of the other political parties. Still in relation to the results, the Pearson's coefficient Linear Correlation showed a very strong classification for the variables Size of the site confronted with received web Mentions and for simple Search confronted with Boolean Search. It is concluded that studies between information retrieval and webometric have correlation, as well as the connections of political parties

Keywords: Information Retrieval; Webometric; Web mentions; Linear Correlation; Political Parties.

1 Introdução

A atual conjuntura política no Brasil fez com que milhares de brasileiros expressassem suas opiniões, seja nas ruas ou nas mídias sociais, de norte a sul do país. Nas ruas os indivíduos contra ou a favor das situações advindas da política nacional promoviam e participavam de manifestações, marchas e grito de ordem. Já nas mídias sociais, além de serem aliadas para divulgação dos eventos oriundo do atual cenário, ocorreu em grande massa compartilhamentos de notícias, bem como a exteriorização de comentários e concepções individuais.

Muitos dos textos, notícias, vídeos compartilhados foram originados de páginas que não possuem credibilidade, fazendo muitas vezes a circulação de dados falsos e consequentemente promovendo o alcance de informações incorretas ao grande público.

Neste sentido, é primordial a procura e acesso a fontes de informação na web confiáveis. Para que cada indivíduo tenha acesso a bons recursos, se faz necessário uma busca de informações que resultam em páginas que tenham fidedignidade.

Essa fidedignidade pode ter bons resultados por meio de uma recuperação da informação elencadas em sistemas de recuperação da informação. Esses, tratam de

> sistemas de operações interligadas para identificar dentro de um grande conjunto de informações (uma base de dados, por exemplo), aquelas que são de fato úteis, ou seja, que estão de acordo com a demanda expressa pelo usuário (ARAÚJO JÚNIOR, 2007, p. 72).

Uma base de dados que recupera vários tipos de fontes de informação é o buscador Google. Numa busca simples, dependendo do assunto em pesquisa, seja por uma palavra ou termo, o Google mostra ao usuário inúmeros resultados. E por varrer toda a web, apresenta resultados de páginas, mídias sociais, blogs, dentre outros suportes e que nem sempre recuperam 100% de informações confiáveis.

Uma forma de recuperar resultados mais autênticos se dá pela webometria, por meio de conexões entre websites e as menções web. De acordo com Orduña-Malea e Aguillo (2014) as menções web tem como características a descrição de um termo/palavra na web de forma textual ou hipertextual, sendo elas encontradas em qualquer parte do documento.

Estudos de menção web com a temática política foi abordado por Lin e Park (2011), que analisaram as menções de 18 membros da Assembleia Nacional da Coréia do Sul. Na coleta de dados foi identificada a visibilidade que cada membro apresentava na web, sendo as menções web localizadas em várias plataformas, como websites de notícias, blogs e imagens. A análise constatou que os membros que tem uma melhor carreira no mundo político apresentam melhor visibilidade no mundo virtual.

Os autores, em outro momento, investigaram a relação dos políticos coreanos do sul com o montante financeiro que receberam em doações do público. Na análise foi verificado que os políticos com mais visibilidade na web tendem a receber mais doações financeiras, revelando assim uma correlação positiva entre a visibilidade das menções web e as doações recebidas (LIN; PARK, 2013).

As menções web podem ser analisadas em um âmbito geral (em toda a web) ou em nichos específicos, como no caso das duas situações analisadas por Lin e Park. Ao se deparar a um nicho ou grupo específico de websites ou um único website, as menções web podem trazer resultados mais confiáveis ao que se quer investigar.

Dessa maneira, a junção de uma menção web atribuída a um website junto ao buscador promove a combinação de artefatos que vem a culminar em resultados mais satisfatórios, pois será apresentado como resultado uma lista de documentos originados de um único website, que contém a menção web desejada. Assim, a fidedignidade das informações descritas nos documentos é certificada pela instituição que promove a divulgação do mesmo.

Em relação a política nacional, de dois em dois anos, os brasileiros acima dos 18 anos têm como obrigação participar das eleições municipais e nacionais, sendo a primeira na escolha de prefeitos e vereadores e a segunda na escolha de deputados estaduais, deputados federais, governadores, senadores e presidente.

