



Da qualidade dos dados à formação de comunidades temáticas: análise de redes de informação em acervos culturais

Henrique Monteiro Cristovão

Doutor em Ciência da Informação, Universidade de Brasília (UnB), Brasília, DF, Brasil.
Professor Associado no Departamento de Arquivologia, Universidade Federal do Espírito Santo (UFES), Vitória, ES, Brasil.

<http://lattes.cnpq.br/5035919384923489>

henrique.cristovao@ufes.br

<https://orcid.org/0000-0003-2011-7022>

Daniela Lucas da Silva Lemos

Doutora em Ciência da Informação, Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, MG, Brasil.

Professora Associada no Departamento de Biblioteconomia, Universidade Federal do Espírito Santo (UFES), Vitória, ES, Brasil.

<http://lattes.cnpq.br/9280443047358807>

daniela.l.silva@ufes.br

<https://orcid.org/0000-0003-1565-7366>

Submetido em: 13/02/2026. Aprovado em: 13/04/2026. Publicado em: 29/05/2026.

RESUMO

Padrões de documentação são usados amplamente nas ações de organização e tratamento da informação para constituir bases de dados com qualidade e, assim, favorecer a produção de informações estratégicas para a gestão de acervos culturais. Este artigo relata a experiência de construção e qualificação das bases de dados do acervo do Fundo de Cultura do Espírito Santo, sob gestão da Secretaria de Estado da Cultura, disponibilizadas na rede de provedores da plataforma MEDIATECA Capixaba. A pesquisa enquadra o percurso metodológico como uma descoberta de conhecimento orientada a redes, articulando a gestão do dado ao longo de seu ciclo de vida (coleta, armazenamento, recuperação e descarte) com etapas analíticas de seleção, preparação, transformação, mineração e interpretação. Metodologicamente, a modelagem de metadados (Dublin Core), o uso de vocabulários controlados e a Matriz de Restrições subsidiaram perfis de aplicação, enquanto rotinas de limpeza, normalização e mapeamento relacional viabilizaram a geração de grafos para exploração no Gephi, complementada por visualizações estatísticas em planilha e Looker Studio. Como resultados, redes de afiliação entre autores contemplados e eixos temáticos evidenciam padrões de participação e permitem identificar agrupamentos coesos por similaridade temática, apoiando hipóteses interpretáveis e insumos para curadorias digitais e decisões de política pública cultural. Conclui-se que a efetividade da análise de redes depende criticamente de escolhas de qualidade e padronização de dados e que a inspeção visual, tratada como parte da interpretação/avaliação, amplia a comunicabilidade dos achados para públicos não especializados.

Palavras-Chave: padrões de documentação; acervos culturais; qualidade e ciclo de vida de dados; análise de redes de informação; descoberta de conhecimento.

INTRODUÇÃO

A padronização é um elemento crucial, discutido amplamente na Ciência da Informação, especialmente na área de Organização e Representação da Informação e do Conhecimento. Padrões de documentação são usados nessa perspectiva para apoiar práticas de catalogação e são essenciais na crescente digitalização de coleções de objetos culturais (Gilliland, 2016). Na plataforma digital de agregação de dados culturais, denominada Midiateca Capixaba (Lemos; Martins; Macedo, 2024), os padrões de documentação são fundamentados em teorias de catalogação descritiva e de assunto, focando na criação e no uso de metadados, vocabulários controlados e regras de catalogação pertinentes a cada perfil de aplicação.

O presente estudo relata a experiência com as bases de dados do acervo do Fundo de Cultura do Espírito Santo (Funcultura), sob a gestão da Secretaria de Estado da Cultura (SECULT/ES), implementadas em um dos repositórios institucionais da rede de provedores da Midiateca Capixaba, para a qual se buscou qualificar seus dados a partir de padrões de documentação na intenção de se obter informações e conhecimentos úteis à gestão de seu acervo cultural.

