XXI ENANCIB
Encontro Nacional de Pesquisa em Ciência da Informação
50 anos de Ciência da Informação no Brasil: diversidade, saberes e transformação social
Rio de Janeiro • 25 a 29 de outubro de 2021

XXI Encontro Nacional de Pesquisa em Ciência da Informação – XXI ENANCIB

GT-8 – Informação e Tecnologia

INTELIGÊNCIA ARTIFICIAL E CIÊNCIA DE DADOS EM CRIS INSTITUCIONAL: MODELO CONCEITUAL

ARTIFICIAL INTELLIGENCE AND DATA SCIENCE IN INSTITUTIONAL CRIS: CONCEPTUAL MODEL

Caio Saraiva Coneglian - Centro Universitário Eurípides de Marília (UNIVEM)
Emanuelle Torino - Universidade Tecnológica Federal do Paraná (UTFPR); Universidade Estadual Paulista (UNESP)
Silvana Aparecida Borsetti Gregorio Vidotti - Universidade Estadual Paulista (UNESP)

Modalidade: Trabalho Completo

Resumo: A crescente disponibilização de dados e informações relacionadas à ecologia de pesquisa, em múltiplos sistemas de informação, culmina em crescente complexidade na atividade de gestão de pesquisa. Como resposta a esta complexidade, são desenvolvidos Sistemas de Informação de Pesquisa Corrente, do inglês *Current Research Information System* (CRIS) que objetivam o gerenciamento de metadados contextuais das atividades de pesquisa relacionadas a determinada instituição, quer seja de pesquisa ou de fomento. Neste estudo, um Modelo Conceitual de CRIS institucional é revisitado, visando o aprimoramento do processo, com o uso de Inteligência Artificial e Ciência de Dados. Como procedimento metodológico, utiliza a revisão bibliográfica para o embasamento teórico-conceitual a fim de contextualizar a Inteligência Artificial e a Ciência de Dados, incorporadas ao estudo. A partir disso, criou-se o modelo revisitado, inserindo uma camada de dados que trata dos aspectos de Ciência de Dados, bem como técnicas e métodos de Inteligência Artificial em todos os processos do CRIS, em especial, inseriram-se Processamento de Linguagem Natural, Visão Computacional, Mineração de Texto e Aprendizagem de Máquina. Conclui-se, assim, que a adaptação apresentada do modelo se mostra como um amadurecimento na própria compreensão que se tem do CRIS, com a inserção de elementos que tornam tal modelo mais atualizado. Assim, os estudos e a criação de modelos e de aplicações de CRIS permitem uma evolução na gestão institucional.

Palavras-Chave: CRIS. Inteligência Artificial; Ciência de Dados; Gestão de Pesquisa.

Abstract: *The increasing availability of data and information related to research ecology, in multiple information systems, culminates in increasing complexity in the research management activity. In response to this complexity, Current Research Information System (CRIS) are developed, which aim to manage contextual metadata of research activities related to a particular institution, whether research or development. In this study, an Institutional CRIS Conceptual Model is revisited, aiming at improving the process, using Artificial Intelligence and Data Science. As a methodological procedure, it uses the bibliographic review for the theoretical-conceptual basis to contextualize Artificial Intelligence and Data Science, incorporated into the study. From that, the revisited model was created, inserting a data layer, which deals with the Data Science aspects, as well as Artificial*

Intelligence techniques and methods in all CRIS processes, in particular, it was inserted Natural Language Processing, Computer Vision, Text Mining and Machine Learning. It is concluded, therefore, that the adaptation of the model presented shows itself as maturation in the very understanding that one has of CRIS, with the insertion of elements that make this model more up-to-date. Thus, the studies and creation of CRIS models and applications allow evolution in institutional management.

Keywords: *CRIS; Artificial intelligence; Data Science; Research Management.*

1 INTRODUÇÃO

A crescente disponibilização de dados e informações relacionadas à ecologia de pesquisa, em múltiplos sistemas de informação (a exemplo de plataforma de currículo, diretórios de grupos de pesquisa, identificadores persistentes de pesquisador, identificadores persistentes de publicações, fontes de publicação ou disponibilização de resultados de pesquisa, editais de fomento à pesquisa, laboratórios e equipamentos de pesquisa, instituições) culmina em crescente complexidade na atividade de gestão de pesquisa.

Como resposta a essa complexidade, são desenvolvidos Sistemas de Informação de Pesquisa Corrente, do inglês, *Current Research Information System* (CRIS) que objetivam o gerenciamento de metadados contextuais das atividades de pesquisa relacionadas a determinada instituição, quer seja de pesquisa ou de fomento.

Dessa forma, visando auxiliar instituições interessadas na implantação de Sistema de Informação de Pesquisa Corrente (CRIS), Torino, Coneglian e Vidotti (2020) propõem a compatibilização de estruturas de representação disponíveis nos diversos sistemas de informação relacionados à atividade de gestão e divulgação de resultados de pesquisa, por meio do uso de tecnologias já existentes para a constituição de CRIS institucional. Tal proposição culmina em um modelo conceitual de CRIS institucional, que possibilita “otimizar a infraestrutura necessária para a gestão de informações de pesquisa visando torná-las acessíveis e reutilizáveis, assegurando ganhos para todos os envolvidos.” (TORINO; CONEGLIAN; VIDOTTI, 2020, p. 21).

