Ferramentas de Coleta de Dados para Pesquisadores da Comunicação e Jornalistas - Desenvolvimento e Aplicação[1]

Márcio Carneiro dos Santos[2]
Universidade Federal do Maranhão, São Luís , MA

RESUMO

Diante da situação de excesso de informações hoje disponíveis no ambiente digital, ao trabalho de profissionais de jornalismo e pesquisadores da área, foi adicionado o desafio de lidar com situações de volume, variedade e velocidade no tratamento de dados. Daí a necessidade de ferramentas capazes de automatizar a parte repetitiva dos processos de coleta e extração de material a partir de bases, sites, APIs e repositórios de diversos formatos, oferecendo ao usuário tempo para dedicar às partes mais complexas da sua atividade. Como resultado de um esforço de pesquisa aplicada, orientada à solução de problemas reais desta ordem, detalha-se, com base na *Design Science*, uma suíte de ferramentas, entre elas o LNEWS, para coleta de conteúdo em portais jornalísticos e o LTWEET para extração de conteúdo da plataforma Twitter.

PALAVRAS-CHAVE: jornalismo guiado por dados; python; bases de dados; extração automatizada; jornalismo computacional.

1. INTRODUÇÃO

A condição contemporânea do ecossistema comunicacional, baseada em redes digitais, massivo crescimento de emissores e explosão informacional, caracterizada por volume, velocidade e variedade, tem sido devidamente registrada nos últimos vinte anos a partir dos trabalhos de Castells (1999), Chwe (2000), Feenberg (2002), Lemos (2002), Santaella (2003), Vilches (2003) entre outros.

O empoderamento público sustentado pelos meios digitais de comunicação criou uma explosão de emissores, reconfigurando as redes de difusão de informação do

[1] Trabalho apresentado no GP Conteúdos Digitais e Convergências Tecnológicas, XX Encontro dos Grupos de Pesquisas em Comunicação, evento componente do 43º Congresso Brasileiro de Ciências da Comunicação.

[2] Professor permanente do programa de Mestrado Profissional da UFMA, email: mcszen@gmail.com

mundo analógico, antes concentradas em grandes *hubs* de atenção, como os veículos de mídia e as fontes oficiais (Santos, 2019, p. 146)

Suas características de interconexão (NEWMAN, 2010; EASLEY, KLEINBERG, 2010), excesso (GLEICK, 2013) e complexidade (SIMON, 1962; MORIN, 2005), aliadas a um conjunto próprio de especificidades que constituem uma verdadeira ontologia dos entes binários, também já mapeada por Manovich (2001) e Santos (2016), indicam um quadro consolidado de percepção de profundas mudanças que, entretanto, parece ainda não ter impactado de forma proporcional os métodos e técnicas utilizados pelos pesquisadores da Comunicação para analisar objetos do ciberespaço.

A iniciativa denominada Métodos Digitais, que tem em Rogers (2013) um de seus principais interlocutores, explicita a necessidade, gerada pelas especificidades dos entes de origem binária, de estuda-los a partir métodos e técnicas originados em sua lógica interna.

Transpondo tal raciocínio ao trabalho de pesquisa, Rogers também separa os métodos eminentemente digitais dos que ele denomina de virtuais, ou seja, que têm sua origem em outros campos e têm sido adaptados para a internet e as redes sociais. A netnografia ou etnografia virtual, os questionários aplicados via e-mail, as entrevistas mediadas pelo computador e pelas redes são algumas das formas adaptadas, diferentes, por exemplo, da mineração e raspagem de dados (data mining e scraping), do acesso direto às APIs das plataformas de mídia social, da utilização de métricas como o Page Rank ou de ferramentas como Open Refine para, respectivamente, coletar, classificar e organizar dados.

Amostras pequenas ou tratadas manualmente pouco podem fazer em movimentos de milhares ou milhões de ações humanas, realizadas muitas vezes de forma quase sincrônica, tais como um conjunto de pessoas publicando *tweets* sobre um tema que momentaneamente arrebata a atenção coletiva (Santos, 2013).

