Uma Análise Comparativa de Abordagens para Extração de Redes de Colaboração Científicas

Bruna Tavares Silva[1], **Rebeca Schroeder**[1]

[1]Departamento de Ciência da Computação
Universidade do Estado de Santa Catarina (UDESC)
Centro de Ciências Tecnológicas – Joinville – SC – Brasil

silvatavares.bruna@gmail.com , rebeca.schroeder@udesc.br

Abstract. *This paper compares approaches related to the extraction and query of data from digital libraries focused on analysis and contruction of scientific collaboration networks. As result, a comparative analysis of related works is presented, which highlights issues related to extraction methods, data models for representation, among other characteristics.*

Resumo. *Este artigo apresenta e discute alguns trabalhos relacionados à extração e consulta de informações de bibliotecas digitais com foco em análise e formação de redes de colaboração científica. Como resultado, uma análise comparativa entre estes trabalhos é apresentada, a qual evidencia questões relacionadas aos métodos de extração, modelos de dados para representação, dentre outras características.*

1. Introdução

Redes estão em todos os lugares [Santos et al. 2017] e podem ser físicas, como as organizações corporativas, ou virtuais. Uma rede social pode ser classificada como virtual quando é composta por indivíduos conectados por meio dos mais variados tipos de relações [Digiampietri 2015, Digiampietri et al. 2017, Menezes 2012]. Entre as redes sociais existentes há àquelas originadas no âmbito científico, que são as Redes de Colaboração Científica formadas por acadêmicos, professores e pesquisadores em geral [Ciacia 2017]. Nestas redes, as interações entre seus elementos são oriundas de trabalhos realizados em conjunto.

As redes de colaboração científica são comumente representadas através de grafos, onde os vértices são os pesquisadores e as arestas representam cooperações entre pesquisadores. Os pesquisadores podem cooperar de diversas maneiras, tais como: participações em projetos, relacionamentos de co-autoria de publicações, orientações em dissertações e teses, participações em banca de defesa de dissertações e teses, produções técnicas, participações em comissões examinadoras, dentre outros tipos de relações científicas [Menezes 2012]. Na Figura 1 é apresentado um exemplo de rede de colaboração científica focada em relacionamentos de co-autoria. As arestas caracterizam publicações entre pesquisadores, cuja espessura indica a quantidade de trabalhos em coautoria e as cores indicam as áreas de atuação [Dias et al. 2016]. Esta representação na forma de grafos habilita a aplicação de diversas métricas aplicadas a esta estrutura de dados. Ademais, algoritmos de classificação, agrupamento e ranqueamento de pesquisadores são aplicados às redes de colaboração científica por diversos trabalhos da literatura [Ciacia 2017].

Figura 1. Exemplo de uma rede de colaboração científica [Dias et al. 2016]

A análise de redes de colaboração científica permite conhecer a comunidade acadêmica, identificar grupos de trabalho, medir o desempenho de pesquisadores e de sua força de trabalho, dentre outras possibilidades. Entretanto, para que esta análise seja possível, a produção destas redes envolve a extração e o processamento de dados obtidos a partir de fontes em que as informações relevantes estejam disponíveis. Em virtude das diversas plataformas científicas que disponibilizam estes dados, existem divergências entre elas com relação às informações que podem ser extraídas, e aos métodos de extração disponíveis. Como resultado, trabalhos que focam na produção de redes de colaboração científicas ou se limitam aos dados disponíveis por sua fonte base, ou consideram um conjunto heterogêneo de fontes para formar sua estrutura de dados. Este artigo apresenta algumas destas abordagens e analisa seus respectivos métodos para extração de dados, criação e visualização de redes de colaboração científica. O objetivo deste trabalho é contribuir com uma análise comparativa de um conjunto de trabalhos representativo da área, avaliando aspectos como tipos de dados extraídos, fontes de dados e diversos aspectos relacionados ao processo de extração e representação das redes.

