Análise de rede sociais de autores e instituições participantes nos anais do WIDAT 2017 a 2023 por coautoria e palavras-chave

Henrique Monteiro Cristovão
Doutor em Ciência da Informação, Universidade Federal do Espírito Santo (UFES), Vitória, ES,Brasil
Professor, Universidade Federal do Espírito Santo (UFES), Vitória, ES, Brasil
http://lattes.cnpq.br/5035919384923489

Lucas dos Santos do Vale
Mestrando em Ciência da Informação, Universidade Federal do Espírito Santo (UFES), Vitória, ES, Brasil
Professor, Secretaria Estadual de Educação (SEDU), Vitória, ES, Brasil
http://lattes.cnpq.br/0502914406473197

Submetido em: 13/02/2024. Aprovado em: 08/05/2024. Publicado em: 18/07/2024.

RESUMO

O Workshop de Informação, Dados e Tecnologia (WIDaT), com suas seis edições de 2017 a 2023, publicou 210 artigos nos anais com a participação de 395 autores vinculados a 80 instituições sediadas em 46 cidades e 7 países. O objetivo da pesquisa é revelar grupos de afinidade entre autores e entre instituições participantes do WIDaT. Pesquisa qualitativa de natureza aplicada, utilizou técnicas de visualização de informação com o apoio de métodos de análise de redes sociais suportados por softwares, tais como: OpenRefine, Gephi e Looker Studio. Obteve como principal resultado a análise oriunda de redes sociais de autores e instituições participantes dos seis eventos. A rede de autores por coautoria mostrou um componente gigante composto pela maioria dos autores, enquanto a rede de autores por palavras-chave amplificou a densidade de conexões, tornando-a um 'mundo pequeno'. A rede de instituições por coautoria revelou uma forte colaboração entre as instituições. A rede de instituições por palavras-chave identificou sete *clusters*, complementados por nuvens de palavras-chave, que revelaram uma coesão semântica dos temas tratados pelas instituições pertencentes a cada *cluster*. A observação das palavras-chave mais incidentes para um determinado *cluster* pode motivar aproximações para colaboração de pesquisa tanto para aquelas já consolidadas quanto para fortalecer o desenvolvimento de novas investigações e estudos.

Palavras-chave: visualização de informação; análise de redes; projeção bipartida; clusterização; WIDaT.

INTRODUÇÃO

O Workshop de Informação, Dados e Tecnologia (WIDaT) é um evento que completou em 2023 a sua sexta edição e, desde 2017,

> [...] foi idealizado com o intuito de reunir as comunidades acadêmicas e industriais que trabalham com dados no Brasil, por meio da oferta de um espaço de discussão e interação entre profissionais da indústria e pesquisadores das áreas de Ciência da Informação, Ciência da Computação, Engenharias e áreas afins (WIDAT, 2023, *online*).

Ao longo das suas seis edições[1], o WIDaT publicou 210 artigos em seus anais, com a participação de 395 autores vinculados a 80 instituições sediadas em 46 cidades e 7 países: África do Sul, Brasil, Canadá, Cuba, Espanha, Holanda e Portugal. A grande maioria das instituições participantes é do Brasil, cobrindo 21 estados da federação.

A análise de redes sociais (ARS) investiga ligações entre entidades sociais, como autores e instituições, e permite descobrir relacionamentos e afinidades não aparentes entre esses atores sociais. A ARS, realizada por inspeção visual, tem forte relação com a visualização de informação. A área da visualização de informação tem como objetivos a revelação de padrões invisíveis a partir de dados abstratos e a possibilidade de obtenção de novas percepções, e não apenas imagens bonitas (Chen, 2013). A representação desses dados abstratos pode ser feita em formato de gráficos e imagens diversas que favorecem a leitura de seus significados. Normalmente, aplicam-se processos de mineração de dados para potencializar revelações de relações não esperadas (Hand *et al*., 2001).

Técnicas inerentes à ARS destacam-se como importantes processos para a mineração de dados e a visualização de informação. A inspeção visual, no contexto da ARS, possui grande capacidade para identificação de características topológicas da rede e a revelação de relacionamentos que seriam difíceis de enxergar diretamente pela observação ou por meio cálculos e inferências realizadas diretamente sobre os dados da base que a originou (Nooy; Mrvar; Batagelj, 2018; Newman, 2010). A topologia da rede, e não os atributos de seus nós, fornece os elementos essenciais para obtenção de bons resultados advindos da ARS (Wasserman; Faust, 1994).

Os nós de uma rede podem ser agrupados segundo critérios diversos. Quando esses critérios são baseados nos seus atributos, ocorre uma classificação. Quando os critérios são baseados nas conexões e arrumações topológicas dos nós, ocorre uma clusterização. Assim, os cálculos para obtenção de uma clusterização remetem a padrões topológicos e, principalmente, o quanto os nós temos de potencial para se juntarem.

---

1 Em 2017, Florianópolis-SC, UFSC, com os anais disponíveis em: https://repositorio.ufsc.br/bitstream/handle/123456789/180265/Anais.do.WIDAT2017.pdf.
Em 2018, João Pessoa-PB, UFPB, com anais disponíveis em:https://dadosabertos.info/enhanced_publications/idt/.
Em 2019, Brasília-DF, UnB, com os anais disponíveis em:http://widat2019.fci.unb.br/index.php/anais-widat-2019.
Em 2021, Belo Horizonte-MG, CEFET-MG, com os anais disponíveis em:https://pub.colnes.org/index.php/anis/issue/view/14.
Em 2022, Vitória-ES, UFES, com os anais disponíveis em:https://widat2022.ufes.br/wp-content/uploads/2023/04/widat-2022-anais.pdf.
Em 2023, Brasília-DF, IBICT, com os anais disponíveis em:https://labcotec.ibict.br/widat/index.php/widat2023. .

A clusterização de autores e instituições, participantes das seis edições do evento, em torno de interesses comuns, pode revelar associações interessantes com potencial capacidade de fomentar futuras coparticipações em projetos de pesquisa e auxiliar no entendimento das temáticas abordadas no evento. Dessa forma, o objetivo da pesquisa é revelar grupos de afinidade entre autores e entre instituições participantes do WIDaT por meio de ARS e visualização de informação.

METODOLOGIA

A pesquisa é qualitativa, de natureza aplicada e se utiliza, sobretudo, de técnicas de visualização de informação com apoio da ARS enquanto ferramenta metodológica (Matheus; Silva, 2006), uma vez que a ARS possibilita enxergar o que outras abordagens não permitem (Wasserman; Faust, 1994; Higgins, Ribeiro; 2018). Processos de descoberta de conhecimento (Fayyad *et al*., 1994) foram aplicados, principalmente baseados em projeções bipartidas[2] sobre as redes bipartidas[3] para torná-las mono partidas[4] e facilitar a sua análise e interpretação (Gao *et al*., 2017).