Os métodos digitais, dessa forma, não pretendem substituir os atuais, mas complementá-los em situações de pesquisa ou profissionais onde se está lidando com objetos cuja ontologia própria exige um tipo de abordagem mais efetiva.

2. DESENVOLVIMENTO ORIENTADO A PROBLEMAS REAIS

> Foram os jornalistas investigativos e os profissionais no ambiente organizacional que primeiro tiveram que lidar com esse tipo de problema. Dados dos portais públicos de transparência e as histórias que podiam conter escondidas atrás dos números; métricas dos indicadores relacionados à presença nas plataformas de mídias sociais; informações trazidas pelas ferramentas de monitoramento; e a enxurrada de dados das soluções de *analytics*, tudo isso trouxe ao campo não apenas um conjunto novo de problemas, mas uma espécie de reação em cadeia que começou a impactar também a necessidade de novas habilidades desses profissionais, diferentes formas de abordagem, a busca por novos modelos de negócio e – por que não também? – a revisão e readequação teórica e epistemológica de um saber que tinha sido construído num mundo analógico, bem diferente do atual (Santos, 2019, p. 147).

A proposta das ferramentas LNEWS e LTWEET vai nessa direção, não como específicas para uma pesquisa e sim como meta-ferramentas de utilização ampla para uma categoria de problemas como propõe a linhagem epistemológica da Design Science.

Design Science (DS) é a "ciência que procura consolidar conhecimento sobre o projeto e desenvolvimento de soluções para melhorar sistemas existentes, resolver problemas e criar novos artefatos" (DRESCH; LACERDA; ANTUNES Jr., 2015, p. 59).

"O artefato, criado pelo homem, representa um intermediador entre um conjunto do conhecimento estabelecido em determinada área e as condições específicas que envolvem o problema que o artefato deverá resolver. (SANTOS, 2016, p.11)".

A DS classifica os artefatos em um continuo que vai de um extremo mais abstrato a outro mais tangível e inclui as categorias de constructos, modelos, métodos e instanciações.

O termo classe de problemas também faz parte dos conceitos importantes da DS. Conjuntos de problemas práticos ou teóricos, que tem já estabelecido um conjunto de soluções ou artefatos a eles ligados, constituem-se numa classe de problemas.

Como exemplo da Comunicação e das Ciências Sociais, poderíamos citar a necessidade geral de coletar dados em repositórios na internet, que poderíamos nomear como coleta de dados digitais.

Seja para a produção de uma matéria jornalística, para um plano de gestão ou para a definição de uma política pública sobre determinado tema, com os processos de digitalização e o crescimento do uso de bases de dados, a necessidade de conseguir tais informações, acessando seus repositórios disponíveis na rede, tais como portais de transparência, por exemplo, caracteriza uma classe de problemas onde operam artefatos como os métodos de scraping (raspagem) e extração automatizada, bem como as instanciações disponíveis exemplificadas pelos algoritmos em determinada linguagem de programação, que operam para resolver tais problemas (SANTOS, 2016, p.13).

Deste modo a proposta das ferramentas do LABCOM se insere na categoria de problemas representada pela necessidade enfrentada pelos profissionais e pesquisadores de extrair conteúdo a partir de sites jornalísticos ou da plataforma Twitter, de forma automatizada, garantindo uma amostragem mais significativa e ganho de tempo, evitando a coleta manual bem mais lenta e cansativa.

Os artefatos, utilizando a terminologia da DS, permitem também ao pesquisador da Comunicação ou ao jornalista a visualização da estrutura interna das editorias/categorias no caso do LNEWS, bem como buscas por perfil ou termo, no caso do LTWEET.

Tal tipo de informação, sem o uso de recursos computacionais, é de acesso bem mais difícil. Por situações deste tipo através do site www.labcomdata.com.br iniciamos o cadastro e acesso de interessados para utilização das soluções que fomos desenvolvendo.

Figura 1 – Tela inicial do site labcomdata.com.br . Fonte: do autor.

