

DATA SCIENCE: CIÊNCIA ORIENTADA A DADOS

Vivenciamos o quarto paradigma da ciência, o qual tem redefinido o *modus operandi* da práxis científica como consequência dos desafios impostos pela produção de dados em larga escala. A era *big data* também revolucionou o mundo dos negócios e vem exigindo uma nova postura das organizações para lidar com o grande volume e variedade de dados tanto estruturados, quanto não-estruturados, produzidos diariamente, de modo a subsidiar melhores decisões estratégicas.

Como resultado destas transformações na ciência e no mundo dos negócios, e como forma de responder às demandas existentes, observamos a expansão de uma área de estudo, relativamente recente, interdisciplinar e intensivamente computacional: a ciência de dados (*data science*). A ciência orientada a dados se vale do potencial de robustas ciberinfraestruturas de informação e comunicação, incluindo tecnologias de grids, e padrões que possibilitam a interoperabilidade e a interligação de dados (*linked data*). Tais padrões e ciberinfraestruturas sustentam as diferentes fases do ciclo de vida de grandes e heterogêneas coleções de dados disponíveis na web e em repositórios de dados digitais, com vistas a atribuir sentido e extrair insights de dados aplicáveis a diferentes domínios e contextos, para a resolução de problemas práticos e reais.

Buscando colaborar para o enriquecimento das discussões acerca da ciência de dados em âmbito nacional, este número temático da revista *Informação & Informação* apresenta um conjunto de reflexões considerando diferentes ângulos deste fenômeno, sob o olhar de pesquisadores da Ciência da Informação e áreas correlatas. Esta edição é composta por 11 contribuições, provenientes de 23 pesquisadores, que versam sobre aspectos conceituais e práticos, bem como desafios e oportunidades que a ciência orientada a dados apresenta à nossa comunidade científica e profissional.

O artigo “Ciência aberta: dimensões para um novo fazer científico”, das autoras Adriana Carla Silva de Oliveira e Edilene Maria da Silva, abre este

número especial, introduzindo o conceito de ciência aberta e dissertando sobre as dimensões que cercam as políticas de estímulo à prática científica, levando em conta os aspectos normativos, legais, institucionais e tecnológicos envolvidos, com vistas a oferecer uma visão multidimensional e multirrelacional da ciência aberta.

Com o objetivo de mapear a produção científica no tema, o artigo “Produção Internacional Sobre Ciência Orientada a Dados: análise dos termos Data Science e E-Science na Scopus e na Web of Science”, de autoria de Leilah Santiago Bufrem, Fábio Mascarenhas e Silva, Natanael Vitor Sobral e Anna Elizabeth Galvão Coutinho Correia, apresenta resultados de uma análise bibliométrica que identifica as principais ênfases temáticas e áreas de concentração das publicações sobre a ciência de dados entre 2006 e 2016.

Ainda sob o viés da comunicação científica, mas com foco nos aspectos das métricas alternativas (altmetrics), o artigo de autoria de Ronaldo Ferreira de Araújo e Ariadne Chloe Mary Furnival intitulado “Comunicação científica e atenção online: em busca de colégios virtuais que sustentam as métricas alternativas” disserta sobre a exposição, o alcance e impacto da produção científica nas mídias sociais, bem como sobre a composição de colégios virtuais para além das comunidades científicas tradicionais, por meio de um estudo exploratório realizado com o artigo científico de maior indicador segundo o Altmetric.com.

Em seguida, os autores Fernando Sayão e Luana Faria Sales apresentam no artigo “Algumas considerações sobre os repositórios digitais de dados de pesquisa”, as principais características, tipologias, funções e infraestruturas dos repositórios digitais científicos, bem como seus benefícios e requisitos funcionais durante todo o ciclo de vida dos dados, para o devido apoio à sustentabilidade da ciência aberta.