Todos os dados coletados, processados e utilizados ao longo da pesquisa, inclusive os arquivos-fonte dos resultados, estão disponíveis no repositório de dados da pesquisa[5].

Os dados (ano, título do artigo, autores, instituições, localizações e palavras-chave) foram coletados diretamente dos metadados nos anais dos seis eventos, organizados em duas planilhas e, posteriormente, pré-processados, limpos e organizados com o software OpenRefine[6].

Quanto à coleta das palavras-chave nos artigos dos anais, verificou-se a possibilidade de considerar termos usados no título. Contudo, essa possibilidade foi descartada tendo em vista que os *templates* para escrita do artigo para cada evento, especificamente quanto às orientações sobre a escrita das palavras-chave, não recomendaram palavras-chave diferentes das que compõem o título.

Com métodos semiautomáticos, baseados em algoritmos de agrupamento disponíveis no software OpenRefine[7], as palavras-chave foram agrupadas. Uma parte desses agrupamentos foi realizado com apoio manual, orientados pela análise de conteúdo de Bardin (1977), onde a categorização dos termos se desenvolve a partir de três etapas: (i) a pré-análise, feita de modo a compreender os termos iniciais de maneira abrangente bem como suas relações; (ii) a estruturação de categorias, realizada a partir do cerne temático, resumindo os termos

2 Projeção bipartida é uma ação que elimina um dos grupos de nós de uma rede bipartida, criando ligações entre os nós do conjunto que permaneceu na rede por meio dos nós eliminados.

3 Rede bipartida possui dois conjuntos de nós, de duas categorias distintas, onde as ligações somente são permitidas entre nós de um conjunto com os nós do outro conjunto.

4 Rede monopartida é formada por nós de uma única categoria e não possui restrição de ligações entre os seus nós, isto é, um nó pode se conectar com qualquer outro nó da rede.

5 Repositório com arquivos de dados e resultados da pesquisa, disponível em: https://github.com/hmcristovao/data_widat_2017_2023.

6 OpenRefine é uma ferramenta de código aberto utilizada para a limpeza e transformação de dados. Disponível em https://openrefine.org/.

7 A descrição dos métodos de agrupamentos semi-automáticos, fornecidos pelo software OpenRefine, está disponível em: https://openrefine.org/docs/manual/cellediting#cluster-and-edit.

resultantes da primeira etapa em categorias mais abrangentes; (iii) a interpretação e a inferência das categorizações a fim de identificar e eliminar possíveis ambiguidades fazendo-se uso da intuição e da análise reflexiva e crítica, conforme recomendam Sousa e Santos (2020).

O resultado dos agrupamentos realizados, foram organizados em dois quadros que encontram disponíveis em sua totalidade no repositório de dados da pesquisa. O primeiro quadro possui 66 termos resultantes de agrupamentos com seus respectivos termos selecionados e métodos empregados. Um recorte dele é mostrado no **Quadro 1**.

Quadro 1 – Recorte do agrupamento de alguns termos

Termos selecionados	Método	Termo equivalente
-dados abertos conectados **-dados conectados abertos**	Algoritmo Colisão de Chaves, Fingerprint	dados abertos conectados
-usuário **-usuários**	Algoritmo colisão de chaves, Ngram-Fingerprint (tamanho 1)	usuário
-ciclo de vida dos dados **-ciclo de vida de dados** **-ciclo de vida**	Algoritmo colisão de chaves, Ngram-Fingerprint (tamanho 1) Daitch-Mokotoff	ciclo de vida de dados
-programas de pós-graduação em ciência da informação **-pós-graduação em ciência da informação**	Algoritmo colisão de chaves, Ngram-Fingerprint (tamanho 1)	pós-graduação em ciência da informação
-dados de pesquisa **-dado de pesquisa**	Algoritmo colisão de chaves, Ngram-Fingerprint (tamanho 1)	dados de pesquisa
-software livre **-softwares livres**	Algoritmo colisão de chaves, Ngram-Fingerprint (tamanho 1)	software livre
-ciência de dados **-ciência dos dados**	Algoritmo colisão de chaves, Ngram-Fingerprint (tamanho 1)	ciência de dados
-análise de redes complexas **-análise de redes de informação** **-análise de redes sociais** **-análise de redes de pesquisa**	Algoritmo Colisão de Chaves, Metaphone3	análise de redes

Fonte: Autoria própria (2023)

O segundo quadro possui 113 termos adicionados a partir dos termos originais ou resultantes dos agrupamentos realizados. Um recorte dele é mostrado no **Quadro 2**.

Quadro 2–Recorte da derivação de alguns termos

Termo original ou resultante do agrupamento	Termo adicionado
acervo artístico universitário	cultura
acervos culturais	cultura
acesso a dados	acesso à informação
acesso à informação	acesso a dados
acesso aberto	acesso à informação
administração pública	setor público
altmetria	infometria
ambientes virtuais de aprendizagem	educação
anais	infometria
análise de redes	redes
análise exploratória visual	visualização
análise métrica da produção científica	infometria

Fonte: Autoria própria (2023).

Percebeu-se que, após as ações de agrupamentos e adição de termos, alguns deles ficaram duplicados para a mesma publicação, conforme mostra o **Quadro 3** que ilustra o problema para a publicação 'A' com a repetição do termo 'infometria'.

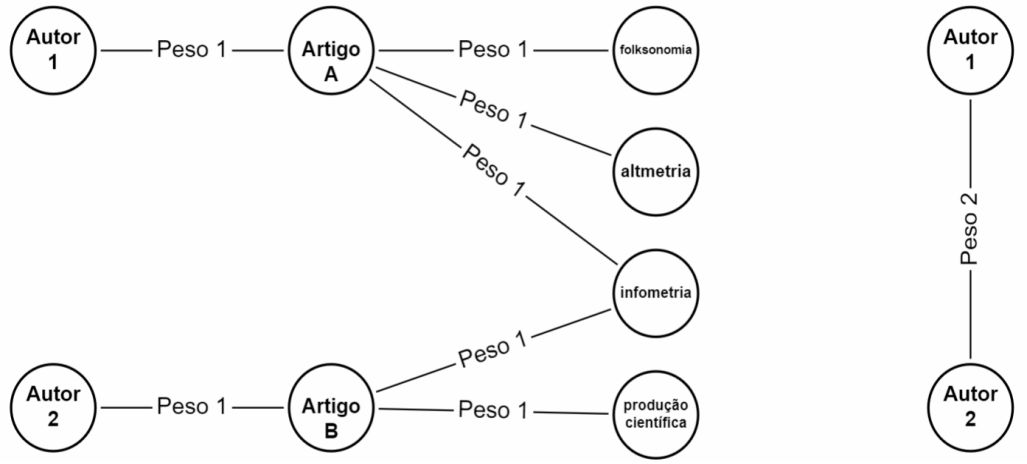
Quadro 3 – Demonstração do problema associado à adição de termos

Artigo	Termo agrupado	Termo adicionado
A	foksonomia	infometria
A	altmetria	~~infometria~~
B	produção científica	infometria

Fonte: Autoria própria (2023).

