VII SINGEP
Simpósio Internacional de Gestão de Projetos, Inovação e Sustentabilidade
International Symposium on Project Management, Innovation and Sustainability
ISSN: 2317-8302

O USO DE DADOS DE TEXTO DISPONÍVEIS PUBLICAMENTE NA CONTRIBUIÇÃO DO DESENVOLVIMENTO TECNOLÓGICO DAS STARTUPS NA ÁREA DA SAÚDE

EVERTON BISPO DOS SANTOS
UNINOVE - Universidade Nove de Julho

CRISTINA DAI PRÁ MARTENS
Universidade Nove de Julho

Este trabalho contou com apoio do Fundo de Apoio à Pesquisa - FAP/UNINOVE

ISSN: 2317-8302

O USO DE DADOS DE TEXTO DISPONÍVEIS PUBLICAMENTE NA CONTRIBUIÇÃO DO DESENVOLVIMENTO TECNOLÓGICO DAS STARTUPS NA ÁREA DA SAÚDE

Resumo

O objetivo deste relato técnico é apresentar um artefato para tratar e utilizar dados de texto disponíveis publicamente. Há diversos *websites* públicos com informações essencialmente textuais. Por meio da *url* de um website público é possível extrair e organizar os dados de texto, a fim de gerar conhecimento e *insights* para o desenvolvimento tecnológico de *startups*. No caso dessa pesquisa da área da saúde usando o método *design science research*, foi possível estabelecer as delimitações da classe de problemas, denominada análise de problemas e apoio à tomada de decisão, e propor um artefato que trata e utiliza os dados de texto que são públicos, a fim de direcionar o processo de geração de conhecimento, baseado em informação textual. Como exemplo de uso do artefato proposto, foram gerados alguns insights que podem ser considerados como oportunidades para o desenvolvimento de novas startups da área da saúde.

Palavras-chave: *startup; Big Data*; dados de texto; saúde; *design science research.*

Abstract

The aims this technological paper is to present an artifact that contribute with the technological development of the startups for the health sector. There are several public websites with information essentially textual. By the url from a public website, is possible extracting and organizing the text data, with the purpose to generate knowledge and insight for technological development of the startups. In case this research from the health area using the design science research method, it was possible to establish the delimitation of the class of problems called problems analysis and decision-making support and to propose an artifact that gives the direction for the process of knowledge generation based in textual information. As an example of the use these artifact, it were generated some insights that can be considered as opportunities for the development of the new startup from health area.

Keywords: startup; Big Data; text data; heath; design science research

1. Introdução

No decorrer dos últimos anos o volume, a velocidade e a variedade de dados tem crescido de maneira expressiva. Todos estes princípios caracterizam o termo *Big Data*, que vem sendo estudado e utilizado por diversas pessoas e empresas que desejam obter dados a fim de aumentar seus resultados em inovação e na exploração de novos mercados. Mesmo que ainda haja o desconhecimento e falta de alinhamento sobre o termo (Mazieri & Soares, 2016). Os dados de texto, por exemplo, são cada vez mais utilizados por estas empresas, pelo fato que as pessoas compartilham suas percepções e experiências em diversas plataformas, sites e mídias sociais, todos os dias (Marcolin, Becker, Wild & Schiavi, 2018).

A vasta criação de conteúdo pelas pessoas anualmente tem um volume de dados desconhecido por grande parte da sociedade e a cada ano esse volume aumenta consideravelmente (Pereira, Alencar, Santos Jr. & Candia, 2017). Não é novidade que muitas empresas estão implantando projetos de *Big Data*, tanto pelo fato da relevância dos dados, como pelo fato do assunto estar em voga (Dutta & Bose, 2015). Além disso, diversos segmentos da indústria estão criando projetos de Big Data (Mazieri & Soares, 2016).

No Brasil já existem mais de 5.000 *startups* (Pereira, Alencar, Santos Jr. & Candia, 2017), com aumento significativo de crescimento a cada semestre. Embora não possam ser comparadas com as empresas tradicionais que são desenvolvidas por meio de um processo diferente, as *startups* têm atraído os olhares de investidores, tanto para desenvolvimento, como para suporte. Em alguns países, são realizadas parcerias entre universidades e empresas a fim de fomentar o empreendedorismo, gerando inovação e a criação de *startups* (Souza, Santos, Rodrigues & Zanfelicce, 2017).