Este artigo está organizado em mais 4 seções. Na seção 2 são apresentados algumas plataformas de dados científicos, a partir das quais as redes de colaboração científicas podem ser extraídas e analisadas. A Seção 3 apresenta abordagens para extração, visualização e análise de redes de colaboração científica. Estas abordagens são comparadas na Seção 4 a partir de diversos aspectos, como a natureza dos dados extraídos, as plataformas e formatos de extração, bem como a estrutura de dados utilizada para representação das redes. Por fim, as conclusões deste artigo são apresentados na Seção 5, juntamente com a identificação de trabalhos futuros na área.

2. Plataformas de Dados Científicos

Atualmente, uma grande quantidade de informações referentes à produção científica está disponível na Web, como por exemplo, publicações científicas, informações sobre projetos de pesquisa e currículos de pesquisadores [Digiampietri 2015]. Essas bases de dados de publicações científicas podem ser específicas por áreas de conhecimento, como o

DBLP [1], a ACM [2], gerais como o IEEE [3], Google Acadêmico [4] ou ainda bases de dados sobre pesquisadores como a Plataforma Lattes.

2.1. DBLP

O projeto DBLP (*Digital Bibliography & Library Project*) é uma biblioteca digital de produções acadêmicas na área de Ciência da Computação. Os dados do projeto DPLP podem ser acessados diretamente pela página do projeto ou é possível copiar a base do projeto em arquivos XML. As informações contidas nos arquivos XML podem ser sobre: artigos, anais de conferências, conferências, livros, capítulos de livros, teses de doutorado, dissertações de mestrado e páginas web. As informações sobre cada tipo de publicação seguem o formato BibTex e podem incluir: nome dos autores, editores, endereço da publicação, título do trabalho, páginas, ano de publicação, veículo de publicação, número, mês, url, instituição, notas, isbn, entre outras.

Além de prover uma interface simples para obter as publicações de um pesquisador ou de um veículo de publicação, a DBLP é conhecida também pela precisão de seus dados, pois uma chave de identificação é atribuída a cada entidade referente às publicações. Este tratamento evita ambiguidades, um problema comum em outras fontes de dados como a Plataforma Lattes [Ciacia 2017].

2.2. ACM

A ACM (*Association for Computing Machinery*) publica, distribui e arquiva produções dos pesquisadores do mundo todo nas áreas de computação e tecnologia da informação. A ACM também publica revistas, boletins informativos e anais de conferência. A Biblioteca Digital da ACM é um dos bancos de dados mais abrangentes do mundo de artigos com textos completos e demais produções bibliográficas da área. Ela armazena citações e textos completos de suas publicações. Cada citação possui links para outras obras do autor, referências clicáveis as suas fontes originais, identificadores de objetos digitais (DOIs) que gerenciam facilmente as conexões eletrônicas. Os dados de cada publicação estão disponíveis nos formatos BibTex, Endnote, ACMRef e CSV.

2.3. IEEE

A IEEE (Instituto de Engenheiros Elétricos e Eletrônicos) é uma associação dedicada ao avanço da inovação e excelência tecnológica. A Biblioteca Digital IEEE Xplore (*IEEE Xplore Digital Library*) é a biblioteca oficial de conteúdo científico e técnico publicado pela IEEE e seus parceiros de publicação. Com aproximadamente 20.000 novos documentos adicionados a cada mês, tornou-se um grande atrativo para fontes de dados de diversas pesquisas.

Um dos diferenciais da Biblioteca Digital IEEE é o sistema de busca avançado, que possibilita a busca através dos campos de referências aos trabalhos, como nome do autor, ano de publicação, tipo de trabalho, palavras-chave, entre outros. Diferentes argumentos de busca podem ser combinados através de operadores lógicos. Além disso é possível exportar os resultados da busca no formato CSV, ou exportar as referências de cada trabalho retornado nos formatos: txt, BibTex, RIS e RefWorks.