O cenário mostrado no **Quadro 3** demonstra a possibilidade de geração de pesos incorretos após a execução de projeções bipartidas na rede. Por exemplo, a eliminação do termo 'infometria' remete a uma correta rede monopartida de autores pela coautoria com o peso 2 na aresta que conecta 'Autor 1' com 'Autor 2', mostrada na **Figura 1**. Caso contrário, se a eliminação de 'infometria' não tivesse ocorrido, o peso entre 'Autor 1' e 'Autor 2' seria 3, e estaria incorreto. Em função disso, todos os termos repetidos, por publicação, foram eliminados da base.

Figura 1 – Rede tripartida corrigida e o resultado da dupla projeção bipartida na rede de autores



Fonte: Autoria própria (2023).

Para mostrar a geolocalização das participações ao longo dos seis eventos, foi realizada uma reconciliação de dados com suporte do OpenRefine para obtenção de latitude e longitude das localidades das instituições, utilizando-se a base de dados da Wikidata[8]. O software Looker Studio[9] foi usado para gerar elementos de visualização de informação, tais como, o mapa de geolocalização das instituições participantes e alguns gráficos estatísticos sobre os dados dos eventos.

As redes foram criadas no formato GML (Graph Modelling Language)[10] por meio de mapeamento realizado no software OpenRefine a partir dos dados coletados. Foram criadas categorias de nós para permitir as ações pontuais de projeção bipartida com o Gephi por meio do plugin Multimode Networks[11]. Os códigos dos mapeamentos realizados são mostrados nos **Quadro 4** e **Quadro 5**.

8 Wikidata é uma base de conhecimento livre que oferece vários serviços gratuitos, entre eles o serviço de reconciliação de dados. Disponível em: https://www.wikidata.org/.

9 Looker Studio é uma plataforma online para visualização de informação que gera gráficos estatísticos e *dashboards*. Disponível em: https://lookerstudio.google.com/.

10 GML (*Graph Modelling Language*) é um formato para representação de grafos de fácil leitura por humanos e com uma capacidade semântica razoável para configurar as características da rede, dos nós e das arestas. Disponível em: https://en.wikipedia.org/wiki/Graph_Modelling_Language/.

11 Multimode Networks é um plugin que adiciona a funcionalidade de projeção bipartida no ambiente do Gephi. Disponível em: https://github.com/jaroslav-kuchar/Multimode-Networks/.

Quadro 4 – Mapeamento para GML a partir dos dados de identificação da publicação e palavras-chave

```
graph [
       comment "rede não direcionada bipartida de chaveTituloAno e termo (repre-
       sentado pelos campos: 'termo', 'termosAdicionado1' e 'termosAdicionado2')"
       directed 0
       node [ id {{jsonize(cells.chaveTituloAno.value)}} categoria "chaveTitulo" ]
       node [ id {{jsonize(cells.termo.value)}} categoria "termo" ]
       node [ id {{jsonize(cells.termoAdicionado1.value)}} categoria "termo" ]
       node [ id {{jsonize(cells.termoAdicionado2.value)}}} categoria "termo" ]
       edge [ source {{jsonize(cells.chaveTituloAno.value)}} target {{jsonize(-
       cells.termo.value)}} ]
       edge [ source {{jsonize(cells.chaveTituloAno.value)}} target {{jsonize(-
       cells.termoAdicionado1.value)}} ]
       edge [ source {{jsonize(cells.chaveTituloAno.value)}} target {{jsonize(-
       cells.termoAdicionado2.value)}} ]
]
```

Fonte: Autoria própria (2023).

Quadro 5 – Mapeamento para GML a partir dos dados dos autores e instituições

```
graph [
       comment "rede não direcionada de 3-partida de chaveTituloAno, autorPseudo,
       instituicao"
       directed 0
       node [ id {{jsonize(cells.chaveTituloAno.value)}} categoria "chaveTitulo" ]
       node [ id {{jsonize(cells.autorPseudo.value)}} categoria "autorPseudo" ]
       node [ id {{jsonize(cells.instituicao.value)}} categoria "instituição" ]
       edge [ source {{jsonize(cells.chaveTituloAno.value)}} target {{jsonize(-
       cells.autorPseudo.value)}}]
       edge [ source {{jsonize(cells.chaveTituloAno.value)}} target {{jsonize(-
       cells.instituicao.value)}}]
]
```

Fonte: Autoria própria (2023).

A execução de ações inerentes à ARS com foco na inspeção visual foi realizada com o software Gephi[12]. Para a formatação das redes, foram usados algoritmos de distribuição do próprio software Gephi bem como o cálculo da métrica de modularidade para a formação dos *clusters* de instituições e de autores.

As nuvens de palavras dos ***clusters*** da rede de instituições, em função dos resultados dos agrupamento e adição de termos sobre as palavras-chave originais das publicações, foram criadas com auxílio da exportação de dados pelo Gephi, oriundos de uma rede bipartida de instituições e palavras-chave, com preparação dos dados em uma planilha e a geração do desenho da nuvem de palavras com o software WordArt[13].

12 GEPHI é um software de código aberto utilizado para visualização, análise e manipulação de redes e grafos. Disponível em https://gephi.org/.