Considerando a necessidade de aperfeiçoar as análises e o processo de tomada de decisão a partir dos dados inseridos no CRIS, como aprofundamento da pesquisa supracitada, os autores, neste estudo, revisitam o Modelo conceitual de CRIS institucional para propor o aprimoramento do processo com o uso de Inteligência Artificial e Ciência de Dados. Ademais, destaca-se que pelo papel do CRIS, bem como pela quantidade de dados que o ambiente reúne, o uso da Ciência de Dados, apoiado pelas técnicas de *machine learning*, pode melhorar o seu uso por partes das instituições em função da inserção de inteligência e aprendizagem no processo.

Como procedimento metodológico, utiliza a revisão bibliográfica para o embasamento teórico-conceitual a fim de contextualizar Inteligência Artificial e Ciência de Dados, incorporadas ao estudo. A utilização da Inteligência Artificial e da Ciência de Dados ocorreu, devido à grande quantidade de dados geridos no ambiente CRIS, o que leva à necessidade do tratamento de tais dados por técnicas como *machine learning*, *Business Intelligence* e Processamento de Linguagem Natural. Como resultado é apresentado um modelo conceitual de integração dos dados de uma ecologia de pesquisa, capaz de subsidiar a composição de um CRIS institucional, utilizando Inteligência Artificial e Ciência de Dados.

2 MODELO DE CRIS INSTITUCIONAL

Tendo em vista a crescente necessidade de gerir informações institucionais, visando aprimorar e otimizar processos, infraestruturas e recursos, no âmbito das atividades de pesquisa, para gerenciar o ciclo de vida da pesquisa é imprescindível a utilização de *Current Research Information System* (CRIS), cuja tradução para o português é Sistema de Informação de Pesquisa Corrente, também conhecido como Gerenciamento de Informações de Pesquisa, tradução de *Research Information Management* (RIM).

CRIS é definido como:

> [...] sistema de informação que pode gerenciar toda a informação relevante da pesquisa, começando com oportunidades de financiamento, passando pelo estágio de redação e submissão de propostas, seguindo com as propostas bem-sucedidas, que se tornam projetos ativos que serão gerenciados até a conclusão – estágio no qual são gerados resultados, muitos dos quais são publicações ou algum outro artefato da atividade de pesquisa. (JOINT, 2008, p. 571, tradução nossa).

Sheppard (2010) destaca que o CRIS precisa interagir com sistemas e processos da instituição, envolvendo recursos humanos e financeiros, informações de alunos e

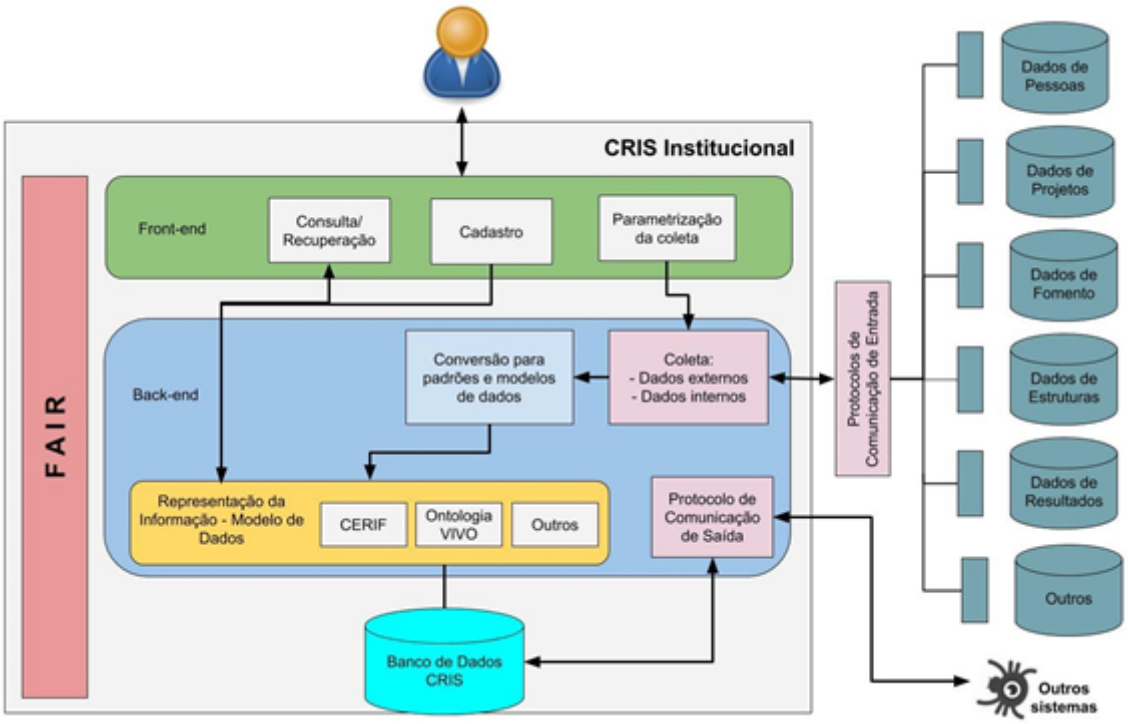
colaboradores, biblioteca, e, interoperar com sistemas externos, especializados em financiamento.