[1]DBLP: `http://dblp.uni-trier.de/`
[2]ACM: `https://libraries.acm.org/`
[3]IEEE: `https://ieeexplore.ieee.org/Xplore/home.jsp`
[4]Google Acadêmico: `https://scholar.google.com.br/`

2.4. Google Acadêmico

Diversos estudos envolvendo o Google Acadêmico (em inglês *Google Scholar*) surgiram nos últimos anos. Ao fazer uma pesquisa, o Google Acadêmico providencia diversos tipos de filtros como: data de publicação (mais novos, mais antigos, seleção de um ano específico), classificação (por relevância - padrão, data, idioma dos resultados). Para cada publicação, são apresentadas informações como o número de citações, artigos relacionados, links para cada um dos autores, link para o arquivo em PDF se disponível, o ano de publicação, a editora, entre outras informações.

As duas principais características do Google Acadêmico a serem destacadas são suas funcionalidades como meta-buscador e índice de citações. Como meta-buscador ele reúne informações das diversas bases de dados do texto completo em uma única interface simples e aberta. Como índice de citações, interliga os diversos documentos a partir de suas referências e revela a rede de conexões [Mugnaini et al. 2008].

2.5. Plataforma Lattes

A Plataforma Lattes é um repositório de dados sobre pesquisadores do Brasil lançada e padronizada pelo CNPq [5] em 1999. Em maio de 2018, a plataforma contava com mais de 569 milhões de currículos cadastrados. Os currículos Lattes são utilizados em diversas pesquisas para a criação e análise de redes de colaboração científicas, por exemplo, análises de redes de pesquisadores de diversas áreas [Gasparini et al. 2017], extração e consulta de informações [Mena-Chalco et al. 2012], criação de modelos para a construção de uma rede social científica multi-relacional [Menezes 2012]. Esses diversos trabalhos indicam que os currículos Lattes são amplamente utilizados na literatura.

Os currículos Lattes são disponibilizados tanto no formato HTML quanto no formato XML. A versão XML é mais adequada para o processamento automático, pois possui todas as seções e campos dos currículos bem delimitados [Digiampietri 2015]. O acesso atual ao currículo Lattes (tanto na versão HTML quanto na XML) ocorre através da Web, no qual é necessário obter uma validação do sistema *Captcha*. Outra possibilidade de acesso aos currículos lattes é através de um *Web Service* provido pelo CNPQ, que é disponibilizado as instituições de ensino brasileiras por meio de um IP único para cada instituição. Os dados de um currículo estão divididos em cinco seções principais, a saber: dados gerais, produção bibliográfica, produção técnica, outras produções e dados complementares.

É necessário ressaltar que as informações presentes nos currículos são preenchidas pelo pesquisador e a atualização das mesmas é de responsabilidade do pesquisador, logo não é possível garantir a atualização das informações. Existem alguns mecanismos de controle de informações, por exemplo obrigatoriedade de certos campos. Entretanto, não há garantias da completude e corretude das informações preenchidas. A plataforma também disponibiliza outras ferramentas como a busca de currículos, na qual pode-se realizar uma busca textual de currículos através da Web ou busca avançada incluindo filtros por assunto, nacionalidade, bolsistas de produtividade do CNPq, formação acadêmica/titulação, atuação profissional, idioma, atividade profissional, dentre outros.

A visualização de redes de colaboração é também um recurso da Plataforma Lattes. Entretanto, além de mostrar apenas a visão de um único pesquisador, o recurso está

[5]Lattes Histórico: `http://memoria.cnpq.br/web/portal-lattes/historico`

disponível apenas através do navegador Internet Explorer. Os Diretórios de Pesquisa, é outro mecanismo provido pela Plataforma Lattes, que provê dados gerais do pesquisador bem como dados sobre os grupos de pesquisa no qual ele está inserido, as linhas de pesquisa em que atua, uma lista com os estudantes (cadastrados no currículo do pesquisador) orientados pelo pesquisador e os indicadores de produção.

3. Abordagens para Extração de Redes de Colaboração Científicas

Esta seção apresenta e discute alguns trabalhos identificados na literatura. Foram selecionados trabalhos que envolvessem a extração e consulta de informações de bibliotecas digitais com foco em análise de redes de colaboração científica ou extração de currículos. Os dados extraídos de todos os trabalhos aqui relacionados são considerados dados abertos, sem restrições de uso e disponíveis para consulta.