13 WordArt é uma ferramenta online para geração de nuvem de palavras. Disponível em: https://wordart.com/.

ANÁLISE E DISCUSSÃO DOS RESULTADOS

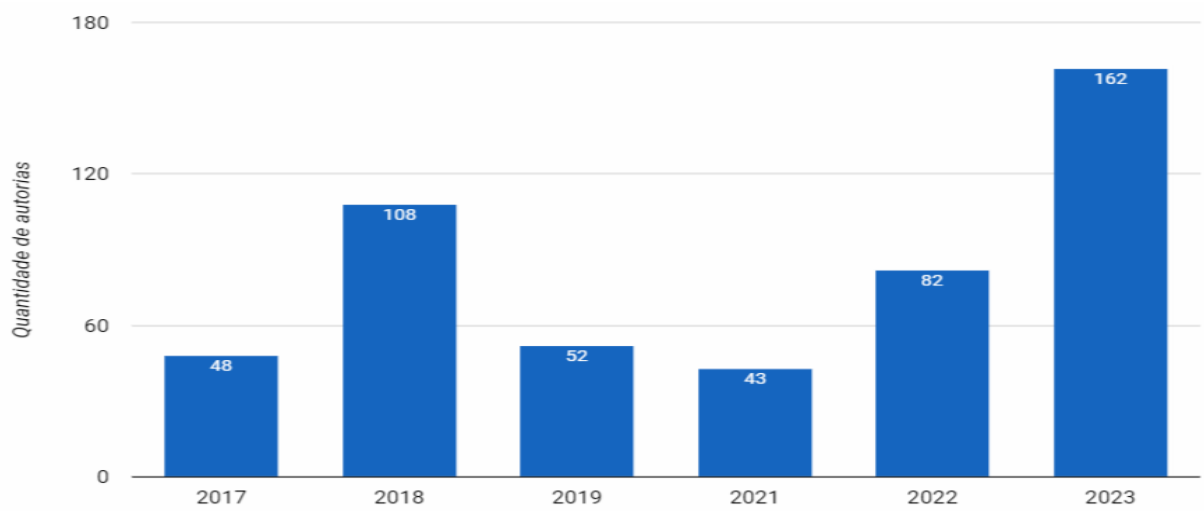
Em busca de um entendimento mais amplo e contextualizado dos *clusters* gerados entre autores e instituições, conforme o objetivo da pesquisa, foram criados alguns elementos visuais secundários. A **Figura 2** mostra a geolocalização das instituições dos autores, com destaque de tamanho proporcional para locais com maior número de autores. Uma forte concentração no Brasil e participações oriundas dos continentes: América do Norte, América Central, Europa e África.

Figura 2 – Mapa de geolocalização das instituições participantes nos seis eventos com destaque de tamanho proporcional à quantidade de autores

Fonte: Autoria própria, com apoio do software Looker Studio (2023).

A distribuição de frequência da participação de autores ao longo dos seis eventos pode ser conferida no **Gráfico 1**. O evento não ocorreu no ano de 2020 devido ao início da pandemia de COVID-19. Desde então, a partir do ano de 2021, houve um forte crescimento no número de participações.

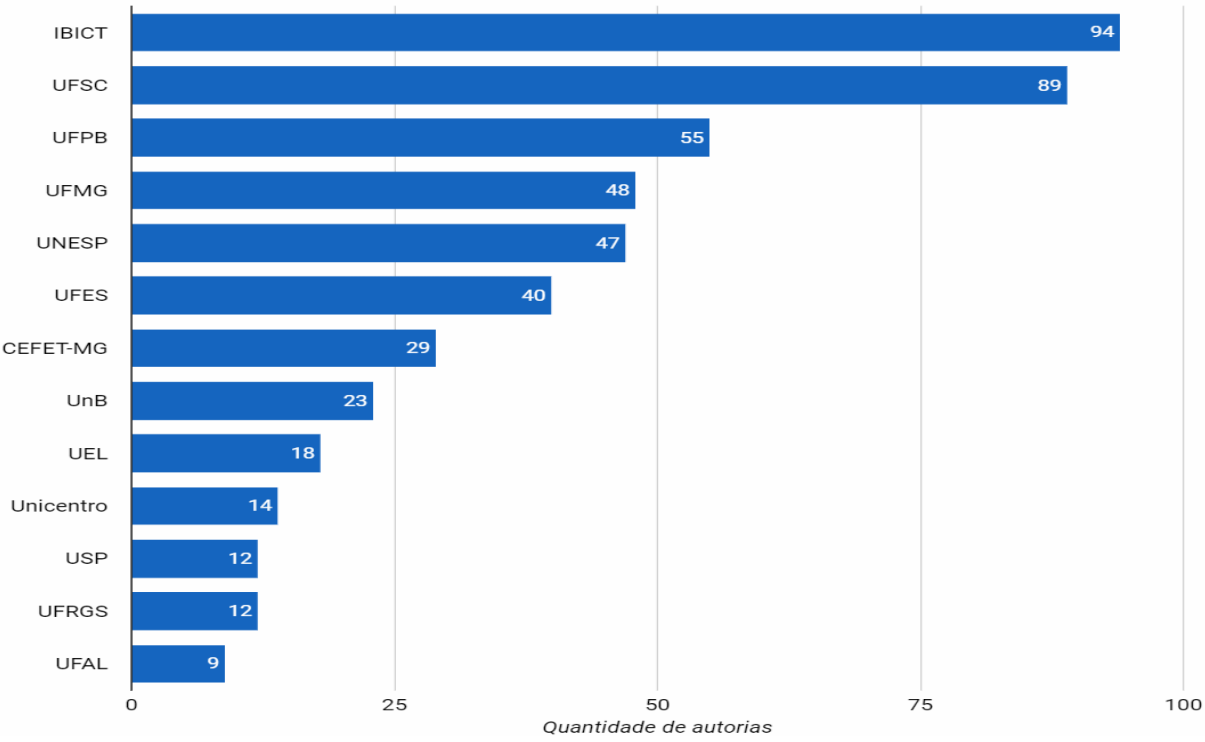
GRÁFICO 1 – Quantidade de autores participantes por evento



Fonte: Autoria própria, com apoio do software Looker Studio (2023).

O **GRÁFICO 2** apresenta as 14 instituições com maior número de participações em autorias nos anais, em um universo de 80 instituições. Uma autoria corresponde à participação de um autor vinculado à instituição em um artigo.