As *startups* surgem também de um artefato de modelagem de negócios estabelecido por Alexander Osterwalder, o *business model canvas*, que além de ser referenciado na literatura, é uma ferramenta de gestão estratégica (Alt & Zimmermann, 2015). Com o advento da internet e com a criação de modelos, como o de Osterwalder, as *startups* avançaram através de ferramentas e conceitos acadêmicos (Alt & Zimmermann, 2015). A formação de uma *startup*, por meio do *artefato* proposto por Osterwalder, mostrou-se eficaz nas primeiras etapas do projeto (Faghih, Dastourian, Sajadi, Henten & Foroudi, 2018).

No entanto, mesmo com o avanço do volume, da velocidade e da variedade da criação de novos dados de texto que possibilitam novos produtos e serviços, as *startups* que surgem geralmente por meio de inovações disruptivas (Mazieri, Quoniam & Santos, 2016), não utilizam estes dados de texto para gerarem o desenvolvimento tecnológico de *startups* para a área de saúde.

Neste sentido, o objetivo deste relato técnico é apresentar um a artefato para tratar/usar dados de texto disponíveis publicamente, que seja capaz de ler, organizar e estruturar dados de texto publicados em sites, blogs e mídias sociais relacionados a saúde. Exemplos de fontes de textos são os sites: www.e-familynet.com, https://bebe.abril.com.br/ ou https://www.minhavida.com.br/. Pretende-se, com o artefato, contribuir para o desenvolvimento tecnológico de *startups* na área da saúde. A abordagem deste artigo tecnológico será qualitativa, utilizando o método de pesquisa *design science research*.

2. Diagnóstico da situação problema

Nesta seção, serão destacados os principais direcionadores deste estudo com um pequeno diagnóstico do problema de pesquisa. Será revisada a utilização de dados de texto, o setor de saúde, o *Big Data* com seus 3 princípios e também as *startups*.

2.1. A utilização de dados de texto

Os dados de texto têm sido utilizados por diversas pessoas em todo o mundo diariamente. Estes dados são imputados em sites ou fóruns de assuntos específicos, mídias

sociais e diversos sites com dados públicos. Os usuários destas mídias sociais ou sites específicos para troca de experiência, inserem suas percepções, opiniões e comentários sobre um determinado assunto (Marcolin et al., 2018).

No entanto, estes dados não têm sido aproveitados para a tomada de decisão das empresas (Marcolin et al., 2018), como as *startups* que são objeto deste estudo, visto que as informações têm um peso cada vez maior no processo de tomada de decisão para geração de novos negócios e o desenvolvimento de produtos e serviços. Empresas de diversos setores têm colhido informações de mídias sociais para entender as preocupações e opiniões dos seus clientes. O uso de dados de texto públicos imprime a opinião generalizada e pode contribuir para o avanço de diversos segmentos do mercado.

Como destacado por Dutta & Bose (2015), o processo de limpeza, organização e disponibilização destes dados não é tão simples, pois tratar uma base de dados não-estruturada depende de ferramentas que não estão acessíveis a usuários e responsáveis de estabelecimentos comerciais, que são grandes interessados no resultado deste processo. Além disso, novas técnicas para ciência e análise de dados foram desenvolvidas.

2.2. Setor de saúde

Os brasileiros e grande parte dos cidadãos do mundo sabem da realidade do sistema de saúde no Brasil. O sistema único de saúde (SUS), é a denominação do sistema público de saúde brasileiro. De acordo com Mendes (2013), mais de 140 milhões de brasileiros são atendidos pelo SUS, no entanto mais de 45 milhões são atendidas pelo sistema de saúde suplementar. Para a quantidade de pessoas atendidas, o serviço público entrega serviço de baixa qualidade segundo, pois investe menos de 50% dos gastos públicos, que está abaixo do investimento dos países desenvolvidos (Mendes, 2013).

2.3. Big Data

O significado do termo *Big Data* ainda é desconhecido por falta de uma definição conceitual. Diversas empresas já utilizam esta nomenclatura com equívocos, pela baixa profundidade das pesquisas relacionadas ao tema, e também pela ausência desta definição (Mazieri & Soares, 2016). Ainda segundo Mazieri & Soares (2016), este termo já é utilizado desde 2001 e seus princípios são: volume de informação, velocidade e variedade, também conhecido como 3Vs (Dutta & Bose, 2015). A tecnologia *Big Data* auxilia empresas a obter novas e aprofundadas ideias por meio do comportamento de usuários e clientes (Dutta & Bose, 2015).