O objetivo deste esforço de pesquisa aplicada é atender a uma demanda crescente de trabalho na área de jornalismo guiado por dados, entre profissionais, contemplando também diversas iniciativas de pesquisa no campo do Jornalismo bem como da Comunicação em geral, mantendo o foco num conjunto de utilizadores mais distantes das práticas de programação e técnicas de raspagem de dados.

Se, em síntese, a pesquisa aplicada é aquela que é orientada à solução de problemas reais, para profissionais e pesquisadores, apreender os movimentos do ecossistema de meios digitais através de ferramentas que lhes permitam coletar e organizar esses dados de forma mais fácil e prática, economizando o seu tempo para as tarefas mais nobres de análise e interpretação, nos parece importante e fundamental diante do contexto contemporâneo.

Links diretos de acesso:

. LNEWS: https://colab.research.google.com/drive/1e9_FdP-7g3KXAjsglqPz-D1ziQ8NpKMI

.LTWEET:https://colab.research.google.com/drive/1wGMvNWqIg6jWCPnSOq8iMNFLpXrNGnpx

. Página do Site com tutoriais e links de acesso: labcomdata.com.br/teste

. Vídeo Tutorial LTWEET: https://youtu.be/z1BUTcxUmJ0

. Vídeo Tutorial LNEWS: https://youtu.be/aTUzx-GVTR4

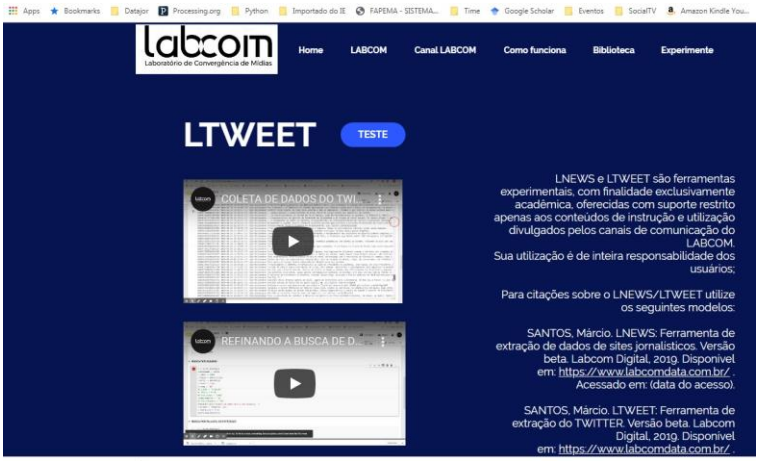

Figura 2 – Página do site com acesso às ferramentas, tutoriais e demais informações.
Fonte:do autor.

3. LNEWS

LNEWS – Ferramenta de Coleta de Conteúdo em Sites Jornalísticos

No seu atual estágio a ferramenta LNEWS permite ao pesquisador:

a) Selecionar um site ou portal jornalístico, com 3 opções pré-selecionadas UOL, ESTADÃO e GLOBO.COM e uma quarta opção onde o usuário pode entrar com um endereço específico para, por exemplo, explorar algum veículo da sua cidade ou estado, com relevância específica para sua pesquisa;
b) A partir da escolha feita no item anterior o software coleta uma lista de links disponíveis para acesso no momento da consulta onde é possível ver a estrutura de organização das editorias/categorias que são utilizadas pelos emissores na publicação do seu conteúdo;
c) Com a lista o usuário pode escolher qualquer uma das matérias recuperadas e vê-la na íntegra com alguns dos seus meta-dados, quando acessíveis, tais como seus autores. Ainda nesse momento é criada uma lista das palavras mais frequentes e feito um sumário (resumo) do texto. Essa última função ainda apresenta pouca eficiência no caso da língua portuguesa.
d) Após essa etapa é oferecido ao pesquisador salvar sua coleta criando dois arquivos, um com a lista de links para as matérias e o segundo com os textos das mesmas, podendo realizar essa última ação adicionando filtros por editoria ou até pela seleção de matérias que apresentem uma palavra específica. Por exemplo, do total de matérias coletadas é possível salvar apenas as da editoria Política em que a palavra "poder" esteja presente.