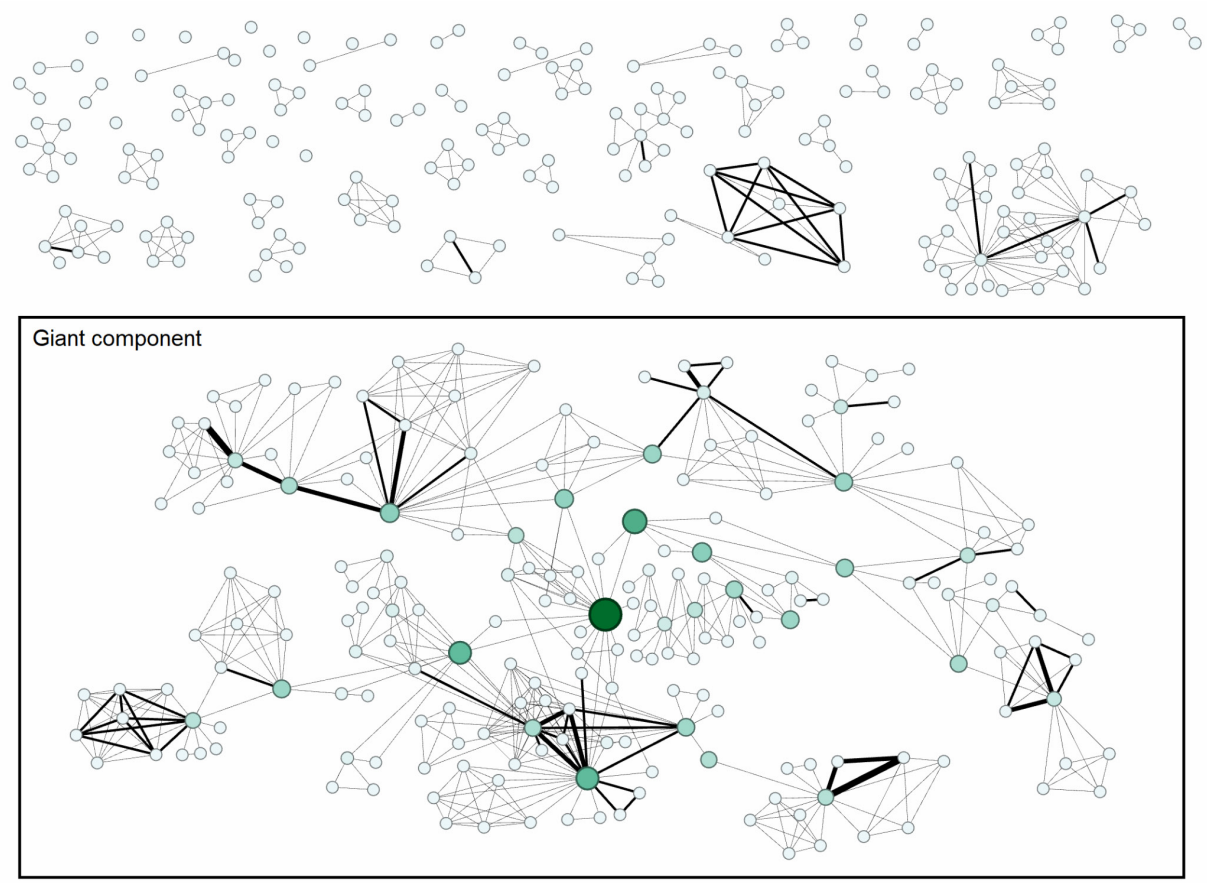
GRÁFICO 2 – Instituições com 9 ou mais autorias



Fonte: Autoria própria, com apoio do software Looker Studio (2023).

Para a etapa da ARS, foi formada inicialmente uma rede 4-partida com as categorias: artigos, autores, instituições e palavras-chave, denominada aqui de rede original. A rede monopartida de autores da **Figura 3** foi obtida com a aplicação de uma projeção bipartida sobre a rede original para conectar autores pela coautoria, considerando os seis eventos.

Figura 3 – Rede de autores conectados pela coautoria com destaque de cor e tamanho proporcional para o nível de intermediação dos autores



Fonte: Autoria própria, com apoio do software Gephi (2023).

A rede da **Figura 3** destaca proporcionalmente, em tamanho e cor, os nós dos autores pelo seu nível de intermediação (*betweenness*)[14]. Um *giant component*[15] é mostrado no retângulo destacado, juntando a maioria dos autores em um componente conectado, ou uma grande comunidade, com conexões de coautoria em suas pesquisas. Contudo, no contexto do *giant component*, o diâmetro[16] é igual a 13 e o caminho geodésico médio[17] é igual a 5,8. Dessa forma, pode-se inferir que, apesar de grande parte dos autores estarem no *giant component* com conexões de coautoria, ainda há uma distância considerável entre

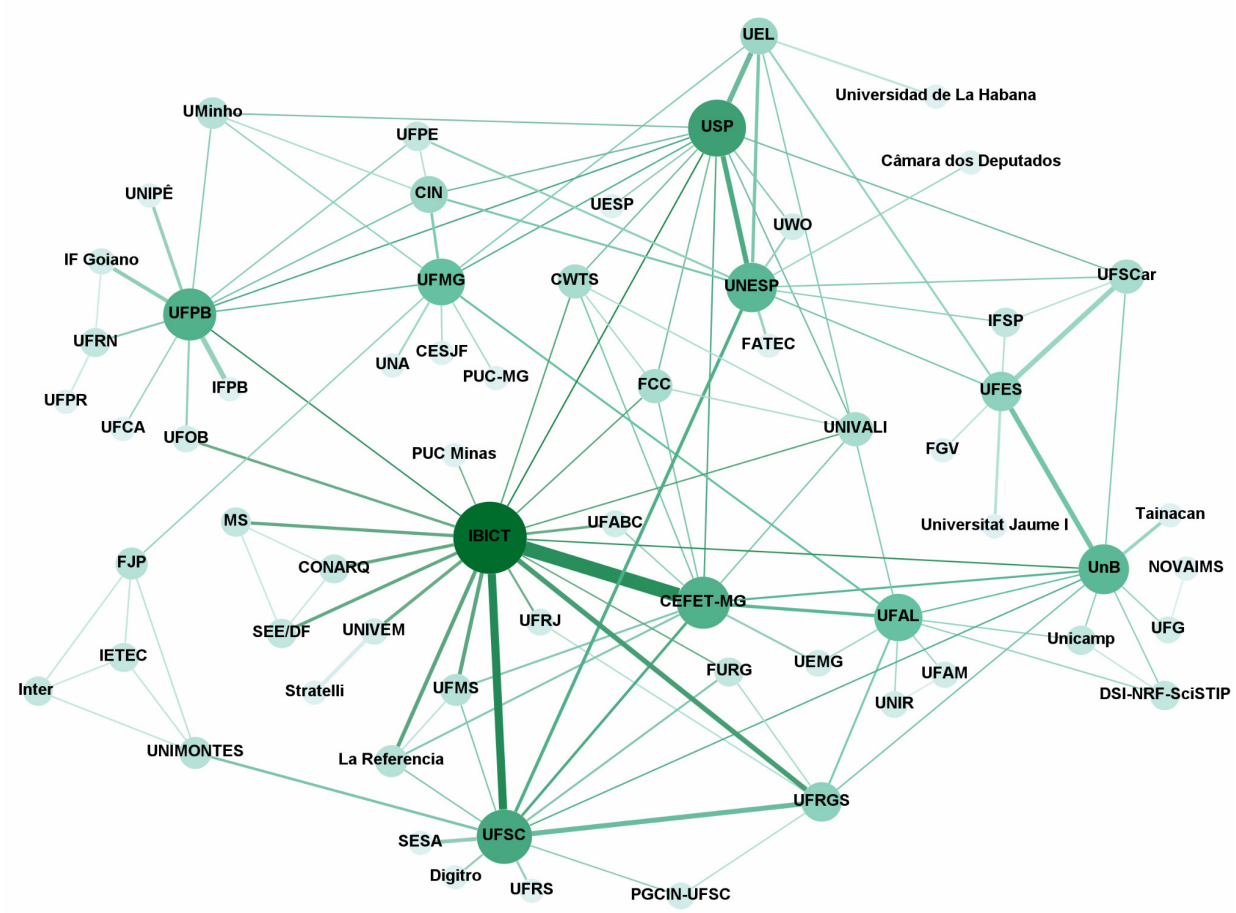
14 *Betweenness* mede a importância de um nó quanto à sua capacidade de intermediar o fluxo com os demais nós da rede.
15 *Giant component*, ou componente gigante, de uma rede, é um componente conectado com proporções muito maiores do que os demais. Componente conectado é um conjunto de nós onde cada um possui pelo menos um caminho para os demais nós da rede.
16 Diâmetro de uma rede é a maior distância existente entre dois nós de uma rede.
17 Caminho geodésico médio é o cálculo da média de todas as distâncias possíveis entre dois pares de nós quaisquer da rede.

eles, evidenciando a existência de subgrupos de coautores quase isolados e conectados por meio de um único autor com elevado grau de intermediação. O restante da rede, fora do *giant component*, mostra alguns autores isolados, díades, tríades e subgrupos, sendo alguns deles bem consolidados, com coautoria frequente representada por arestas mais espessas.