Volume de dados - Ao decorrer dos anos, o aumento da quantidade de dados está fora do controle das empresas, e cada vez que a quantidade de dados aumenta, o valor dos dados diminui proporcionalmente (Laney, 2001). A área de mineração de dados é constantemente aperfeiçoada para tratar o grande volume de dados (Marcolin et al., 2018). Este volume de dados cada vez maior, pode gerar ideias e sobretudo apoiar as empresas no planejamento das ações, a fim de ofertar uma gama de serviços inovadores (Marcolin et al., 2018).

Variedade de dados - Um dos maiores desafios para o uso e o gerenciamento de dados é a variedade de formato, estrutura, inconsistências e semântica (Laney, 2001). Devido à complexidade de gestão da variedade de dados, muitas discussões são levantadas, a fim de gerar novas maneiras para processar e gerir esta variedade de dados em diferentes tipos de negócios. O fator custo-benefício também é posto em discussão, pois novas maneiras de gerenciar estes dados dará mais visibilidade e apoiará a tomada de decisões. (Mazieri & Soares, 2016).

Velocidade de dados - É a dinâmica da taxa de dados que são criados e armazenados, em sites de interação entre cliente e empresas e também em mídias sociais, nos quais consumidores expressam suas opiniões (Marcolin et al., 2018). O princípio de velocidade que

faz parte do conceito Big Data, também trata do tempo/ período que são captados/ transacionados uma quantidade específica de parâmetros (Dutta & Bose, 2015).

2.4. Startup

Com uma característica diferente das empresas tradicionais que são planejadas por meio de um plano de negócios, as *startups* são fundadas geralmente por indivíduos sem experiência na indústria na qual estão empreendendo (Klepper, 2009). As *startups* nas fases iniciais, apesar de desfrutarem de aportes e suporte dos "anjos", são vulneráveis ao fracasso, devido aos conflitos entre os sócios (Pereira, Alencar, Santos Jr., Candia, 2017).

Alexander Osterwalder desenvolveu o *business model canvas*, que contribuiu e ainda contribui para a fundação de *startups* (Alt & Zimmermann, 2015). O modelo de negócios como conceito mais amplamente difundido após os trabalhos de Osterwalder, teve sua primeira aparição acadêmica em 1957 e começou a ser mais recorrente após o final de 1990 (Osterwalder, Pigneur e Tucci, 2005). Os modelos anteriores não tinham como resultado o desenvolvimento de novos modelos como o artefato proposto por Osterwalder, que pode melhorar as mensurações, análises e as comparações em organizações (Osterwalder, Pigneur e Tucci, 2005).

O desenvolvimento tecnológico de *startup* (Osterwalder, 2004) na área da saúde, por exemplo, pode trazer contribuições relevantes para a comunidade que não é assistida pelo sistema único de saúde brasileiro (Mendes, 2013).

3. Método de pesquisa

Para o desenvolvimento deste relato técnico, foram realizadas diversas pesquisas por meio das bases de dados da scopus, google scholar e web of science, na qual foram realizadas buscas por palavras-chave, como: "startups em saúde", "design science research startup", "sistema único de saúde", "artifact startups", dentre outras como "big data & startups". Além disso, mais de 60 artigos foram revisados e por meio da seção resumo selecionados, a fim de identificar sinergia com o tema deste relato técnico, que futuramente será um subsídio à geração de uma pesquisa utilizando o método de pesquisa design science research (DSR), que ao longo dos anos tem atingido grande potencial como um legítimo paradigma da pesquisa de sistema de informação, no entanto ainda não atingiu todo o potencial impacto na geração de artefatos (Gregor & Hevner, 2013). Ainda de acordo com Gregor & Hevner (2013), o mal entendimento e falta de clareza sobre o método DSR, fazem com ele não tenha a influência que lhe é devida. O que não é claramente compreendido, é que o DSR incorpora três ciclos que estão relacionados: ciclo de relevância – que apresenta artefatos de pesquisa em ambiente de teste; ciclo de rigor – que fornece teorias de métodos e incrementa novos conhecimentos gerados pela pesquisa; e ciclo de desenho central – que objetiva construir e avaliar artefatos e processo de desenho (Hevner, 2007).