Uma prática realizada durante a montagem de uma campanha política são as articulações de partidos para unir forças por meio de coligações partidárias. Essas coligações partidárias consistem "[...] na união de dois ou mais partidos que apresentam os seus candidatos em conjunto para uma determinada eleição." (TODA POLÍTICA, 2015). Neste sentido, é muito comum que os partidos políticos mantenham relações entre eles no mundo real além das coligações partidárias, e essas relações podem também ocorrer no meio virtual por menções web.

2 Objetivos

O objetivo geral desta pesquisa é aplicar a webometria e identificar as conexões na web entre os partidos políticos com maior representação no Senado Federal, para tanto os objetivos específicos são:

(a) quantificar o tamanho do website de cada partido político;

(b) aplicar técnica de recuperação da informação por busca simples e por operador Booleano entre os partidos políticos;

(c) analisar as conexões entre os partidos políticos por meio de seus websites e suas menções web;

(d) calcular a correlação dos indicadores averiguados.

3 Procedimentos Metodológicos

A pesquisa pretende descrever e explorar por meio das técnicas de webometria os fenômenos entre as conexões dos partidos políticos pesquisados. Para tanto, o método de abordagem tem como característica o quantitativo, com a finalidade de quantificar e interpretar os dados resultantes da técnica apresentada.

A definição dos partidos políticos sucedeu em uma busca no site do Senado Federal (http://www12.senado.leg.br), em virtude de ser a entidade que compõe senadores com representação de todos os estados do Brasil. De acordo com a Senado (2017) o pleito é composto de 81 senadores com representação de 17 partidos políticos.

Para essa análise foram selecionados os cinco partidos políticos que tem a maior representação numérica no Senado, que totalizaram aproximadamente 65,43% (53 senadores) de todo o montante. A Tabela 1 apresenta os partidos selecionados.

Tabela 1 – Partidos políticos selecionados

Partido Político	Representação
Partido do Movimento Democrático do Brasil (PMDB)	21
Partido da Social Democracia Brasileira (PSDB)	11
Partido do Trabalhadores (PT)	9
Partido Progressista (PP)	6
Partido Socialista Brasileiro (PSB)	6
Demais partidos	27
Sem partido	1
Totais	**81**

Fonte: Adaptado do Senado Federal (2017).

Diante dos partidos políticos selecionados ocorreu a busca pelo website que corresponde a cada partido. Os respectivos websites estão representados no Quadro 1.

Quadro 1 – Websites dos partidos políticos

Partido	Website
PMDB	http://pmdb.org.br/
PSDB	http://www.psdb.org.br/
PT	http://www.pt.org.br/
PP	http://www.pp.org.br/
PSB	http://www.psb40.org.br/

Fonte: Elaborado pelos autores (2017).

A partir das informações preliminares ocorreu o início da análise quantitativa. A coleta de dados compreendida nos três primeiros objetivos específicos teve como buscador escolhido o Google, realizada no dia 29 de agosto de 2017.

Para o primeiro objetivo específico 'tamanho do site' atribui a busca o conector 'site:'; O segundo objetivo específico direcionado à recuperação da informação foi utilizado a busca simples e o operador Booleano 'AND' seguidos do conector site:br; e para o terceiro objetivo específico que relaciona um website com a menção web aplicou o conector 'site:' seguido da menção web desejada.

O quarto objetivo específico teve como propósito a aplicação da correlação de Person sobre os indicadores coletados. Para esse objetivo foi utilizada a junção de duas variáveis e aplicadas no software SPPS versão 22.

Ademais, todos os dados da pesquisa foram tabulados em planilhas do Excel 2013 e quanto ao grafo apresentado em forma de rede utilizou os softwares Ucinet e Netdraw.

4 Resultados

Todos os dados foram recuperados pela sigla de cada partido político em virtude que os mesmos são mais conhecidos dessa maneira do que seus nomes por extenso. Os resultados apresentados estão divididos conforme os objetivos específicos propostos.

4.1 Tamanho de site

O primeiro indicador analisado foi o tamanho do site, que consiste em averiguar o número de páginas que cada website possui. Para tanto, foi atribuído o conector 'site:' atribuído do website desejado conforme o exemplo: "site:partidopolitico.org.br". Os montantes estão apresentados na Tabela 2.