É importante destacar nessa rede de autores da **Figura 3** que, na visão de Granovetter (1973), os laços fortes são estabelecidos com as pessoas mais próximas, que no contexto dessa rede seriam as próprias coautorias das publicações. Por outro lado, Granovetter chama a atenção para a importância dos laços fracos que, no contexto dessa rede, seriam as conexões com autores de outros grupos de pesquisa por meio dos autores de nível mais elevado de intermediação. Nessa ótica, seriam abertas possibilidades de conexões com outras realidades e oportunidades de inovação e, consequentemente, de novas colaborações.

A **Figura 4** mostra uma rede de instituições pela coautoria de seus filiados com destaque proporcional para o grau da instituição, isto é, com maior número de conexões de coautoria. O caminho geodésico médio igual a 2,9 revela uma forte colaboração entre as instituições participantes do WIDaT.

Figura 4 – Rede de instituições conectadas pela coautoria de seus filiados com destaque proporcional de tamanho e cor para grau das instituições



Fonte: Autoria própria, com apoio do software Gephi (2023).

Dentre as instituições destacadas com maior número de autorias no **Gráfico 2**, observa-se que elas possuem um valor de centralidade de grau elevado, conforme mostra a rede da **Figura 4**, refletindo em bom nível de conexão com outras instituições. Além disso, destacam-se USP e UFAL pela razão maior que 1 entre o valor da centralidade de grau e a quantidade de autorias, isto é, a quantidade de conexões com outras instituições é maior do que a quantidade de autorias. Observa-se também que a Unicentro, apesar de estar entre as instituições com maior número de autorias, **Gráfico 2**, não possui conexões de autoria com outras instituições e, portanto, não apareceu na rede da **Figura 4**.

A investigação sobre as palavras-chave nos seis eventos resultou em 465 termos únicos que, cuja distribuição de frequência foi representada na nuvem de palavras da **Figura 5**. Grande parte dos termos citados na chamada dos trabalhos do evento, como tópicos de interesse[18], aparecem representados nessa nuvem de palavras, tendo o termo 'ciência da informação' com maior destaque. O **Quadro 6** apresenta as 15 palavras-chave mais frequentes.

Figura 5 – Nuvem de palavras-chave dos seis eventos com destaque de tamanho proporcional para a quantidade de ocorrências.

Fonte: Autoria própria, com apoio do software WordArt (2023).

18 Tópicos de interesse da chamada de trabalhos do último evento, em 2023, disponível em: https://widat2023.ibict.br/chamada-de-trabalhos/.

Quadro 6 – As 15 palavras-chave mais frequentes nas publicações dos seis eventos

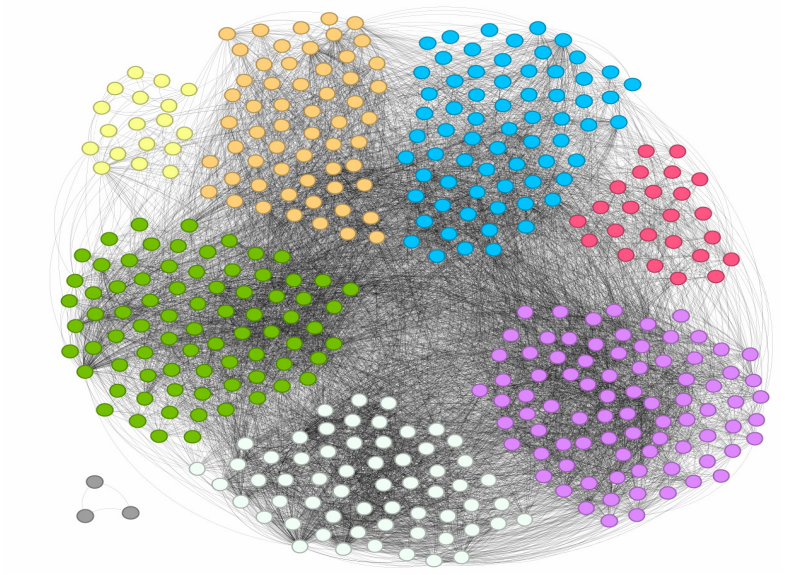
Palavra chave	Quantidade de ocorrências
ciência da informação	53
dados abertos	51
produção científica	49
ciência de dados	48
infometria	43
repositório de dados	43
gestão	35
metadados	35
educação	33
plataforma lattes	33
dados de pesquisa	32
big data	28
gestão de dados	27
saúde	27
ciência aberta	26

Fonte: Autoria própria (2023).

A rede monopartida com 395 autores, **Figura 6**, foi formada após uma projeção bipartida sobre a rede original utilizando-se as palavras-chave como nós intermediários de conexão. Ela apresenta oito *clusters* de autores organizados por semelhança do uso de palavras-chave em seus artigos.

Comparativamente, enquanto a rede de coautoria, **Figura 3**, possui diâmetro igual a 13 e caminho geodésico médio igual a 5,8, a rede de autores por afinidade de palavras-chave, **Figura 6**, possui diâmetro igual a 5 e caminho geodésico médio igual a 2,0. Isso demonstra que a rede social de autores por palavras-chave se aproxima do conceito de mundo pequeno (*small world*), conforme Watts e Strogatz (1998), onde um autor poderia facilmente acessar outro autor qualquer da rede por poucas conexões relativas a afinidade de suas pesquisas.