Elucida-se assim que o CRIS gerencia dados relacionados à atividade de pesquisa e seus relacionamentos, que inclui: projetos; pesquisadores; organização; resultados, que consistem em publicações, dados, patentes, produtos; indicadores; equipamentos; instalações; serviços; financiamento; medição (LOPATENKO, 2001; SALES; SAYÃO, 2015; euroCRIS, 2020).

Torino, Coneglian e Vidotti (2020) apontam que, embora essas informações estejam disponíveis, em geral, estão em sistemas de informação distintos. Inegavelmente, a gestão em um único ambiente possibilita melhorar a visão acerca do cenário da ecologia de pesquisa de uma instituição. Assim, destacam os autores que o CRIS pode se constituir como plataforma única ou por meio da integração de dados de múltiplos sistemas de informação que armazenam dados relacionados ao ciclo de vida da pesquisa.

Deste modo, para auxiliar na estruturação de um ecossistema de pesquisa que se constitua como CRIS institucional, a partir da coleta de dados provenientes de múltiplas fontes internas e externas à instituição e da compatibilização das estruturas de representação da informação, Torino, Coneglian e Vidotti (2020), propõem um Modelo de CRIS institucional (Figura 1). O referido modelo considera padrões de referência, como o *Common European Research Information Format* (CERIF), e outras tecnologias da web semântica.

Figura 1 - Modelo de CRIS institucional



Fonte: Torino, Coneglian e Vidotti (2020).

O modelo conceitual proposto por Torino, Coneglian e Vidotti (2020) possibilita a coleta e o armazenamento de dados provenientes de múltiplas fontes que gerenciam dados e informações concernentes ao ciclo de vida da pesquisa, como: os dados de organização, infraestrutura para pesquisa, projetos de pesquisa, grupos de pesquisa, pesquisador (pessoa), fomento, resultados de pesquisa, e os compatibiliza visando constituir um CRIS.

O modelo (Figura 1) é constituído de *front-end* e *back-end*, responsáveis por diferentes processos. No *front-end* é possível realizar entradas manuais de dados, mediante 'Cadastro', bem como realizar a 'Parametrização das Coletas' por meio da qual a alimentação automática ocorre. Enquanto no *back-end* as entradas automáticas são processadas de acordo com a parametrização com o uso dos 'Protocolos de Comunicação de Entrada', que aciona a coleta de dados provenientes de bancos de dados internos e externos à instituição, nos quais são coletados dados de pessoas, de projetos de pesquisas, de fomento à pesquisa, de estruturas para o desenvolvimento de pesquisa e de resultados de pesquisa (como: publicações, dados de pesquisa, patente, *software* e demais objetos relacionados; status, identificador persistente, fonte de publicação). Os dados coletados passam por um processo de conversão para os padrões e modelos de dados adotados no CRIS e, assim, são armazenados no banco de dados. Ainda no *back-end*, na camada de 'Representação da

Informação', está contido o modelo de dados que pode utilizar ontologias, como a VIVO e o CERIF, não limitando-se a elas. A 'Consulta/Recuperação' aos dados armazenados pode ser realizada por humanos via interface gráfica no *front-end*, e por aplicações computacionais por meio do 'Protocolo de Comunicação de Saída' no *back-end*.

Assim, a implantação de um CRIS institucional requer a definição de um modelo de dados estruturados, considerando ontologias, padrões de metadados, protocolos de comunicação entre diferentes sistemas e o uso de tecnologias que possibilitem a sua construção.

No presente estudo, o Modelo de CRIS institucional é revisitado, visando aprimorar o processo com o uso de Inteligência Artificial e Ciência de Dados.

3 INTELIGÊNCIA ARTIFICIAL E CIÊNCIA DE DADOS

A Inteligência Artificial (IA) é uma área de estudos que há muitos anos vem sendo pesquisada, renovada e aprofundada. Em especial, nas últimas décadas a IA passou a ser parte do cotidiano das pessoas e foco de estudo para diversas áreas do conhecimento. Áreas tradicionais como medicina e direito estão se apropriando das tecnologias geradas a partir da IA e estão realizando um grande avanço ao criar equipamentos e ambientes autônomos.

No entanto, estudos que se assemelham ao que chamamos hoje de Inteligência Artificial começaram a ser realizados na década de 1950, em especial com pesquisas vinculadas à linguística, na busca de aprimorar a representação do conhecimento por parte das máquinas. Desde então, a IA vem sendo atualizada e expandida, com uma série de outras aplicações e definições.