Tabela 2 – Tamanho de site dos partidos políticos

Estratégia atribuída	Tamanho de site
PMDB	42600
PSDB	162000
PT	18000
PP	493
PSB	9320

Fonte: Elaborado pelos autores (2017).

Os três partidos com maior representação política no senado são os três com a maior quantidade de páginas na web. Sendo o maior destaque para o PSDB que tem mais que o dobro de páginas somando os demais partidos analisados.

Também foi observado um fenômeno diferente no website do PP, cujo o número de páginas é bem inferior comparado com os demais partidos.

4.2 Recuperação da informação Booleana

Com intuito de recuperar as conexões entre os partidos políticos em toda web foi submetida uma busca simples para cada par de partidos e uma busca com o operador Booleano AND aliado a expressão “site:br”, para recuperar as páginas em websites brasileiros.

O panorama das conexões da busca simples e com o operador Booleano podem ser apreciados na Tabela 3 e Tabela 4 respectivamente.

Tabela 3 – Conexão pela busca simples

Estratégia atribuída	Resultado
PMDB PSDB	3600000
PMDB PT	3160000
PMDB PP	1910000
PMDB PSB	2090000
PSDB PT	3120000
PSDB PP	1820000
PSDB PSB	5420000
PT PP	3490000
PT PSB	2090000
PP PSB	1400000

Fonte: Elaborado pelos autores (2017).

A seguir os resultados da estratégia de busca com o operado Booleano AND.

Tabela 4 – Conexão pelo operador Booleano AND

Estratégia atribuída	Resultado
PMDB AND PSDB	3440000
PMDB AND PT	3160000
PMDB AND PP	1920000
PMDB AND PSB	2140000
PSDB AND PT	3120000
PSDB AND PP	1820000
PSDB AND PSB	2140000
PT AND PP	3890000
PT AND PSB	1900000
PP AND PSB	1370000

Fonte: Elaborado pelos autores (2017).

O maior resultado obtido em conexão de duas siglas foi a dos partidos PSDB e PSB com 5420000 na busca simples e PT e PP com 3890000 na busca Booleana.

O segundo maior resultado das conexões foi dos partidos políticos que tem o maior tamanho de site (PMDB e PSDB), com 3600000 nas buscas simples e 3440000 nas buscas com o operador Booleano.

Das dez estratégias de busca sete obtiveram resultados diferentes, sendo quatro com maior expressão resultados de estratégias aplicados a buscas simples.

Baseado nos resultados apresentados entre as conexões atribuídas a busca simples e ao operador Booleano apresentamos a soma de páginas que cada sigla apareceu, conforme Tabela 5.

Tabela 5 – Somatório individual de cada partido

Partidos	Busca simples	Op. Booleano AND
PMDB	10760000	10660000
PSDB	13960000	10580000
PT	11860000	11860000
PP	8620000	9000000
PSB	11000000	7610000

Fonte: Elaborado pelos autores (2017).

Diante do resultado o que podemos perceber que quatro dos cinco partidos políticos tiveram na busca Booleana um montante de resultado menor, ou seja, por meio do operador Booleano ocorre uma redução nos resultados, que podem culminar a informações mais precisas e verídicas do que se procura, porém foram resultados muitos próximos.

Na busca simples o PSDB foi o maior somatório e na busca Booleana o PT é o partido com maior número de páginas. Já o

partido com o menor montante foi o PSB em ambas as buscas.

4.3 Websites e Menção web

Com a finalidade de recuperar informações mais precisas acerca de conexões entre partidos políticos utilizou a técnica webométrica entre a junção do website e a menção web.

Para tanto realizou estratégias de busca com um dos websites dos partidos políticos com a sigla de outro partido político da população em estudo, conforme o exemplo: "site:partidopolítico.org.br Partido A". Desse modo os resultados apresentados estão dispostos em uma matriz conforme Tabela 6.

Tabela 6 – Matriz das conexões por menção web

Website	Menção web				
	PMDB	PSDB	PT	PP	PSB
PMDB		1200	2520	745	598
PSDB	9340		27000	3060	3560
PT	1740	3790		642	639
PP	10	8	5		3
PSB	567	412	803	133	

Fonte: Elaborado pelos autores (2017).