3.1. Digiampietre e Silva (2011)

O trabalho de [Digiampietri and Silva 2011] apresenta um *framework* para análise e visualização de redes sociais de pesquisadores, com o objetivo de identificação e análise de redes sociais de pesquisadores que possuem currículo na plataforma Lattes. O *framework* executa um processo completo de descoberta de conhecimento, iniciando pela busca da URL dos currículos de uma lista de pesquisadores. Para tanto, as APIs de pesquisa do Google e do Bing são utilizadas, onde são pesquisados o nome do pesquisador seguido pela informação "currículo lattes". Uma vez identificada a URL de cada currículo, os mesmos são obtidos e suas informações extraídas. Este processo é executado para cada pesquisador da lista, para posterior identificação de relacionamentos entre os pesquisadores. Dois tipos de relacionamentos são identificados: coautorias e áreas de interesse correlatas. Com essas informações, uma rede social é produzida e pode ser manipulada por meio de uma ferramenta gráfica. Um resumo da produção intelectual de cada pesquisador também pode ser gerado pela ferramenta.

O sistema de visualização apresenta as redes sociais na forma de um grafo, onde os nós são pesquisadores que associam-se a outros pesquisadores por coautorias e áreas de atuação correlatas. Na representação de coautorias, a espessura das arestas é proporcional à quantidade de publicações que os pares de pesquisadores realizaram em conjunto. Além disto, a ferramenta permite verificar algumas métricas sobre a rede social, como por exemplo, a distância entre um nó selecionado, que representa um pesquisador, e sua vizinhança. Por fim, o relatório de resumo da rede contém informações sobre a Rede Social de Pesquisadores considerando as regras definidas pelo Comitê de Computação da CAPES, com o objetivo de combinar a produção intelectual dos pesquisadores com a classificação de periódicos e conferências, juntamente com os demais critérios de avaliação dos programas de pós-graduação.

3.2. Mena-Chalco, Digiampietri e Cesar-Jr (2012)

No trabalho de [Mena-Chalco et al. 2012] foram desenvolvidos algoritmos para identificação automática de coautorias em produções bibliográficas, bem como a caracterização topológica de redes de coautoria de grupos de pesquisadores da Plataforma Lattes.

Um estudo sobre os algoritmos desenvolvidos considerou 176.114 currículos Lattes de pesquisadores associados à Grande Área de Ciências Exatas e da Terra. Esses dados

foram obtidos da plataforma Lattes em maio de 2011, a partir de um *parser* desenvolvido pelos autores que baixava os currículos automaticamente no formato HTML. Como resultado, foi obtida uma rede de coautoria e os resultados de algumas métricas relacionadas a grafos. A maior componente conexa dessa rede é composta de 63.066 pesquisadores (aproximadamente 36% da rede), e o diâmetro do grafo é 1.037, o que quantifica o número de arestas entre os dois nós mais distantes no grafo. Outras sete redes de coautoria foram criadas, sendo essas discretizadas por triênios, o que permitiu inspecionar o incremento na colaboração acadêmica ao longo do tempo.

3.3. SOS Lattes

A ferramenta SOS Lattes (*Semantic Ontology-based Script Lattes*), tem por objetivo auxiliar nas tarefas de extração e consulta de informações da Plataforma Lattes nos moldes da Web Semântica. O SOS Lattes combina a ferramenta ScriptLattes para extração de dados, sendo esta modificada para atuar como um serviço e exportar os dados no formato OWL-XML (*Web Ontology Language* XML).