O Funcultura foi criado pela Lei Complementar Estadual nº458 de 21/10/2008 e, desde então, é o principal mecanismo do Governo do Estado para o fomento e incentivo a programas, projetos e ações culturais (Espírito Santo, 2023). Os recursos do Fundo são aplicados em projetos, ações e programas baseados em três dimensões culturais: cultura como expressão simbólica de um povo, cultura como direito e cidadania, e cultura como economia que gera renda e trabalho. Nesse ensejo, há uma necessidade real de otimização do acesso à cultura frente ao mundo contemporâneo, com sua dinâmica digital que rege o cotidiano da sociedade e de grande parte dos cidadãos. Assim, a entrada deste acervo no projeto Midiateca Capixaba é justamente para oportunizar estudo, análise, pesquisa e organização do conjunto de dados acerca dos editais, projetos e produtos culturais em bases de dados qualificadas, possibilitando a sua exploração para fins de difundir informações úteis à sociedade bem como conhecimentos que permitam aos gestores tomarem decisões pertinentes para a elaboração de políticas públicas que sejam de interesse ao cidadão capixaba, em especial.

A análise de redes de informação tem sido usada para a projeção visual de descoberta de conhecimento em bases de dados e, principalmente, por ser uma ferramenta metodológica que viabiliza enxergar o que outras abordagens não permitem (Wasserman; Faust, 1994; Higgins, Ribeiro; 2018), sendo também uma boa estratégia para investigar conexões entre elementos sociais e culturais. Por exemplo, um estudo para revelar grupos de afinidade entre autores e instituições participantes de evento científico, evidenciando como redes de coautoria e de palavras-chave podem apoiar leituras de agrupamentos e padrões por inspeção visual, foi empregada a combinação de análise de redes de informação e visualização de informação (Cristovão; Vale, 2024). Contudo, ela precisa estar suportada por redes com

dados bem estruturados, com conteúdos bem formatados e com integridade, além de contar com o apoio de informações estatísticas para auxiliar na interpretação, o que depende de dados bem descritos.

Assim, o problema de pesquisa encontra-se em saber de que maneira o uso de padrões de documentação consegue favorecer o tratamento de dados endereçados à análise exploratória de dados, de modo a entender suas características, identificar padrões, detectar *outliers* e formular hipóteses preliminares. Por conseguinte, o objetivo geral da pesquisa é identificar agrupamentos coesos nos autores contemplados nos editais por meio de similaridades nos eixos temáticos de seus respectivos projetos, a partir da análise exploratória das bases de dados do Funcultura.

Para situar essa estratégia no campo e explicitar a passagem da qualificação de dados para a extração de padrões, a seção seguinte apresenta um enquadramento de descoberta de conhecimento orientada a redes, compatibilizando o processo de descoberta de conhecimento em bases de dados (Fayyad; Piatetsky-Shapiro; Smyth, 1996; Hand; Mannila; Smyth, 2001) com o ciclo de vida dos dados (Sant'Ana, 2016) e com dimensões de qualidade orientadas ao uso (Wang; Strong, 1996), particularmente relevantes quando a ausência de dados e inconsistências podem distorcer medidas e comunidades em redes (McKnight *et al.*, 2007).

DESCOBERTA DE CONHECIMENTO ORIENTADA A REDES: da qualidade e do ciclo de vida dos dados à formação de comunidades em acervos culturais