Em uma clássica visão que reúne diversas definições da Inteligência Artificial, Russell e Norvig (2016, p. 5, tradução nossa) apontam quatro categorias principais: "[1] Sistemas que pensam como humanos. [2] Sistemas que agem como humanos. [3] Sistemas que pensam racionalmente. [4] Sistemas que agem racionalmente."

As categorias apresentadas demonstram como a Inteligência Artificial é um amplo campo de estudo que envolve diferentes visões e aplicações, que podem reunir desde algoritmos de aprendizagem de máquina, passando por instrumentos capazes de reconhecer vozes e imagens, até a automação de processos humanos por robôs.

A partir da visão apresentada, podemos verificar dois exemplos de IA que têm objetivos e concepções totalmente distintas: o *chatbot* que busca agir com um humano, por

meio de conversas; e os sistemas de reconhecimento de placas que pensam e agem de forma racional, uma vez que o objetivo é apenas identificar quais são as placas dos veículos de forma automatizada.

Diante desse cenário, podemos compreender a Inteligência Artificial por meio dos seus diversos campos de pesquisa que possuem objetivos diferentes. Nesse contexto, destacamos os campos de pesquisa relacionados a este estudo: Processamento de Linguagem Natural e Aprendizagem de Máquina.

Primeiramente, destaca-se a área de Processamento de Linguagem Natural. Vieira e Lopes (2010, p. 184) apontam que: "Processamento de Linguagem Natural (PLN) é uma área de Ciência da Computação que estuda o desenvolvimento de programas de computador que analisam, reconhecem e/ou geram textos em linguagens humanas, ou linguagens naturais.". Essa definição insere a área de PLN dentro da Ciência da Computação, tendo, porém, interdisciplinaridades, em especial com a Linguística, pela necessidade de compreensão e tratamento da linguagem natural e humana.

Outra definição afirma que

> O Processamento de Linguagem Natural é uma gama teoricamente motivada de técnicas computacionais para analisar e representar textos que ocorrem naturalmente em um ou mais níveis de análise linguística com a finalidade de obter processamento de linguagem semelhante ao humano para uma série de tarefas ou aplicações. (LIDDY, 2001, p. 1, tradução nossa).

Outro importante campo de estudo é o *machine learning* ou aprendizagem de máquina. Esse é um dos campos mais estudados e aplicados atualmente. Existem diversos sistemas que inserem tais tecnologias no dia a dia das pessoas, criando uma série de facilidades para diversas funções e se tornando um elemento indispensável para as organizações. Por exemplo, assistentes virtuais de bancos e operadores de telemarketing virtual utilizam outros elementos da Inteligência Artificial, como o processamento de linguagem natural, mas tem no aprendizado de máquina a sua principal função: a capacidade de "aprender".

Uma definição de Aprendizagem de Máquina é:

> O aprendizado de máquina é uma disciplina focada em duas questões inter-relacionadas: "Como construir sistemas de computador que melhoram automaticamente com a experiência? e Quais são as leis fundamentais da teoria estatística da informação computacional que governam todos os sistemas de aprendizagem, incluindo computadores, seres humanos e organizações?" (JORDAN; MITCHELL, 2015, p. 255, tradução nossa).

A partir do que foi apresentado pelos autores, é possível identificar que a aprendizagem de máquina traz a premissa da autonomia das máquinas para favorecer a criação de sistemas mais eficientes para as pessoas. Ademais, a definição apresentada insere aspectos vinculados a estatística, que está bastante presente nos estudos de aprendizagem de máquina.

Outro campo de estudos, vinculado à área de Inteligência Artificial, é a mineração de texto (em inglês *text mining*). Tal campo é definido como: "[...] a descoberta por computador de novas informações anteriormente desconhecidas, extraindo automaticamente informações de diferentes recursos escritos." (HEARST, 2003, p. 1, tradução nossa).

Por meio dessa definição, verifica-se que a área de mineração de textos tem como objetivo encontrar novas informações que, por vezes, não estão explícitas no texto para um agente computacional. Com o apoio de técnicas de *machine learning*, a mineração de textos pode aumentar a compreensão e a identificação de elementos desconhecidos inicialmente.

Outra temática relacionada à IA é a Visão Computacional apontada, em um clássico texto da área, por Poggio, Torre e Koch (1985) como um campo de estudos da Inteligência Artificial centrado nos estudos de processamento de informações visuais. Tal definição demonstra a amplitude deste tema, demonstrando que o foco está no processamento da informação, abarcando pesquisas e técnicas que, de alguma forma, tratam as imagens por meio de técnicas computacionais.

A área de Ciência de Dados ganhou destaque nos últimos anos devido ao contexto da expansão da quantidade de dados disponíveis, bem como do uso dos dados para apoiar processos de tomada de decisão.