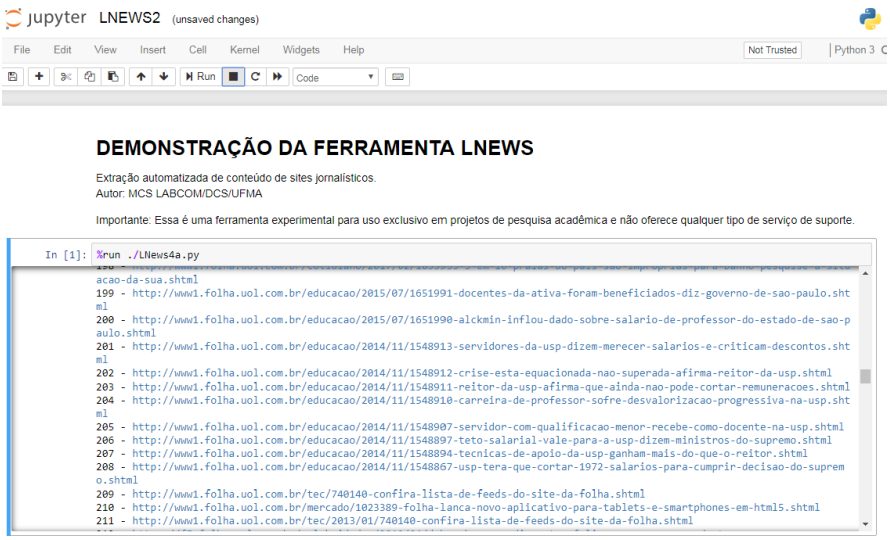

Figura 3 – Tela com lista de links extraídos. Fonte: do autor.

Outro módulo da solução, denominado LNEWS Análise, já trabalha não mais com os sites diretamente e sim com os arquivos de textos extraídos. A partir deles estão sendo desenvolvidas funções para calcular e visualizar métricas tais como a diversidade léxica, que traduz a riqueza na utilização das palavras do texto e as palavras frequentes que podem ser também listadas ou apresentadas num gráfico.

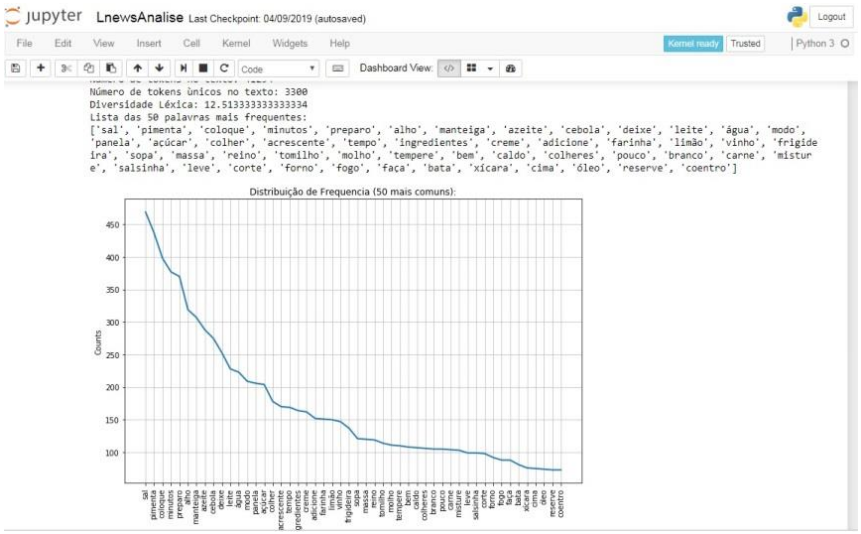

Figura 4 – Tela Lnews Análise com palavras frequentes do texto e gráfico. Fonte : do autor .