Ainda sobre o DSR, é destacado que o método de pesquisa está apto para ser protagonista nas pesquisas de sistemas de informação e cabe à comunidade de pesquisadores, fomentar novas pesquisas que tragam experiência e segurança para a escolha e utilização deste método (Hevner, 2017). Autores, revisores e editores, fazem o esforço necessário para entender, difundir e dar credibilidade para o DSR como um paradigma que traz conhecimento, mesmo que ainda possa ser considerado como distinto é um método legítimo. O DSR é uma estratégia de pesquisa focada em resolver problemas de uma forma analítica e seu propósito é investigar e estudar o que é artificial, além de pesquisar o comportamento deste elemento. É denominado como um processo de alto rigor para gerar artefatos, modelos e métodos, com o propósito de solucionar problemas (Lacerda, Dresch, Proença e Junior, 2014).

3.1. Relevância

O artefato para leitura e extração de dados de texto será relevante ao desenvolvimento tecnológico de *startups* para o setor da saúde, especificamente porque por meio de acesso aos dados de textos públicos trará grandes contribuições para a sociedade. Conforme Hevner (2007), a relevância da pesquisa científica em desenho associada às atividades do projeto de pesquisa poderá contribuir significativamente para a geração de conteúdo e insights para a tomada de decisão de startups na área da saúde.

3.2. Seleção e coleta de dados

A fim de contribuir para melhoria da saúde das pessoas, bem como gerar um processo de geração de valor dos dados, o artefato para leitura e extração de dados de texto será utilizado para que o desenvolvimento tecnológico de *startups* na área da saúde seja mais ágil e assertivo, tendo como base os dados de texto que são compartilhados publicamente em sites que tratam sobre a saúde, como o www.e-familynet.com, alvo deste estudo. Este site foi escolhido por apresentar diversas interações entre os usuários, por meio de dados de texto e por estar há mais 15 anos apoiando a sociedade com informações relevantes à saúde de bebês crianças.

A coleta de dados de texto dar-se-á através de uma ferramenta de leitura e extração de dados, que após inserir a *url* do *website* e realizar a parametrização dos padrões de extração, encontrará estes dados, iniciará o processo de coleta e simultaneamente a gravação destes dados localmente, a fim de posteriormente executar os comandos necessários para organizar os dados gravados na base.

4. Resultados

O artefato denominado processo de extração e organização de dados de texto (Figura 1), que é o resultado deste relato técnico, foi estruturado e desenvolvido através do entendimento das pesquisas realizadas, principalmente com base nos artigos Marcolin et al. (2018) e Mazieri et al. (2016), que tratam sobre *webscraping* e acesso às bases de patentes, respectivamente. O artefato demonstra em um passo-a-passo como obter dados públicos e sobretudo como trata-los e utilizá-los. Com a aplicação de cada uma das tarefas deste artefato, espera-se obter acesso à conteúdo altamente relevante para a tomada de decisão e (Marcolin et al., 2018), também para o desenvolvimento tecnológico das *startups* em saúde, que poderão contribuir com serviços e informações para a sociedade, que são ou não atendidas pelo sistema único de saúde (SUS).

Este artefato possui 9 tarefas e a tarefa 1 reflete uma simples pesquisa que pode ser realizada em sites de busca, a fim de encontrar sites específicos sobre o tema pesquisado. Já a tarefa 2 do artefato se refere à identificação apenas dos sites que possuem dados públicos, cujo foi o resultado da tarefa 1. A tarefa 3 especificamente irá possibilitar o processo de *webscraping* (Marcolin et al., 2018) que será detalhado nas próximas tarefas.

Para a tarefa 4 do artefato, a *startup* precisará acessar uma ferramenta específica que consegue realizar a extração de dados de texto através da *url* do site público. Esta ferramenta por ser online ou instalada em um *desktop*. A tarefa 5 é o início do processo de interação entre a ferramenta utilizada por meio da *startup* e o site público que contém os dados de texto que serão extraídos e organizados. Esta interação será realizada por meio de acesso à base de dados do site público via comando de *javascript*.

Na tarefa 6 do artefato, as funções de *javascript* que serão parametrizadas pela *startup*, farão tanto a leitura como a captura dos dados de texto, como por exemplo: palavras-chave, repetições de palavras específicas, semântica, assuntos mais discutidos, tópicos com mais *likes*, tópicos com mais acesso e interações, dentro outros. A tarefa 7 é mais complexa tecnicamente, pois necessitará de conhecimentos específicos e ferramentas de apoio para

estruturar a saída e compilação dos dados de texto coletados, por meio de tabela, linha ou matriz com os dados já estruturas conforme os parâmetros definidos na tarefa 6.