Os estudos da área de Ciência de Dados estão fortemente vinculados à compreensão sobre o *Big Data*. Nesse contexto:

> Antes do big data, nossa análise geralmente se limitava a uma pequena quantidade de hipóteses que definíamos bem antes de coletarmos os dados. Quando deixamos que os dados falem por si, podemos gerar conexões que nem sabíamos que existiam. Assim, alguns fundos hedge [forma de investimento alternativa] usam o Twitter para prever o desempenho do mercado de ações. A Amazon e a Netflix baseiam suas recomendações de produtos nas diversas interações em seus sites. Twitter, LinkedIn e Facebook mapeiam o 'gráfico social' das relações entre os usuários para aprender mais sobre suas preferências. (MAYER-SCHÖNBERGER; CUKIER, 2013, p. 9).

Diante do exposto, verifica-se que há uma mudança no modo como a tomada de decisões acontece, no contexto da Ciência de Dados, em que por meio de técnicas e algoritmos de Inteligência Artificial é possível extrair padrões, realizar predições e apoiar a construção de sistemas computacionais que obtém mais valor de seus dados.

Nesse sentido, a Ciência de Dados contempla todo o processo vinculado à captura, ao tratamento, à análise, ao processamento e à disponibilização dos dados. Complementando tal visão:

> O que diferencia a ciência dos dados das estatísticas é que a ciência dos dados é uma abordagem holística. Estamos cada vez mais encontrando dados na natureza, e cientistas de dados estão envolvidos com a coleta de dados, massageando-o em uma forma tratável, fazendo-o contar sua história, e apresentando essa história para outras pessoas. (BARLOW, 2011, p. 3).

A partir do que foi apresentado, evidencia-se que a Ciência de Dados possui um papel amplo no processo de manuseio dos dados nas mais diversas fases. A abordagem holística apresentada pelo autor demonstra a importância e a necessidade de aplicar a Ciência de Dados quando tem-se cenários de Big Data que, entre outros aspectos, possui grandes volumes de informações.

Neste trabalho utilizou-se algumas ferramentas que se vinculam à Ciência de Dados, sendo eles: *Business Intelligence* (BI), processamento distribuído e *data lake.*

Primeiramente, o BI é apontado como "[...] o processo de transformação de dados brutos em informações utilizáveis para maior efetividade estratégica, insights operacionais e benefícios reais para o processo de tomada de decisão nos negócios." (DUAN; XU, 2012).

Outro aspecto vinculado ao contexto da Ciência de Dados é o processamento paralelo ou distribuído de grandes conjuntos de dados. Em suma, busca-se utilizar diferentes unidades de processamento, ou mesmo diferentes unidades computacionais, para promover o processamento dos dados existentes. Nesse sentido, a promoção de técnicas e algoritmos que permitam o processamento distribuído é essencial para que grandes quantidades de dados possam ser tratadas.

Por fim, os *data lakes* são apresentados como ambientes que reúnem conjuntos de dados brutos para que a partir deles sejam aplicadas técnicas de *machine learning* e estatística para extrair novos conhecimentos. Nesse contexto, Mathis (2017, p. 293, tradução

nossa) aponta que: “Mover dados para um local de armazenamento central para simplificar a análise e gerar valor a partir dos dados é a ideia principal por trás dos *data lakes*.”

Desta forma, unindo a Inteligência Artificial e a Ciência de Dados, pode-se trazer evoluções para diversas áreas e temáticas de estudos, como no caso o CRIS que é apresentado a seguir.

4 USO DE INTELIGÊNCIA ARTIFICIAL E CIÊNCIA DE DADOS EM CRIS

Estudos apontam a relevância da utilização de CRIS para gestão de dados de pesquisa e Ciência Aberta (BALL; BROWN; MOLLOY; VAN den EYNDEN; WILSON, 2015; SCHÖPFEL; PROST; REBOUILLAT, 2017; BIESENBENDER; PETERSOHN; THIEDIG, 2019) o que torna mais latente a necessidade de implantação de CRIS institucional para a gestão das atividades de pesquisa.