O ScriptLattes é uma ferramenta livre e de código aberto sob a licença GNU. Como entrada para a ferramenta é necessário incluir uma lista de identificadores de Currímpculos Lattes. A ferramenta então faz o *download* automático dos currículos Lattes em formato HTML de um grupo de pessoas de interesse. Como resultado são geradas páginas HTML com listas de produções e orientações. Adicionalmente são criados diversos grafos (redes) de colaboração entre os membros do grupo de interesse e um mapa de geolocalização dos membros e alunos (de pós-doutorado, doutorado e mestrado) com orientação concluída [Mena-Chalco and Junior 2009].

O funcionamento do SOS Lattes inicia com técnicas de organização dos dados em ontologias utilizando o *Semantic Lattes*, uma ferramenta que realiza a importação de currículos e lista de veículos de publicações científicas e transforma para arquivos padrões de referência como XML entre outros [da Costa and Yamate 2009]. Além disto, a ferramenta utiliza a base de conhecimento do OntoLattes[Bonifacio 2002], e incrementa esta ontologia que comporta dados de currículos Lattes com a inclusão de novas regras da ferramenta de inferência e atualização das bibliotecas.

3.4. Digiampietri (2015)

Uma análise da rede social de pesquisa do evento BraSNAM foi desenvolvida pelos autores de [Digiampietri 2015]. No estudo, foram apresentados conceitos de rede social, grafos, métricas utilizadas na análise de redes sociais e medidas relacionadas a análise bibliométrica. O estudo também abordou aspectos relacionados à Plataforma Lattes, como a obtenção e organização de dados de currículos Lattes, atualização, corretude e completude dos dados, bem como aspectos relacionados à obtenção, organização e refinamento dos dados. O estudo foi realizado antes da adoção do sistema *Captcha* e utilizou de algoritmos automáticos para extração dos currículos Lattes.

Entre as etapas da análise da rede social brasileira, estão a obtenção, organização e refinamento dos dados [Digiampietri 2015]. O trabalho apresentou três estratégias para obtenção dos identificadores dos currículos, incluindo um processamento inicial dos currículos, a criação de um banco de dados relacional, o enriquecimento do conjunto de dados por meio de: fator de impacto JCR (*Thompson's Journal Citation Reports*); índice SJR (*Scimago Journal Rank*); e Qualis.

3.5. Ciacia (2017)

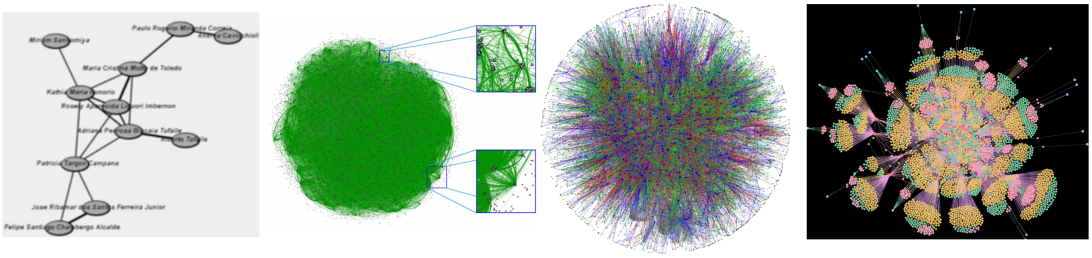
O trabalho de [Ciacia 2017] analisou e selecionou os 29 autores mais prolíficos da área de IHC. O estudo teve como objetivo analisar as redes de colaboração científica da comunidade brasileira de IHC. Neste trabalho, foram também utilizadas técnicas de KDD (*Knowledge Discovery in Database*) ou descoberta/extração de conhecimento.

Para criação de redes de colaboração, os currículos da plataforma Lattes foram baixados manualmente em formato XML. Em seguida, foram armazenados em um banco de dados relacional para a análise, limpeza e padronização dos dados. Após essa etapa, as tabelas foram exportadas, em formato CSV, para um banco de dados orientado a grafos, o Neo4J. Ao considerar todos os dados providos pela plataforma (por exemplo informações sobre o autor, orientações, dados do resumo de cada autor, publicações, entre outros) foram identificadas cerca de 2.949 coautorias entre autores das publicações.