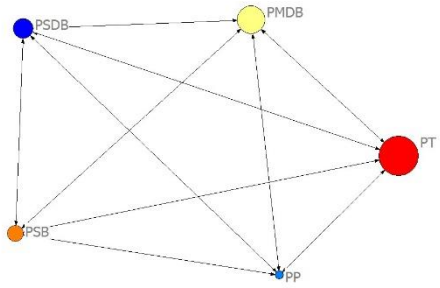
De imediato podemos afirmar que todos os websites possuem conexões por menção web como todos os partidos em estudo.

O PSDB foi o website que mais mencionou os demais partidos. Vale lembrar que o tamanho de site do PSDB colaborou com esse feito.

Já o PP por tem um tamanho de site inferior em relação aos demais foi o que menos mencionou os partidos estudados.

O panorama das menções web pode ser observado na Figura 1.

Figura 1 – Conexões entre os partidos políticos



Fonte: Elaborado pelos autores (2017).

A figura demonstrou que o PT foi o partido que mais recebeu menções, ou seja, é o partido com maior número de conexões junto de outro partido fora de seu website. De acordo com a Tabela 6 o PT é o mais mencionado em três websites (PMDB, PSDB e PSB).

O PP assim como é o website que menos menciona os partidos é também o que menos recebe menção em outros websites.

A Tabela 7 representa o quantitativo de todas as menções emitidas e recebidas pelas estratégias de busca atribuída de cada partido político.

Tabela 7 – Quantitativo das menções web

Partidos	Menções web	
	Emitidas	Recebidas
PMDB	5063	11657
PSDB	42960	5410
PT	6811	30328
PP	26	4580
PSB	1915	4800

Fonte: Elaborado pelos autores (2017).

Apenas o website do PSDB emitiu mais menções web dos demais partidos que recebeu menções deles. Sendo sua emissão mais que o triplo de todas as menções emitidas dos demais websites dos partidos.

Com 42960 menções emitidas significa dizer que aproximadamente 27% do montante de 162000 páginas do PSDB está relacionada com ao menos um partido da população em estudo.

Já o PP foi o menor website que mencionou outro partido político, com 26 menções web o montante representa aproximadamente 5% de relações com os demais partidos em sua página (total de páginas de 493).

4.4 Correlação entre os indicadores

Após os indicadores coletados teve-se como proposta entender o nível de correlação de cada variável quantitativa. Para tanto será aplicado o teste de Correlação Linear de Pearson.

Segundo Barbetta (2014) a Correlação Linear de Pearson estará sempre entre o intervalo de 1 e -1. Quando mais próxima das extremidades é considerado uma correlação forte, quando próxima a zero é considerada

uma correlação fraca. Entre a correlação forte e fraca também existe a correlação moderada que são os carros em que o índice não se aproxima de 1, 0 e -1.

O Quadro 2 apresenta a diferenciação entre Correlação Linear bem fraca, fraca, moderada, forte e bem forte ou inexistência aplicada nesta pesquisa.

Quadro 2 – Classificação de Correlação Linear

Índice (r) (+ ou -)	Descrição
(r) = 0	Não existe correlação linear.
0,001 ≤ (r) < 0,199	Correlação linear bem fraca.
0,200 ≤ (r) < 0,399	Correlação linear fraca.
0,400 ≤ (r) < 0,699	Correlação linear moderada.
0,700 ≤ (r) < 0,899	Correlação linear forte.
(r) ≥ 0,900	Correlação linear bem forte.

Fonte: Adaptado de SHIMAKURA (2006).

A aplicação de correlação linear teve como base cinco variáveis quantitativas, compostas pelos indicadores criados na análise de dados: Tamanho do site, Busca simples, Busca pelo operador Booleano AND, Menções Emitidas e Menções recebidas.

O Quadro 3 apresenta os cruzamentos das variáveis selecionadas.

Quadro 3 – Classificação de Correlação Linear

Variáveis	Código
Tamanho de site x Busca simples	CL1
Tamanho de site x Operador AND	CL2
Tamanho de site x Menções Emitidas	CL3
Tamanho de site x Menções Recebidas	CL4
Busca simples x Operador AND	CL5
Busca simples x Menções Emitidas	CL6
Busca simples x Menções Recebidas	CL7
Operador AND x Menções Emitidas	CL8
Operador AND x Menções Recebidas	CL9
Menções Emitidas x Menções Recebidas	CL10

Fonte: Elaborado pelos autores (2017).