Também acerca da gestão de dados, Ricardo César Gonçalves Sant’Ana propõe no artigo “Ciclo de vida dos dados: uma perspectiva a partir da Ciência da Informação” um novo olhar para o ciclo de vida dos dados, amparando-se nos conceitos e contribuições da Ciência da Informação, mas também levando em conta o papel de outras áreas-chave como a Ciência da Computação, de modo a discutir todas as fases que compõem esse ciclo e

seus fatores de interferência, considerando os dados como elementos centrais do processo.

Os autores Moisés Lima Dutra e Douglas Dyllon Jeronimo de Macedo apresentam em “Curadoria Digital: proposta de um modelo para Curadoria Digital em ambientes Big Data baseado numa abordagem semi-automática para a seleção de objetos digitais”, técnicas para a seleção e avaliação de objetos digitais voltadas para curadoria digital, considerando o volume, a velocidade, a variedade, a veracidade e o valor dos dados coletados em múltiplos domínios do conhecimento, tomando por base métodos heurísticos que possam ser aplicados tanto por curadores de dados humanos, quanto por agentes de software.

Sob o prisma da mineração de dados e dos dados interligados, os autores Carlos Henrique Marcondes, Leonardo Cruz da Costa, e Sérgio de Castro Martins buscam contextualizar no artigo “Descoberta de conhecimento em artigos digitais em Ciências Biomédicas” a questão das publicações ampliadas e discutir questões sobre os modelos semânticos para artigos digitais, e sobre a aplicabilidade destes modelos para composição de publicações científicas com base em recursos interligados que sejam acessíveis e processáveis de forma automática.

Na mesma seara, mas com foco na atribuição de sentido aos dados disponíveis na web, no artigo “Web Semântica e Ontologias: um estudo sobre construção de axiomas e uso de inferências”, os autores José Eduardo Santarem Segundo e Caio Saraiva Coneglian discorrem sobre linguagem OWL, suas contribuições para a Web Semântica e para o Linked Data, e reportam testes de análise de inferências e de aplicação de axiomas como forma de agregar lógica e de aperfeiçoar a contextualização de dados textuais.

Aspectos relacionados à ligação de entidades e às dificuldades de mineração e atribuição de sentido automático ao tratar grandes quantidades de dados textuais em linguagem natural também são abordados por Eduardo Habib Bechelane Maia e Marcello Peixoto Bax no artigo intitulado “Um estudo Bibliográfico sobre Ligação de Entidades”, no qual são sintetizadas as principais abordagens de ligação de entidades e os principais gargalos destas técnicas.

Também relacionado às contribuições das técnicas e padrões atuais para a interligação de dados, o autor Rogério Aparecido Sá Ramalho em seu artigo “BIBFRAME: modelo de dados interligados para bibliotecas” apresenta o modelo Bibliographic Framework para descrição formal de relacionamentos de recursos de dados, e discute os principais desafios e perspectivas de uso deste novo modelo de dados em bibliotecas.

A ciência orientada a dados está se estabelecendo como uma área do conhecimento e de formação, balizada, em parte, nas técnicas e abordagens relatadas nos artigos supracitados. Desse modo, Renata Gonçalves Curty e Jucenir da Silva Serafim encerram este número temático com o artigo “A formação em ciência de dados: uma análise preliminar do panorama estadunidense”, no qual são identificados e discutidos os aspectos formativos do cientista de dados, incluindo as habilidades e competências esperadas e desejadas destes profissionais, por meio da análise de 93 cursos ofertados por instituições dos EUA. O estudo objetiva contribuir para a formação do cientista de dados em âmbito nacional, considerando o papel da Ciência da Informação nesse processo formativo.

Esperamos que a leitura do conjunto de reflexões que compõem este número temático gere novos e frutíferos debates, e traga maior visibilidade a aspectos concernentes à ciência de dados.

Renata Gonçalves Curty
Brígida Maria Nogueira Cervantes

Departamento de Ciência da Informação – Universidade Estadual de Londrina (UEL).