Os resultados foram obtidos através de análises bibliométricas, estatísticas e análises das redes sociais criadas. Utilizou-se também de uma adaptação da medida do Número de Erdös para formar uma medida de distância de um autor qualquer ao autor mais influente da comunidade. Verificou-se que as publicações em trabalhos em eventos atingiu mais de 74% de todas as publicações identificadas, sendo 75% dessas na forma de artigos completos, evidenciando uma preocupação em publicar resultados finais dos trabalhos. Além dos dados bibliométricos, diversas análises foram efetuadas sobre as redes de colaboração geradas pelos 29 autores mais prolíficos da comunidade de IHC.

4. Análise Comparativa

Nas seções anteriores foram apresentados alguns trabalhos relacionados à extração e geração de redes de colaboração científica. Constatou-se nos trabalhos a existência de etapas fundamentais dentre elas: Seleção dos dados a serem extraídos; Extração dos dados a partir das plataformas; Limpeza e padronização dos dados; Persistência dos dados e Geração de redes de colaboração científica. Na Figura 2 é apresentada a saída em forma de grafos das redes de colaboração de alguns trabalhos analisados. O trabalho de [Ferreira Galego and Wassermann 2013] não pode ser incluído pois os resultados são dados na forma de ontologias.



(1) Digiampietri and Silva 2011 **(2)** Mena-Chalco et al. 2012 **(3)** Digiampietri 2015 **(4)** Ciacia 2017

Figura 2. Redes de coautorias geradas pelos trabalhos analisados

A Tabela 1 compara os trabalhos relacionados de acordo com a fonte de dados e os dados selecionados para extração. A biblioteca dominante dentre os trabalhos foi a Plataforma Lattes. Outras como IEEE, DBLP e Capes também foram listadas. Além disso

alguns trabalhos utilizaram de outros tipos de dados para enriquecimento da extração, como o Qualis para classificação, o Google *Scholar* para contagem de citações, entre outros.

Tabela 1. Bibliotecas e Tipos de Dados Extraídos

Trabalhos	Dados Extraídos	Bibliotecas
Digiampietri and Silva 2011	Produções Bibliográficas, Áreas de interesse	Lattes, Capes
Mena-Chalco et al. 2012	Produções Bibliográficas	Lattes
Ferreira Galego and Wassermann 2013	Produções Bibliográficas, Orientações	Lattes, IEEE
Digiampietri 2015	Currículos, Formações, Áreas de Atuação, Projetos de Pesquisas, Produções Bibliográficas e Orientações	DBLP, Lattes, Google Scholar, Qualis, Índice JCR, Scimago Journal Rank, Microsoft Academic Search, Thompson's Journal Citation Reports
Ciacia 2017	Produções Bibliográficas, Orientações, Formação, Resumo CV, Endereço, Idioma, Prêmios	Lattes

Sobre os dados extraídos identificou-se a predominância das produções bibliográficas, cuja característica já era esperada devida à natureza das redes de colaboração científica. Outros tipos como orientações e áreas de interesse também são listados. Nota-se pela Tabela 1 que a Plataforma Lattes é amplamente utilizada, indicando a sua relevância para pesquisas e validade ao longo dos anos, já que os trabalhos estão distribuídos no período de 2011 a 2017. Outros trabalhos utilizaram de bases de dados adicionais para enriquecer as redes como a Capes, Qualis, DBLP, *Google Scholar*, entre outras [Digiampietri and Silva 2011, Digiampietri 2015, Ferreira Galego and Wassermann 2013].