4. LTWEET

Considerando a importância atual da plataforma Twitter, seja para acompanhamento dos fatos, principalmente da política, que na plataforma se refletem e aparecem com enorme rapidez, bem como de diversos outros aspectos da produção de sentido humana que, através de tweets, vão se revelando, a necessidade de ferramentas para extrair e organizar esse conteúdo é fundamental, seja para jornalistas como também para pesquisadores.

Hoje a ferramenta LTWEET permite a extração de conteúdo por perfil ou termo (tag) incluindo também uma essencial funcionalidade de especificar o período de interesse do profissional ou pesquisador.

Figura 5 – Tela do LTWEET com suas opções de coleta. Fonte: do autor .

O trabalho com o LTWEET também já gerou frutos adicionais que temos colocado na página do LABCOM DATASETS, ou seja, na página disponibilizamos conjuntos de tweets já extraídos para consulta ou análise incluindo, entre outros, o do Governador do Maranhão, Flávio Dino e o do presidente Bolsonaro.

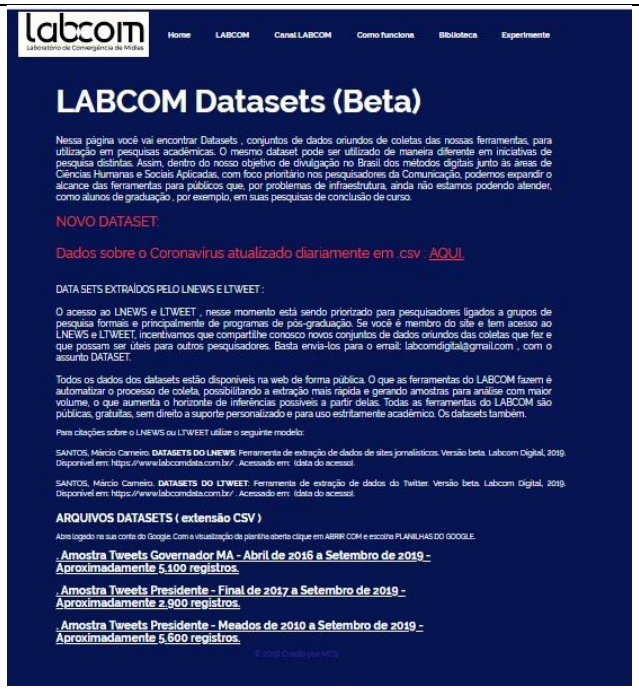

Figura 6 – Tela da página do site LABCOM DATASETS que oferece conjunto de dados já extraídos utilizando a ferramenta LTWEET. Fonte : do autor .

Hoje, apenas para controle, o acesso à página das ferramentas LNEWS e LTWEET é feito mediante cadastro em formulário que nos permite avaliar o grau de utilização e a diversidade dos usuários interessados nas mesmas. Atualmente quase 100 pesquisadores, de todas as regiões do país e representando diversos grupos de pesquisa e interesses diversos, já se cadastraram para utilizar.

Para alunos e profissionais de jornalismo também temos disponibilizado oportunidades de contato com as ferramentas através dos canais de comunicação do LABCOM bem como em eventos como Intercom Nordeste, Intercom Nacional e Encontro Nacional da ABCiber .

Figura 7 – Apresentação das ferramentas para alunos do Mestrado Acadêmico da UFMA em Imperatriz. Fonte: do autor.

O interesse pelas ferramentas entre pesquisadores e jornalistas também pode ser identificado através dos registros do Google Analytics do site do LABCOM – www.labcomdata.com.br .

	Página	Visualizações de página	Porcentagem do Visualizações de página
1.	/coleta-lattes	210	50,85%
2.	/code	51	12,35%
3.	/	32	7,75%
4.	/dados-covid	31	7,51%
5.	/metodos-digitais	16	3,87%
6.	/datasets-beta	11	2,66%
7.	/code?fbclid=IwAR0ZEvmvECKUupjLmOHUtDmhOGIm7xY_6nt0CJqbszzFcSdLg5jnGDoFcXM	8	1,94%
8.	/perguntas-frequentes	6	1,45%
9.	/lattes	5	1,21%
10.	/apresentacoes-e-eventos-labcom	4	0,97%

visualizar relatório completo

	País	Usuários	Porcentagem do Usuários
1.	Brazil	185	92,50%
2.	United States	11	5,50%
3.	Portugal	3	1,50%
4.	Mexico	1	0,50%

visualizar relatório completo

Figuras 8 e 9 – Telas com métricas do Google Analytics a partir do site do LABCOM, indicando o crescimento do tráfego no site após o lançamento das ferramentas e os países de origem dos acessos. Fonte: do autor.