A descoberta de conhecimento em bases de dados (KDD: Knowledge Discovery in Databases) pode ser entendida como um processo, e não apenas um conjunto de técnicas, voltado a transformar dados em representações úteis para apoiar interpretações e decisões. Fayyad, Piatetsky-Shapiro e Smyth (1996) conceituam a KDD como um processo não trivial de identificação de padrões válidos e potencialmente úteis, destacando também a ênfase em padrões compreensíveis e interpretáveis como conhecimento. O processo sugerido pelos autores é composto de cinco etapas: (i) *Selection*: seleção da base de dados e identificação de possíveis problemas e/ou objetivo da análise; (ii) *Preprocessing*: pré processamento dos dados, preparação limpeza e ajustes diversos; (iii) *Transformation*: transformação dos dados para uma base que seja aceita pelos algoritmos de data mining; (iv) *Data mining*: mineração de dados pela aplicação dos algoritmos escolhidos conforme o contexto, problema e dados para extrair padrões e relacionamentos; e (v) *Interpretation/evaluation*: interpretação e avaliação dos resultados obtidos pela mineração de dados. Contudo os autores alertam que a aplicação dos métodos de KDD sem análise ou julgamento crítico pode produzir padrões espúrios e interpretações inválidas. Hand, Mannila e Smyth (2001) enfatizam a etapa *data mining* como a análise de conjuntos observacionais para revelar relações não esperadas e sumarizar dados de maneiras novas, compreensíveis e úteis, observando que os resultados podem assumir a forma de modelos ou padrões diversos, tais como regras, clusters, grafos

etc. Essa caracterização é especialmente pertinente quando a organização de dados culturais visa apoiar tanto a descrição quanto a extração de regularidades (padrões de participação, distribuição de recursos, recorrências temáticas e possíveis assimetrias).

Paralelamente às etapas do KDD, o ciclo de vida dos dados, conforme proposto por Sant'Ana (2016), é organizado em quatro etapas: (i) Coleta: definição das necessidades e obtenção dos dados; (ii) Armazenamento: registro e estruturação dos dados para persistência e reúso; (iii) Recuperação: disponibilização dos dados para acesso, análise e visualização; e (iv) Descarte: eliminação de dados desnecessários ou inviáveis de manter ou processar.

Na perspectiva da Ciência da Informação, é possível compatibilizar o processo de KDD de Fayyad, Piatetsky-Shapiro e Smyth (1996) com o ciclo de vida dos dados de Sant'Ana (2016) ao tratá-los como dois recortes complementares do mesmo fenômeno: enquanto o KDD explicita o encadeamento analítico que transforma dados em padrões interpretáveis, o modelo de Sant'Ana organiza a gestão informacional do dado ao longo de quatro fases orientadas pelo uso e permeadas por fatores transversais como a qualidade, a integração, a privacidade, a disseminação e a preservação. Assim, a coleta pode ser alinhada às etapas iniciais do KDD em que se define o domínio/problema e se cria o conjunto-alvo, isto é, selecionam-se dados pertinentes ao objetivo informacional; o armazenamento se aproxima do pré-processamento e da transformação (limpeza, tratamento de ausências/ruído, redução e projeções) que preparam um acervo de dados consistente para análise e reúso; a recuperação encontra paralelo com a execução do data mining e com a exploração iterativa que depende de acesso/consulta aos dados para enumerar padrões e modelos; e o descarte, por sua vez, pode ser compatibilizado com a fase de interpretação/avaliação e ação do KDD, pois envolve decisões informacionais de retenção/eliminação (e, por contraste, de preservação e disseminação) tanto sobre dados quanto sobre resultados, distinguindo o que tem valor para o uso institucional do que deve ser depurado, arquivado ou encerrado no ciclo.