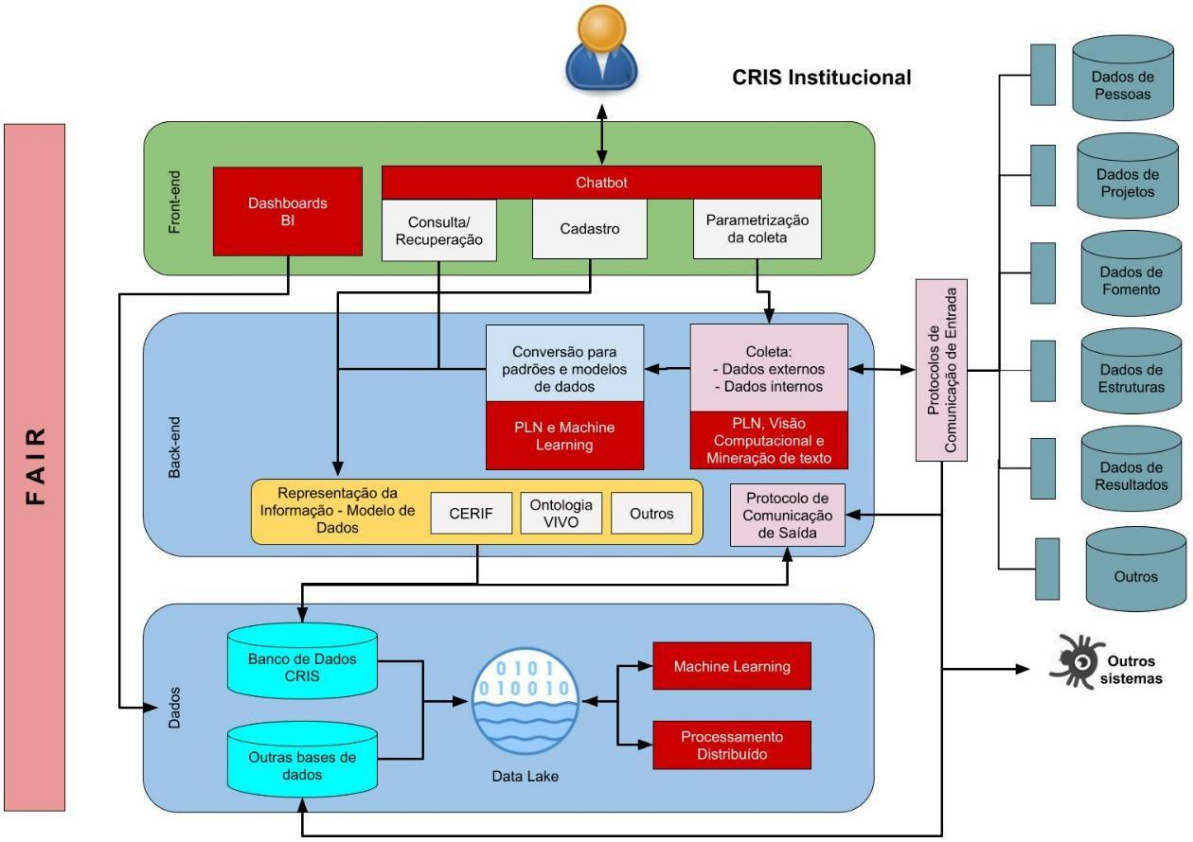
No Brasil, as discussões acerca das temáticas de gestão de dados de pesquisa e da Ciência Aberta são crescentes. Contudo, a disponibilização de CRIS ainda é inexpressiva. Tal afirmação tem como base a busca realizada no *Directory of Research Information Systems* (DRIS), que registra 866 CRIS em todo o mundo, sendo apenas um deles brasileiro, a Plataforma Sucupira cadastrada pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

Assim, visando auxiliar as instituições brasileiras interessadas na disponibilização de CRIS, Torino, Coneglian e Vidotti (2020) apresentaram um modelo conceitual para a integração de dados provenientes de múltiplos sistemas de informação que gerenciam dados e informações do ciclo de vida da pesquisa.

A partir da compreensão do Modelo de CRIS institucional, refletiu-se sobre a inserção da Inteligência Artificial e da Ciência de Dados ao modelo, buscando aprimorar a proposta e trazer mais autonomia para o processo.

A Figura 2 apresenta o Modelo de CRIS institucional com a inserção dos aspectos da Inteligência Artificial e da Ciência de Dados.

Figura 2 - Modelo de CRIS institucional com o uso de Inteligência Artificial e Ciência de Dados



Fonte: Elaborado pelos autores a partir de Torino, Coneglian e Vidotti (2020).

A Figura 2 apresenta alguns elementos essenciais para a compreensão desta adaptação do Modelo de CRIS institucional, que ocorreram nas camadas de *front-end*, *back-end* e, principalmente, na inclusão da camada de dados.

Ainda que no modelo anterior já existisse uma base de dados que tratava dos dados armazenados no CRIS institucional, no novo modelo os dados passam a ter um maior protagonismo, buscando aprimorar análises de dados que possam ser realizadas. Dessa forma, essa nova camada busca permitir que o CRIS se torne um ambiente informacional digital que otimize significativamente a tomada de decisão e, sob a perspectiva das atividades de pesquisa, seja um aglutinador dos dados de uma instituição.

Assim, a camada de dados visa permitir que as informações internas do CRIS e informações externas, coletadas diretamente de outras bases institucionais ou de ambientes informacionais externos, possam ser reunidas em um *data lake*. Um *data lake* é um ambiente que reúne os dados de forma bruta, potencializando a mineração e a descoberta de conhecimento. Dessa forma, tem-se neste ambiente tanto dados tratados que são

oriundos da base do CRIS, quanto dos ambientes externos, sem tratamento, o que favorece a descoberta de conhecimento e, quando apoiado por algoritmos de *machine learning* e técnicas de processamento distribuído, permitem o processamento computacional de uma grande massa de dados, além de favorecer as análises de dados.