Tabela 2. Comparação quanto à Extração dos Dados e suas Características

Trabalhos	Objetivo	Método de Seleção	Formato de Extração	Critérios de Extração
Digiampietri and Silva 2011	Extração Produção de Redes Sociais de Pesquisadores	Automático	HTML	Lista de Pesquisadores
Mena-Chalco et al. 2012	Extração Produção de Redes de Coautoria	Automático	HTML	Áreas de Conhecimento ID's Correlatos
Ferreira Galego and Wassermann 2013	Extração Consulta de Informações da Plataforma Lattes	Automático	HTML XML (ScriptLattes)	Lista de Pesquisadores
Digiampietri 2015	Extração Produção de Redes de Colaboração Científica	Automático	HTML XML	Lista de Pesquisadores (Doutores no Brasil)
Ciacia 2017	Extração Produção e Análise de Redes de Colaboração Científica	Manual	XML	Autores Mais Prolíficos de IHC

Um fato importante a ser considerado é o da escolha pela automatização da seleção e extração dos dados, perceptível na Tabela 2. Entretanto, é importante observar que a automatização destes processos era devida à utilização de softwares que operavam antes da inclusão do sistema *Captcha* na plataforma Lattes, fazendo com que esse método esteja indisponível atualmente.

Quanto aos critérios de extração, 'Lista de Pesquisadores' é amplamente elegido pelos trabalhos relacionados aqui descritos. Entretanto, outros tipos, como 'Eventos e Grupos de Pesquisa' poderiam agregar dados às redes de colaboração e favorecer análises adicionais. Nos trabalhos relacionados não foram elencadas essas opções, instigando assim um interesse especial nos possíveis resultados gerados.

Na Tabela 3 foram analisadas as estruturas de formação das redes de colaboração. Notou-se que o modelo de representação das redes de colaboração em grafos é o mais utilizado, exceto pelo trabalho de [Ferreira Galego and Wassermann 2013] que está focado na produção de ontologias. Outra característica identificada são os diferentes tipos de dados modelados como nós. Por exemplo, o trabalho de [Ciacia 2017] representou como nós: veículo de publicação, instituições e pesquisadores. Sobre os atributos dos nós, apenas dois incluíram dados como atributos. Já em relação aos atributos das arestas, ficou evidente a omissão quanto a esta categoria. Uma possibilidade de atributo dos relacionamentos/arestas é incluir a quantidade de produções onde dois pesquisadores participaram como co-autores, indicando a quantidade de cooperações desta natureza.

Tabela 3. Comparação quanto à Geração de Redes de Colaboração Científicas

Trabalhos	Modelo	Entidades	Relacionamentos	Atributos das Entidades	Atributos Relacionamentos
Digiampietri and Silva 2011	Grafos	Pesquisadores	Coautorias, Subáreas Compartilhadas	-	-
Mena-Chalco et al. 2012	Grafos Listas de adjascências	Pesquisador	Coautorias	-	-
Ferreira Galego and Wassermann 2013	Ontologias	Pesquisador, Orientação, Publicação	Coautorias	Equivalentes aos atributos de cada tipo de publicação no Lattes	-
Digiampietri 2015	Grafos Relacional	Pesquisadores, Cidades	Coautorias	-	-
Ciacia 2017	Grafos, Relacional	Pesquisador, Orientando, Coautor, Veículo, Instituição	Coautorias, Doutorado, Mestrado, 'Publicou', Vínculo	Pesquisador: LattesID, Nome, País, Estado, Cidade, Data Atualização, Grau Instrução, Instituição, Endereço, Idiomas, nPaper	-

5. Conclusão

Este artigo relacionou um conjunto representativo de trabalhos que propõem abordagens para extração de dados e produção de redes de colaboração científicas. Estes trabalhos foram comparados sob diferentes aspectos como: tipos de dados extraídos plataformas de dados para extração, métodos de seleção e extração, e o modelo de dados empregado pelas redes de colaboração. Uma constatação foi a importância da Plataforma Lattes como uma fonte para geração das redes. Isto é devido pela quantidade de dados e detalhes disponíveis aos pesquisadores do Brasil. Entretanto, a obtenção automática de currículos da plataforma se tornou indisponível após o ano de 2015 com a adoção do sistema *Captcha*. Para tanto, considera-se que o restabelecimento desta possibilidade é fundamental para a correta execução de alguns dos trabalhos avaliados, bem como de trabalhos futuros.