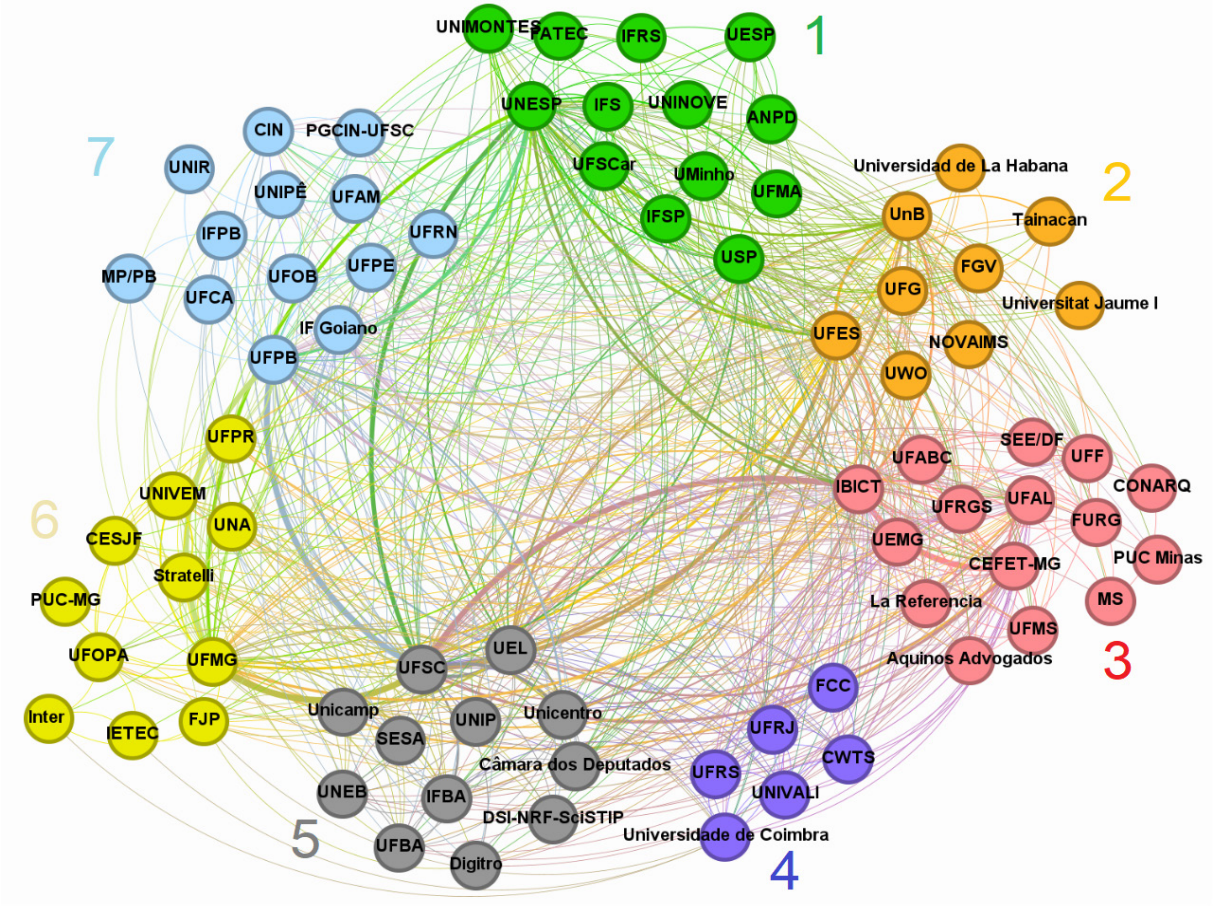
FIGURA 6 – Rede de autores conectados pelas palavras-chaves de seus artigos com destaques de cores para *clusters* formados por suas afinidades de pesquisa



Fonte: Autoria própria (2023), com apoio do software Gephi.

A rede monopartida de instituições, mostrada na **FIGURA 7**, foi obtida após projeção bipartida sobre a rede original, utilizando-se as palavras-chave como elementos de conexão. Ela mostra sete *clusters* numerados de instituições pela afinidade das palavras-chave dos artigos publicados por autores vinculados às suas respectivas instituições.

Figura 7 – Rede de instituições conectadas pelas palavras-chaves dos artigos de seus autores filiados com destaques de cores para *clusters* formados por semelhança de pesquisas



Fonte: Autoria própria (2023), com apoio do software Gephi.

Uma das instituições participantes do WIDaT, o Arquivo Nacional, não apareceu na rede da **Figura 6**, pois as palavras-chave usadas não tinham equivalência conceitual com as demais, das outras instituições.

Analisando comparativamente as redes de instituições por coautoria (**Figura 4**) e por palavras-chave (**Figura 7**), é possível inferir que essa última evidencia um potencial de proximidade entre autores vinculados às instituições pertencentes a um mesmo *cluster*. Isso pode ser corroborado pela leitura das nuvens de palavras para cada um dos sete *clusters*, Figuras 8 a 14.

A **Figura 8** apresenta a nuvem de palavras-chave do *cluster* 1 (verde) com destaque proporcional de tamanho dos principais termos. Destacam-se: **repositório de dados** (20 ocorrências referentes a 4,7% do total), **metadados** (18 ocorrências referentes a 4,2% do total) e **fusão de dados** (16 ocorrências referentes a 3,8% do total).

FIGURA 8 – Nuvem de palavras-chave do *cluster* 1, cor verde

Fonte: Autoria própria, com apoio do software *WordArt* (2023).

A **FIGURA 9** apresenta a nuvem de palavras-chave do *cluster* 2 (laranja) com destaque proporcional de tamanho dos principais termos. Destacam-se: **ciência de dados** (18 ocorrências referentes a 4,2% do total), **cultura** e **museu** (16 ocorrências cada, referentes a 3,7% do total).

FIGURA 9 – Nuvem de palavras-chave do *cluster* 2, cor laranja

Fonte: Autoria própria, com apoio do software *WordArt* (2023).

A **Figura 10** apresenta a nuvem de palavras-chave do *cluster* 3 (vermelho) com destaque proporcional de tamanho dos principais termos. Destacam-se: **produção científica** (33 ocorrências referentes a 4,8% do total), **plataforma Lattes** (30 ocorrências referentes a 4,4% do total) e **brcris** (23 ocorrências referentes a 3,4% do total).

Figura 10 – Nuvem de palavras-chave do *cluster* 3, cor vermelha

Fonte: Autoria própria, com apoio do software *WordArt* (2023).

A **Figura 11** apresenta a nuvem de palavras-chave do *cluster* 4 (roxo) com destaque proporcional de tamanho dos principais termos. Destacam-se: **infometria** (8 ocorrências referentes a 18,2% do total), **métricas**, **educação** e **portal** (3 ocorrências cada, referentes a 6,8% do total).

Figura 11 – Nuvem de palavras-chave do *cluster* 4, cor roxa

Fonte: Autoria própria, com apoio do software *WordArt* (2023).

A **Figura 12** apresenta a nuvem de palavras-chave do *cluster* 5 (cinza) com destaque proporcional de tamanho dos principais termos. Destacam-se: **ciência de dados** (28 ocorrências referentes a 4,5% do total), **dados abertos** (24 ocorrências referentes a de 3,8% do total) e **educação** (17 ocorrências referentes a 2,7% do total).

Figura 12 – Nuvem de palavras-chave do *cluster* 5, cor cinza

Fonte: Autoria própria, com apoio do *software WordArt* (2023).