O principal elemento ligado à camada de dados presente no *front-end* e que foi inserido nessa nova visão do modelo é o *Business Intelligence* (BI), por meio de seus *dashboards*. Compreende-se, aqui, por BI os principais elementos que permitem o acesso e a análise de dados para apoiar o processo de tomada de decisões. Nesse contexto, o BI seria uma plataforma que, ligada a camada de dados, possibilita uma visão ampla sobre os dados disponíveis da instituição, aprimorando o desempenho institucional ao apoiar a tomada de decisões, por exemplo, no credenciamento de programas de pós-graduação, criação de novos cursos, contratação de docentes, destinação de recursos, investimento em infraestrutura, entre outros.

Ao unir BI e camada de dados a um CRIS, busca-se tornar as instituições cada vez mais orientadas a dados, ou *data-driven*. A partir da compreensão de que o CRIS reúne informações vitais da parte central de uma instituição que atua com pesquisa, tem-se a possibilidade de aprimorar a tomada de decisão ao adotar esse modelo que apresenta elementos de Inteligência Artificial e Ciência de Dados.

Adicionalmente, outro elemento inserido no âmbito do modelo proposto é o *chatbot*, presente no *front-end*. Esse é um instrumento utilizado para promover a conversação entre um humano e um agente computacional que realiza tarefas, responde perguntas e busca informações. Os *chatbots* têm se popularizado justamente para facilitar o atendimento aos usuários não dependendo de pessoas, mas também são muito utilizados como uma nova forma mais natural e humana de se relacionar com os usuários.

Desta forma, propõe-se que o CRIS tenha como uma interface adicional um *chatbot* para responder questões, inserir novos dados e trazer as informações que o usuário necessita. O *chatbot* é capaz, portanto, de executar as funções normais do CRIS, porém utilizando um outro tipo de interface e relacionamento com o usuário.

Adicionalmente, na camada do *back-end* há duas aplicações principais da Inteligência Artificial e áreas vinculadas. Primeiramente, na coleta de dados, a aplicação de visão computacional e de processamento de linguagem natural com o intuito de, a partir de objetos digitais, extrair conteúdo e metadados existentes. Também na coleta de dados, o

processo de mineração de texto pode apoiar na descoberta de palavras e conceitos significativos que são parte dos documentos, podendo enriquecer a representação da informação.

Na sequência, no contexto da conversão dos dados coletados para os padrões definidos de representação da informação, utiliza-se dois campos de aplicações principais vinculados à Inteligência Artificial. Por meio do Processamento de Linguagem Natural, apoiado por técnicas de *machine learning*, é possível realizar uma conversão mais aprimorada dos dados coletados para padrões e modelos de dados.

Por meio da identificação de conceitos com uso de PLN, compreensão do sentido com o apoio de *machine learning*, entre outras técnicas, tem-se uma melhor classificação e identificação das classes e entidades a serem inseridas nas ontologias e outros instrumentos.

Diante do modelo apresentado, verifica-se que algumas tecnologias, métodos e técnicas vinculados à Inteligência Artificial e à Ciência de Dados foram a base para a adaptação apresentada no modelo. Em especial, as técnicas de *machine learning* propiciam análises de dados capazes de extrair mais elementos e *insights.* Além delas, o uso de *chatbot* e PLN permite com que algoritmos avançados para compreensão dos termos criem um ambiente digital capaz de se relacionar de forma mais natural com os humanos.

Outro elemento essencial está na Visão Computacional, que se vincula à Inteligência Artificial, e permite que documentos digitalizados possam ser inseridos no CRIS e que o significado de dados e metadados possam se tornar explícitos para a máquina, em especial, quando apoiados por técnicas de mineração de texto.

Por fim, a camada de dados criada para apoiar os processos de Ciência de Dados busca contribuir com o modelo ao aprimorar o processo de tomada de decisão com os dados pertencentes ao domínio do CRIS. Dessa forma, a partir de técnicas de Ciência de Dados, junto ao uso de aprendizagem de máquina, tem-se um modelo mais adaptável, inteligente e orientado a dados.

5 CONSIDERAÇÕES FINAIS

O CRIS pode ser utilizado para o gerenciamento de diferentes aspectos da pesquisa e fornece elementos para o planejamento estratégico de instituições de pesquisa ou de fomento, uma vez que provê aos envolvidos, sejam eles pesquisadores, técnicos de apoio, gestores de projetos de pesquisa ou de agências de fomento, informações sobre o

andamento e o avanço da pesquisa, bem como indicadores de desempenho. Ademais, a utilização de Inteligência Artificial e Ciência de Dados amplia os benefícios do uso de CRIS, dentre os quais destacam-se a visão abrangente da atividade de pesquisa, a redução de custos e de esforços, além da otimização do processo de tomada de decisão.

Neste trabalho, o Modelo de CRIS institucional com o uso de Inteligência Artificial e Ciência de Dados apresenta contribuições para tornar esse ambiente orientado a dados, trazendo uma tendência da sociedade atual em que a análise de dados, com apoio de aprendizagem de máquina, passa a ser essencial nos mais diversos sistemas de informação.

Vale destacar, ainda, que a adaptação apresentada do modelo mostra-se como um amadurecimento na própria compreensão que se tem do CRIS, com a inserção de elementos que enriquecem o modelo.