Além das duas ferramentas detalhadas aqui, a suíte de soluções do LABCOM também inclui a LCLattes que gera relatórios de produção para a plataforma Sucupira e a LQUALIS que apresenta as avaliações Qualis das revistas acadêmicas da CAPES.

Dados adicionais sobre a suíte de ferramentas desenvolvida pela LABCOM:

LNEWS — ferramenta de coleta de conteúdo em sites jornalísticos. Extrai textos, links e títulos das matérias filtradas por editoria ou palavra-chave.

LTWEET — ferramenta de coleta de tweets que podem ser filtrados por usuário ou palavra-chave.

Link de acesso para ambas: https://www.labcomdata.com.br/teste . Observação: é necessário se cadastrar em formulário na home do site e depois solicitar acesso de membro já que o acesso é feito logado no site.

LTEXTO — Faz análise de arquivos de texto gerando visualização de palavras frequentes e dispersão léxica. Ainda em desenvolvimento. Utilização mediante solicitação de link.

LQUALIS — Verifica o QUALIS Periódicos de publicações da área de Comunicação e Informação oferecendo a pontuação nas tabelas antiga e nova (ainda não oficial) e consulta em lote.

Link de acesso: https://www.labcomdata.com.br/code

LCLATTES — Gera relatórios detalhados sobre a produção dos pesquisadores com base nos dados do CNPq. Inclui hoje a produção acadêmica em revistas, trabalhos em eventos, livros e capítulos.

Link de acesso: https://www.labcomdata.com.br/coleta-lattes

Vídeos Tutoriais para as ferramentas são encontrados no canal do LABCOM no YOUTUBE em

https://www.youtube.com/channel/UCz9xhatSPQ9MZFPf1e9kDUQ/videos

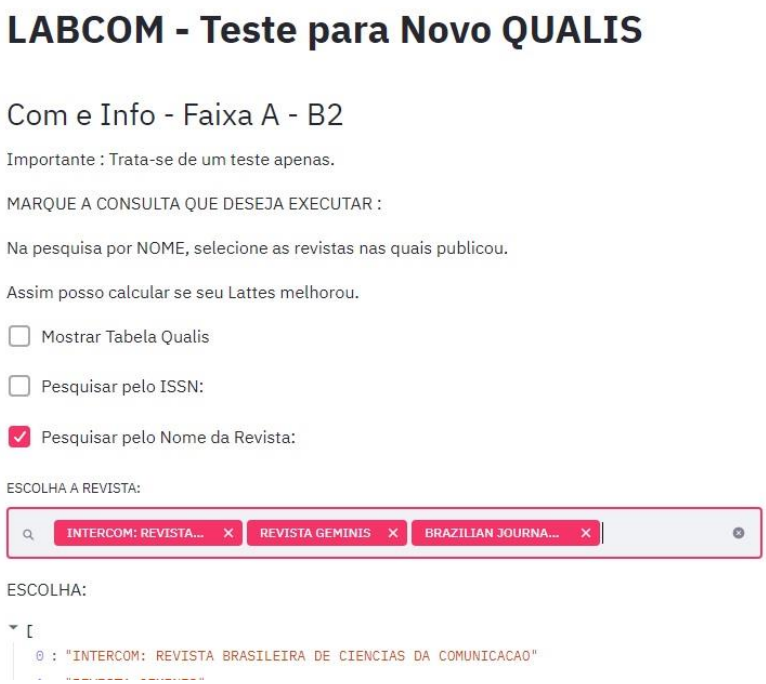

Figura 10 – Tela da solução LQUALIS que também pertence à suíte de ferramentas do LABCOM. Fonte: do autor.