Desse modo o Quadro 4 apresenta o resultado de cada variável confrontada e a classificação da correlação linear.

Quadro 4 – Efeito da Correlação Linear de Pearson

Código	Valor (r)	Classificação
CL1	0,833	Correlação Linear forte.
CL2	0,333	Correlação Linear fraca.
CL3	0,985	Correlação Linear bem forte.
CL4	-0,230	Correlação Linear fraca.
CL5	0,439	Correlação Linear moderada.
CL6	0,855	Correlação Linear forte.
CL7	0,174	Correlação Linear bem fraca.
CL8	0,335	Correlação Linear fraca.
CL9	0,794	Correlação Linear forte.
CL10	-0,175	Correlação Linear bem fraca.

Fonte: Elaborado pelos autores (2017).

Das 10 correlações averiguadas uma recebeu destaque com índice de correlação linear bem forte.

A CL3 que cruzou a variável Tamanho de Site e Menções web recebida ficou bem próximo de 1 com 0,985. Isto mostrou na análise que quando um website tem um número de páginas expressivo tende a emitir mais menções web.

O CL1, CL6 e o CL9 obtiveram como classificação os coeficientes de 0,833, 0,855 e 0,794 respectivamente, que corresponde a uma Correlação Linear Forte. Nesses três casos todas as variáveis estão presentes, demonstrando assim a forte ligação entre as variáveis estudadas.

Dois dos cruzamentos das variáveis apresentaram correlação negativa, vistos nos casos CL4 e CL10, a primeira como correlação linear fraca e a segunda com correlação linear bem fraca. Embora classificadas nesses índices, o negativo indica que se houver alterações nas variáveis a tendência é uma diminuir quando a outra aumentar.

5 Considerações Finais

Quanto ao objetivo principal da pesquisa que tinha como propósito aplicar a webometria nos partidos políticos com representação no Senado Federal foi atingido pois existe conexões na web entre eles.

Foi constatado que a maneira que se busca uma informação entre as conexões propostas ocorre resultados diferentes, sendo a técnica de webometria eficaz no que diz respeito a resultados mais precisos quando se analisa conexões entre duas entidades, pois a técnica confronta uma menção web e um website específico.

O que foi observado também que a webometria se correlaciona com a recuperação da informação, visto que nos confrontos das variáveis entre características de recuperação da informação com webometria os índices em metade dos cruzamentos foram expressivos, ou seja, entre correlação bem forte a moderada.

Em relação a temática Política, sugere-se estudos mais aprofundados de correlação dos partidos políticos com portais de notícias, bem

como informações do mundo real com o virtual.

Referências

ARAÚJO JÚNIOR, R. H. **Precisão no processo de busca e recuperação da informação**. Brasília: Thesaurus, 2007.

BARBETTA, P. A. **Estatística**. Aplicações às Ciências Sociais. Florianópolis: Editora UFSC, 2014.

LIM, Y. S.; PARK, H. W. How do congressional members appear on the web? Tracking the web visibility of South Korean politicians. **Government Information Quarterly**, v. 28, n. 4, p. 514-521, out. 2011.

_____. The structural relationship between politicians' web visibility and political finance networks: a case study of South Korea's National Assembly members. **New Media & Society**, v. 15, n. 1, p. 93-108, fev. 2013.

ORDUÑA-MALEA, E.; AGUILO, I. F. **Cibermetría**: Midiendo el espacio red. Barcelona: Editora UOC, 2014.

TODA POLÍTICA. **Como funciona uma coligação partidária**. 2015. Disponível em: <https://www.todapolitica.com/como-funciona-uma-coligacao-partidaria/>. Acesso em: 09 jul. 2017.

SENADO FEDERAL. **Senadores**. Disponível em: <http://www25.senado.leg.br/web/senadores/em-exercicio>. Acesso em: 11 jul. 2017.

SHIMAKURA, S. E. **Interpretação do coeficiente de correlação**. 2006. Disponível em: <http://leg.ufpr.br/~silvia/CE003/node74.html >. Aceso em: 19 jul. 2017.