A **Figura 13** apresenta a nuvem de palavras-chave do *cluster* 6 (amarelo) com destaque proporcional de tamanho dos principais termos. Destacam-se: **ontologia** (20 ocorrências referentes a 5,3% do total), **gestão** (18 ocorrências referentes a 4,8% do total), **dados abertos** (12 ocorrências referentes a 3,2% do total).

Figura 13 – Nuvem de palavras-chave do *cluster* 6, cor amarela

Fonte: autoria própria, com apoio do software *WordArt* (2023).

A **Figura 14** apresenta a nuvem de palavras-chave do *cluster* 7 (azul) com destaque proporcional de tamanho dos principais termos. Destacam-se: **ciência da informação** (15 ocorrências referentes a 4,9% do total), **gestão de dados** (10 ocorrências referentes a 3,2% do total), **lei geral de proteção de dados** e **privacidade de dados** (9 ocorrências cada, referentes a 2,9% do total).

Figura 14 – Nuvem de palavras-chave do *cluster* 7, cor azul

Fonte: Autoria própria, com apoio do software *WordArt* (2023).

As nuvens de palavras dos sete *clusters* sugerem coesão semântica dos temas tratados pelas instituições pertencentes a cada um dos *clusters*. A observação das palavras-chave mais frequentes em um determinado *cluster* não limita a pesquisa das instituições pertencentes ao *cluster*, pelo contrário, pode motivar aproximações inter-instituições para colaboração e fortalecimento de pesquisa, tanto para aquelas já consolidadas quanto para as que se encontram em fase inicial de investigação e estudo.

CONCLUSÕES

A pesquisa revelou grupos de afinidade entre autores e entre instituições participantes do WIDaT por meio de ARS e visualização de informação com gráficos, mapa de geolocalização e redes sociais, conforme previsto no objetivo proposto.

A metodologia empregada requereu um trabalho minucioso com as palavras-chave dos artigos publicados nos anais, sendo necessária a realização de análise de conteúdo de Bardin (1977) para execução de agrupamentos e a determinação de termos equivalentes. O trabalho conjunto com os softwares OpenRefine e Gephi, no mapeamento para a criação

das redes, foi determinante para a execução de projeções bipartidas que possibilitaram a criação de redes mono partidas de autores e instituições, tanto por coautoria quanto por palavras-chave.

Observaram-se contrastes importantes entre as redes de autores e instituições oriundas da coautoria e as redes oriundas das palavras-chave. Apesar de muitos grupos de coautoria estarem bem consolidados, houve maior aproximação de autores e instituições por meio das palavras-chave, evidenciando o fenômeno de “mundo pequeno”, e mostrando o grande potencial de colaborações em trabalhos futuros e o fortalecimento em pesquisas em andamento.

As nuvens de palavras dos sete clusters sugerem coesão semântica dos temas tratados pelas instituições pertencentes a cada um dos clusters. A observação das palavras-chave mais frequentes em um determinado cluster pode trazer motivação para aproximações inter-instituições na colaboração e fortalecimento de pesquisa, tanto para aquelas já consolidadas quanto para as que se encontram em fase inicial de investigação e estudo.

Os resultados da pesquisa e, principalmente, a nuvem de palavras-chave geral das publicações dos seis eventos mostrou que o WIDaT vem se consolidando como um espaço aberto, forte e capaz de reunir as comunidades que trabalham com temáticas diversas que envolvem dados e tecnologia, e com intersecções na Ciência da Informação.

Considerando a possibilidade de aplicação da metodologia desta pesquisa em outros eventos, indica-se, como continuidade do trabalho, a criação de rotinas automatizadas de extração de metadados de anais para diminuir o trabalho manual de coleta de dados.

REFERÊNCIAS

BARDIN, L. **Análise de conteúdo**. Lisboa: Edições 70, 1977.

CHEN, C. **Mapping scientific frontiers**: the quest for knowledge visualization. 2. ed. London: Springer Science & Business Media, 2013.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI Magazine**, Palo Alto, v. 17, n. 3, p. 37-54, Mar. 1996. DOI: 10.1609/aimag.v17i3.1230.

GAO, M.; CHEN, L.; LI, B.; LI, Y.; LIU, W.; XU, Y.-c. Projection-based link prediction in a bipartite network. **Information Sciences**, New York, v. 376, p. 158-171, Jan. 2017. Disponível em: https://doi.org/10.1016/j.ins.2016.10.015. Acesso em: 29 maio 2023.

HAND, D. J.; MANNILA, H.; SMYTH, P. **Principles of data mining**. Cambridge: MIT Press, 2001.

HIGGINS, S. S.; RIBEIRO, A. C. A. **Análise de redes em Ciências Sociais**. Brasília: Enap, 2018. Disponível em: https://repositorio.enap.gov.br/bitstream/1/3337/1/Livro_Analise%20de%20Redes%20em%20Ci%C3%AAncias%20Sociais.pdf. Acesso em: 29 maio 2023.

MATHEUS, R. F.; SILVA, A. B. O. Análise de redes sociais como método para a Ciência da Informação. **DataGramaZero** – Revista de Ciência da Informação, Belo Horizonte, v. 7, n. 2, abr. 2006.

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

NOOY, W.; MRVAR, A.; BATAGELJ, V. **Exploratory social network analysis with Pajek**: revised and expanded edition for updated software. 3. ed. Cambridge: Cambridge University Press, 2018.

SOUSA, J. R.; SANTOS, S. C. M. Análise de conteúdo em pesquisa qualitativa: modo de pensar e de fazer. **Pesquisa e Debate em Educação**, Juiz de Fora, v. 10, n. 2, p. 1396-1416, jul./dez. 2020. Disponível em: https://periodicos.ufjf.br/index.php/RPDE/article/view/31559. Acesso em: 29 maio 2023.

GRANOVETTER, M. S. The strength of weak ties. **American Journal of Sociology**, Chicago, v. 78, n. 6, p. 1360–1380, May. 1973. DOI: 10.1086/225469. Disponível em: http://www.journals.uchicago.edu/doi/10.1086/225469. Acesso em: 29 maio 2023.

WASSERMAN, S.; FAUST, K. **Social network analysis**: methods and applications. Cambridge: Cambridge University Press, 1994. 868 p.

WATTS, D. J.; STROGATZ, S. H. Collective dynamics of ‘small-world’ networks. **Nature**, Nova York, v. 393, n. 6684, p. 440–442, June. 1998. DOI: 10.1038/30918. Disponível em: http://www.nature.com/nature/journal/v393/n6684/abs/393440a0.html. Acesso em: 29 maio 2023.

WIDAT. **Home page**. [*S. l.*], 2023. Disponível em: https://widat2023.ibict.br/. Acesso em: 29 maio 2023.

AGRADECIMENTOS

Fundação de Amparo à Pesquisa e Inovação do Espírito Santo (FAPES).