Alguns dos problemas envolvidos nos dados obtidos de plataformas digitais incluem a ambiguidade de dados, o preenchimento incorreto, a desatualização dos dados, entre outros. Estes problemas podem ser decorrentes dos próprios autores, assim como dos mecanismos de atualização das próprias plataformas. Para minimizar estas questões, acredita-se que a utilização de diversas plataformas em conjunto possa ser uma solução, pois habilitaria o cruzamento, confirmação e complementação dos dados.

Referências

[Bonifacio 2002] Bonifacio, A. S. (2002). Ontologias e consulta semântica : uma aplicação ao caso lattes. Dissertação de mestrado, Universidade Federal do Rio Grande do Sul. Instituto de Informática. Programa de Pós-Graduação em Computação.

[Ciacia 2017] Ciacia, F. (2017). Geração das redes de colaboração científica da comunidade acadêmica de ihc. Trabalho de conclusão de curso, Universidade do Estado de Santa Catarina, Joinville, Brasil.

[da Costa and Yamate 2009] da Costa, A. P. and Yamate, F. S. (2009). Semantic lattes: uma ferramenta de consulta de informações acadêmicas da base lattes baseada em ontologias. Trabalho de conclusão de curso, Escola Politécnica da Universidade de São Paulo, São Paulo.

[Dias et al. 2016] Dias, T. M. R., Moita, G. F., and Dias, P. M. (2016). Adoção da plataforma lattes como fonte de dados para caracterização de redes científicas. *Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação*, 21(47):16.

[Digiampietri 2015] Digiampietri, L. A. (2015). *Análise da Rede Social Acadêmica Brasileira*. PhD thesis, Escola de Artes, Ciências e Humanidades, Universidade de São Paulo.

[Digiampietri et al. 2017] Digiampietri, L. A., Mugnaini, R., Pérez-Alcázar, J. J., Delgado, K. V., Tuesta, E. F., Us, J., and Mena-Chalco, P. (2017). Análise da evolução, impacto e formação de redes nos cinco anos do BraSNAM. In *BraSNAM 2017*, pages 497–503.

[Digiampietri and Silva 2011] Digiampietri, L. a. and Silva, E. E. (2011). A Framework for Social Network of Researchers Analysis. *Iberoamerican Journal of Applied Computing*, 1(1):1–24.

[Ferreira Galego and Wassermann 2013] Ferreira Galego, E. and Wassermann, R. (2013). Extração e Consulta de Informações do Currículo Lattes Baseadas em Ontologias. In *X Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*, page 12.

[Gasparini et al. 2017] Gasparini, I., de Mendonça, F. C., Silveira, M. S., Diniz, S., Barbosa, J., and Schroeder, R. (2017). Crossing the borders of IHC. In *IHC 2017*, pages 1–10.

[Mena-Chalco et al. 2012] Mena-Chalco, J. P., Digiampietri, L. a., and Cesar-Jr, R. M. (2012). Caracterizando as redes de coautoria de currículos Lattes. *BraSNAM - Brazilian Workshop on Social Network Analysis and Mining*, 1(1):12.

[Mena-Chalco and Junior 2009] Mena-Chalco, J. P. and Junior, R. M. C. (2009). scriptLattes: an open-source knowledge extraction system from the Lattes platform. *Journal of the Brazilian Computer Society*, 15(1):31–39.

[Menezes 2012] Menezes, V. S. d. A. (2012). Análise de Redes Sociais Científicas. *Coppe/Ufrj*, 33:140–142.

[Mugnaini et al. 2008] Mugnaini, R., Strehl, L., and Strehl, L. (2008). Recuperação e impacto da produção científica na era google: uma análise comparativa entre o google acadêmico e a web of science 10.5007/1518-2924.2008v13nesp1p92. *Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação*, 13(1):92–105.

[Santos et al. 2017] Santos, D. V., Cunha, T. C., Silva, A. B. O., Parreiras, F. S., and Gomes, O. A. (2017). Comparação de Técnicas de Predição de Links em Sub-redes de Coautoria Formada por Currículos da Plataforma Lattes. In *BraSNAM*, pages 611–622.