Enquanto trabalho futuro, busca-se aplicar o modelo apresentado visando aprimorar a ecologia de pesquisa institucional e contribuir com a gestão estratégica e orientada a dados.

REFERÊNCIAS

BALL, A.; BROWN, C.; MOLLOY, L.; VAN den EYNDEN, V.; WILSON, D. **Using CRIS to power research data discovery**. Disponível em: http://dspacecris.eurocris.org/handle/11366/378. Acesso em: 01 jun. 2020.

BARLOW, M. **What is Data Science**. [*S.l.*]: O'REILLY, 2011.

BIESENBENDER, S.; PETERSOHN, S.; THIEDIG. C. Using Current Research Information Systems (CRIS) to showcase national and institutional research (potential): research information systems in the context of Open Science. **Procedia Computer Science**, n. 146, p. 142-155, 2019. Trabalho apresentado na 14ª International Conference on Current Research Information Systems, CRIS, 2018. DOI: https://doi.org/10.1016/j.procs.2019.01.089. Disponível em: https://www.sciencedirect.com/science/article/pii/S1877050919300948. Acesso em: 04 jan. 2021.

DUAN, L; XU, L. D. Business intelligence for enterprise systems: A survey. **IEEE Transactions on Industrial Informatics**, v. 8, n. 3, p. 679-687, 2012. euroCRIS. CERIF support blog. Disponível em: https://www.eurocris.org/eurocris_archive/cerifsupport.org/. Acesso em: 04 jul. 2020.

HEARST, M. What is text mining. **SIMS**, UC Berkeley, v. 5, 2003. Disponível em: https://www.jaist.ac.jp/~bao/MOT-Ishikawa/FurtherReadingNo1.pdf. Acesso em: 09 jun. 2021.

JOINT, N. Current research information systems, open access repositories and libraries: ANTAEUS. **Library Review**, v. 57, n. 8, p. 570-575, 2008. Disponível em: https://doi.org/10.1108/00242530810899559. Acesso em: 19 set. 2017.

JORDAN, M. I.; MITCHELL, T. M. Machine learning: Trends, perspectives, and prospects. **Science**, v. 349, n. 6245, p. 255-260, 2015.

LIDDY, E. D. Natural language processing. *In*: KENT, A. (ed.). **Encyclopedia of Library and Information Science**. 2nd. New York: Ed. NY. Marcel Decker, Inc., 2001.

LOPATENKO, A. S. Information retrieval in Current Research Information Systems. 2001. Disponível em: https://arxiv.org/abs/cs/0110026. Acesso em: 22 jun. 2020.

MATHIS, C. Data lakes. **Datenbank-Spektrum**, v. 17, n. 3, p. 289-293, 2017.

MAYER-SCHÖNBERGER, V; CUKIER, K. **Big data**: como extrair volume, variedade, velocidade e valor da avalanche de informação cotidiana. 1. ed. Rio de Janeiro: Elsevier, 2013.

POGGIO, T.; TORRE, V.; KOCH, C. Computational vision and regularization theory. **Readings in computer vision**, p. 638-643, 1987.

RUSSELL, S. J.; NORVIG, P. **Artificial intelligence**: a modern approach. Malaysia: Pearson Education Limited, 2016.

SALES, L. F.; SAYÃO, L. F. Ciberinfraestrutura de informação para a pesquisa: uma proposta de arquitetura para a integração de repositórios e sistemas CRIS. **Informação & Sociedade: Estudos**, João Pessoa, v. 25, n. 3, p. 163-184, set./dez. 2015. Disponível em: https://periodicos.ufpb.br/ojs/index.php/ies/article/view/23998. Acesso em 11 ago. 2020.

SCHÖPFEL, J.; PROST, H.; REBOUILLAT, V. Research data in Current Research Information Systems. **Procedia Computer Science**, v. 106, p. 305-320, 2017. DOI: https://doi.org/10.1016/j.procs.2017.03.030. Disponível em: https://www.sciencedirect.com/science/article/pii/S1877050917302983. Acesso em: 01 jun. 2020.

SHEPPARD, N. Learning how to play nicely: repositories and CRIS. **Ariadne**, n. 64, jul. 2010. Disponível em: http://www.ariadne.ac.uk/issue64/wrn-repos-2010-05-rpt. Acesso em: 01 ago. 2020.

TORINO, E.; CONEGLIAN, C. S.; VIDOTTI, S. A. B. G. Estruturas de representação para reuso de dados no contexto da ecologia de pesquisa: CRIS Institucional. **Informação & Informação**, Londrina, v. 25, n. 3, p. p. 1-27, jul./set. 2020. DOI: http://dx.doi.org/10.5433/1981-8920.2020v25n3p1. Disponível em: http://www.uel.br/revistas/uel/index.php/informacao/article/view/41946. Acesso em: 24 nov. 2020.

VIEIRA, R; LOPES, L. Processamento de Linguagem Natural e o tratamento computacional de linguagens científicas. **EM CORPORA**, p. 183, 2010.