De volta ao ambiente da *startup* agora na tarefa 8 do artefato, os dados serão transformados em frases, ideias, ocorrências e sentenças, para que através da tarefa 9, um relatório possa ser elaborado com informações consistentes sobre o site público, com evidências para o desenvolvimento tecnológico de *startups* em saúde.

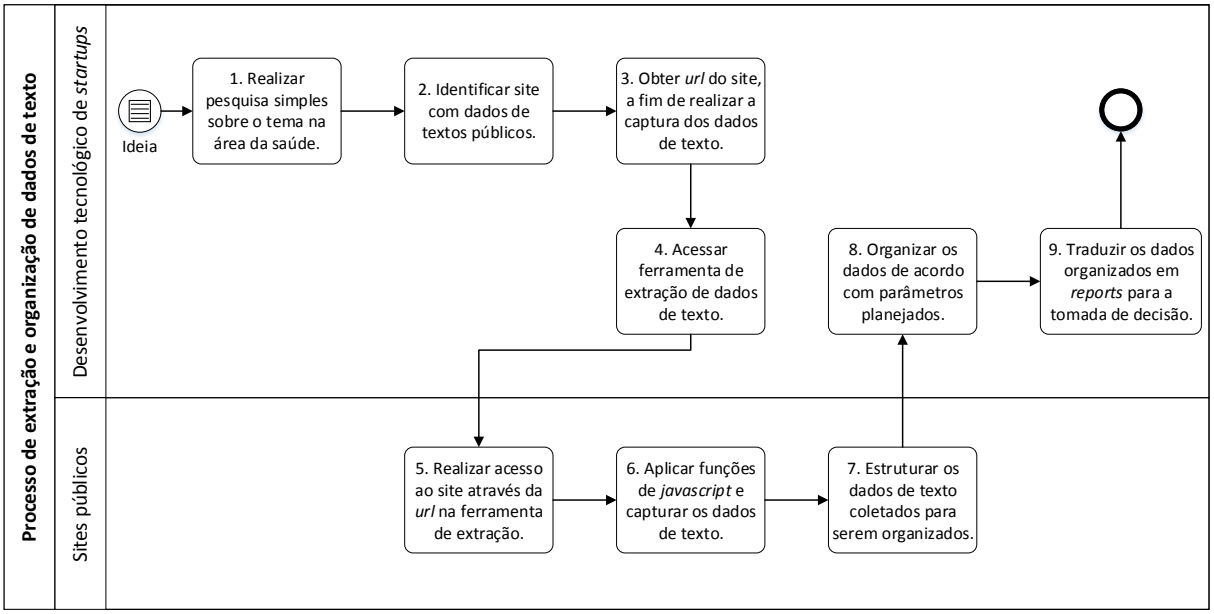

Figura 1: Artefato para leitura e extração de dados de texto
Fonte: Elaborado pelo autor com base em Marcolin, Becker, Wild e Schiavi (2018) e Mazieri, Quoniam e Santos (2016).

A classe de problemas denominada análise de problemas e apoio à tomada de decisão, está totalmente conectada ao artefato para leitura e extração de dados, porque é propícia para analisar o problema de tratamento e utilização de dados de texto públicos por *startups*, e apoiar na estruturação da solução deste gap com o artefato proposto. Esta classe de problemas visa a generalização da sua utilização para o tratamento de problemas dessa natureza, que por sua vez, poderão se beneficiar da proposição deste artefato. Para este projeto de pesquisa, esta classe de problema apoiou na identificação do problema de disponibilidade e utilidade de dados de texto, e na construção do processo sistêmico como apoio na tomada de decisão, por meio da geração do artefato.

Através da proposição deste artefato para leitura e extração de dados de texto, especificamente utilizado para o site www.e-familynet.com/, foram gerados insights para responder a questão central desta pesquisa. Este website, dedicado à área da saúde, dispõe de grande conteúdo de dados de textos públicos, que poderiam gerar aplicativos para que mães pudessem compreender melhor as principais ocorrências dos filhos recém-nascidos, com informações sobre medicamentos, especialistas em saúde e os principais hospitais por região; outra possível proposição seria identificar a quantidade de ocorrências de uma determinada patologia em um *range* de idade, a fim de fornecer subsídios para tomada de decisão na concepção de novas pesquisas com o intuito de gerar vacinas e tratamentos específicos através desta volumetria de informações.

5. Análise

Após a realização das pesquisas para chegar até o artefato demonstrado na Figura 1, foi diagnosticado que a área da saúde necessita de produtos e serviços, para eliminar problemas declarados e compartilhados em *websites* públicos, por meio de dados de texto.