5. CONSIDERAÇÕES FINAIS

As características de volume, variedade e velocidade impactam diretamente a efetividade de abordagens não orientadas a lidar com sua presença, interferindo em resultados, diminuindo as possibilidades de inferências mais sólidas e, inclusive, inviabilizando a apreensão de fenômenos cada vez mais importantes na lista dos pesquisadores da Comunicação.

Iniciativas, ainda em desenvolvimento, como o novo LWhats, voltam-se agora a um dos mais promissores nichos de pesquisa ou investigação jornalística, considerando a centralidade e o nível amplo de adoção da ferramenta Whats App, por indivíduos e organizações.

Através dessa nova linha, abre-se um espaço ainda pouco explorado e com grande interseção com debates, incluindo políticos, dentro das discussões na nova esfera pública digital.

A nova ferramenta automatiza e coleta uma série de métricas sobre as conversas em grupos de WhatsApp, incluindo número de publicações totais, classificação dos tipos de postagens, contando a parte de texto, número de arquivos multimídia, utilização de emojis, gerando também relatórios individuais dessas mesmas características de todos os participantes do grupo.

Além disso, há possibilidade de geração de visualizações como nuvem de palavras mais usadas e gráficos, indicando dias e horários com maior utilização.

Este conjunto de ferramentas consolida do LABCOM – Laboratório de Convergência de Mídias, no campo da pesquisa aplicada, com orientação à solução de problemas reais de jornalistas e pesquisadores da área, articulando teoria, conhecimento empírico e desenvolvimento de artefatos via software, num esforço também de inovação, considerando a área de Ciências Sociais Aplicadas e, mais especificamente, Comunicação e Informação.

REFERÊNCIAS

CASTELLS, M. **A sociedade em rede. São Paulo**: Paz e Terra, 1999.

CHWE, M. Communication and coordination in social networks in: **Review of Economic Studies**, 67, p. 128-156, 2000.

DRESCH, A.; LACERDA, D.; ANTUNES J.. **Design Science Research**: método de pesquisa para avanço da ciência e tecnologia. Porto Alegre: Bookman, 2015.

EASLEY, D.; KLEINBERG, J. **Networks, Crowds and Markets**: reasoning about a highly connected world. Nova York: Cambridge University Press, 2010.

FEENBERG, A. **Transforming technology**: a critical theory revisited. New York: Oxford University Press, [E-book], 2002.

GLEICK, J. **A informação**: uma história, uma teoria, uma enxurrada. São Paulo, Companhia das Letras, 2013.

LEMOS, A. **Cibercultura**: tecnologia e vida social na cultura contemporânea. 4. ed. Porto Alegre: Sulina, 2002.

MANOVICH, L. **The language of new media**. Massachusetts: Mit Press. 2001.

MORIN, E. **Introdução ao pensamento complexo**. 4.ed. Porto Alegre: Sulina, 2005.

NEWMAN, M. E. **Networks: an introduction**. Nova York: Oxford University Press, 2010.

ROGERS, R. **Digital Methods**. Cambridge: Mit Press. [E-book], 2013.

SANTAELLA, L. **Culturas e artes do pós-humano**: da cultura das mídias à cibercultura. São Paulo: Paulus, 2003.

SANTOS, M. **Comunicação digital e jornalismo de inserção**: como big data, inteligência artificial, realidade aumentada e internet das coisas estão mudando a produção de conteúdo informativo. São Luís: LABCOM Digital, 2016.

_____. **A datificação de um campo de conhecimento**: como algoritmos, números e abordagens quantitativas estão mudando a comunicação. In: Organicom , Ano 16, nº 31, 2019.

SIMON, H. **The architecture of complexity**. In: Proceedings of the American Philosophical Society. v. 106, n. 6. dez, 1962.

VILCHES, L. **A migração digital**. São Paulo: Loyola, 2003.