Conforme já destacado na seção de diagnóstico da situação problema, o setor da saúde tem um déficit de atendimento à população que utiliza o sistema único de saúde (SUS). Este problema poderá ser atenuado ou eliminado também, com o aproveitamento do conhecimento gerado por meio destes dados, pois hoje são compartilhados, mas não há desenvolvimento tecnológico para suprir esta demanda. Este artefato contribuirá com esta classe de problema, que analisa os *gaps* e fornece apoio à tomada de decisão, pois especificamente apresenta uma lógica de fácil entendimento, por meio do processo de pensamento sistêmico, fazendo associações e correlações com as tarefas que são dependentes e sequenciais. Com a aplicação deste artefato, espera-se contribuir com a atenuação do problema de geração de conhecimento para o desenvolvimento tecnológico de *startups* em saúde e após a sua materialização, também ser aplicado para problemas de outros setores que podem ter o desenvolvimento tecnológico de *startups*.

Esta pesquisa teve como escopo apenas da geração de um artefato para tratar/usar dados de texto disponíveis publicamente. Desta forma, não foi realizada identificação e análise de ferramentas que poderão executar a parametrização e também como serão estabelecidas a extração e a organização dos dados de texto, que ficará a cargo da *startup* que irá utilizar o artefato, na melhor forma.

É sugerida a elaboração de outros estudos para aprofundamento das ferramentas mais efetivas, capaz de realizar a captura, leitura, extração e organização dos dados de texto, bem como estudos que possam identificar padrões de linguagens de desenvolvimento de *websites* que possuem dados de textos públicos, para que estes possam ser uteis na geração de conhecimento e insights para o desenvolvimento tecnológico de *startups*.

Referências

Alt, R., & Zimmermann, H. D. (2014). Status of business model and electronic market research: An interview with Alexander Osterwalder. *Electronic Markets*, 24(4), 243-249.

Cohen, W. M., & Levinthal, D. A. (2000). Absorptive capacity: A new perspective on learning and innovation. *In Strategic Learning in a Knowledge economy* (pp. 39-67).

Dutta, D. Bose, I. (2015). Managing a big data project: The case of Ramco Cements limited. *International Journal of Production Economics*.

Faghih, N., DASTOURIAN, B., Sajadi, S. M., Henten, A., Foroudi, P. (2018). A Framework for Business Model with Strategic Innovation in ICT Firms: The Importance of Information. *The Bottom Line* (pp. 16- 41).

Gregor, S., & Hevner, A. R. (2013). Positioning and presenting design science research for maximum impact. *MIS quarterly*, 37(2).

Hevner, A. R. (2007). A three cycle view of design science research. *Scandinavian journal of information systems*, 19(2), 4.

Klepper, S. (2010). The origin and growth of industry clusters: The making of Silicon Valley and Detroit. *Journal of Urban Economics*, 67(1), 15-32.

Lacerda, D. P., Dresch, A., Proença, A., Júnior, J. A. V. A. (2014). Design Science Research: método de pesquisa para a engenharia de produção. *Gest. Prod*. vol.20 no.4 São Carlos 2013.

Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. *META Group Research Note*.

Mazieri, M., Soares, E. D. (2016). Conceptualization and theorization of the Big Data. *International Journal of Innovation*, v. 4, p. 23-41.

Mazieri, M., Quoniam, L. (2016). Inovação a partir das informações de patentes: proposição de modelo Open Source de Extração de Informações de Patentes (Crawler). *Revista Gestão & Tecnologia*, v. 16, p. 76-112.

Mendes, E. V. (2013). 25 anos do Sistema Único de Saúde: resultados e desafios. *Estud. av*. vol.27 no.78 São Paulo.

Osterwalder, A., Pigneur, Y., & Tucci, C. L. (2005). Clarifying business models: Origins, present, and future of the concept. Communications of the association for Information Systems, 16(1), 1.

Pereira, J. C., Alencar, L. M., Junior, H. A. G. S., Candia, J. J. (2017). Aspectos relevantes para investimento em startups. *Simpósio Internacional de Gestão de Projetos, Inovação e Sustentabilidade*.

Souza, R. M., Santos, T. A., Rodrigues, W. M., Zanfelicce, R. L. (2017). Cooperação universidade-empresa nos EUA: Fomento ao Empreendedorismo no MIT (Massachusetts Institute of Technology) e CIC (Cambridge Innovation Center). *Simpósio Internacional de Gestão de Projetos, Inovação e Sustentabilidade*.