Nesse enquadramento, a análise de redes de informação se enquadra como o artefato analítico que materializa a transição dos dados aos padrões nas duas abordagens, uma vez que permite tratar relações como estruturas analisáveis, explorando padrões de conexão que não emergem com a mesma clareza em tabelas ou listagens. No KDD, a passagem de tabelas e registros para um grafo corresponde principalmente às etapas de *Transformation* (quando entidades e relações são expressas como nós e arestas, frequentemente em estruturas bipartidas de afiliação, e de *Data mining* quando se aplicam medidas e algoritmos para extrair regularidades como centralidades, coesões e comunidades (Newman, 2010; Borgatti; Halgin, 2014). Em redes bipartidas, a projeção para uma rede monopartida pode ser entendida como uma transformação que viabiliza leituras de proximidade por coocorrência, tornando possível aplicar rotinas de detecção de comunidades e análises estruturais no grafo resultante (Newman, 2010; Banerjee; Jenamani; Pratihari, 2017). Já no ciclo de vida dos dados, a análise de redes se ancora na Coleta/Armazenamento, onde a consistência de identificadores, a padronização e a qualidade determinam a fidelidade das ligações representadas, e se concretiza na Recuperação, quando os dados são disponibilizados não

apenas para consulta, mas também para exploração relacional e visual, de modo a apoiar interpretação. A inspeção visual e a navegação exploratória funcionam como suporte cognitivo para compreender padrões e validar hipóteses sobre agrupamentos e conexões relevantes (Nooy; Mrvar; Batagelj, 2018), reforçando o papel da visualização como meio de ampliar compreensão e comunicação de conhecimento entre públicos com diferentes repertórios técnicos (Eppler; Burkhard, 2004; Burkhard, 2005).

O reconhecimento dos resultados de uma etapa analítica dependem diretamente de decisões anteriores: modelagem, padronização semântica, normalização, controle de vocabulário e documentação de transformações. Em contextos institucionais, essa perspectiva ajuda a justificar por que a pesquisa não se reduz à visualização da rede, mas envolve uma engenharia de dados e metadados que antecede e condiciona a leitura dos padrões.

Essa dependência é ainda mais crítica quando se considera qualidade de dados. Wang e Strong (1996) propõem uma visão orientada ao usuário baseada na noção de *fitness for use*, isto é, qualidade como adequação ao uso pretendido; a partir desse ponto de vista, organizam dimensões em quatro categorias: intrínseca (qualidade em si), contextual (adequação à tarefa), representacional (clareza e forma de representação) e acessibilidade (condições de acesso). Complementarmente, Eckerson chama atenção para o impacto econômico e organizacional da baixa qualidade, observando que erros, inconsistências e degradação ao longo do tempo podem comprometer aplicações analíticas e decisórias, e cita estimativas amplamente utilizadas na literatura sobre custos agregados da má qualidade (Eckerson, 2002). No plano metodológico, isso sustenta a necessidade de explicitar rotinas de validação (completude, consistência, duplicidades, padronização) e escolhas de normalização que afetam diretamente medidas de rede e a estabilidade dos agrupamentos detectados.

Um ponto específico na qualidade de dados e com grande efeito em redes é a ausência de dados que, segundo McKnight *et al.* (2007) impedem a interpretação correta do fenômeno que está sendo estudado. Em análise de redes, esse ponto se traduz em consequências diretas: faltas de preenchimento e inconsistências de identificação podem quebrar conexões, reduzir graus, fragmentar componentes e alterar a composição de comunidades — o que reforça a importância de discutir integridade e rastreabilidade antes da interpretação estrutural.

Newman (2010) observa que a decomposição de redes em comunidades, formadas por subgrupos coesos, é uma noção de importância prática: redes sociais tendem a se organizar em grupos densamente conectados internamente, com conexões mais fracas entre grupos; examinar comunidades pode revelar divisões de interesse e níveis de organização difíceis de perceber sem dados e métodos de rede. Assim, quando o objetivo é identificar agrupamentos coesos por similaridade temática, ou por coocorrência de vínculos com editais/temas, a formação de comunidades opera como uma forma de sumarização estrutural: reduz complexidade e oferece uma hipótese interpretável sobre quem se relaciona com quem e por qual critério relacional.

Quando os dados assumem a forma de afiliações, por exemplo, autores vinculados a temas/editais, o modelo adequado frequentemente é uma rede bipartida, com dois tipos de vértices e ligações apenas entre tipos diferentes. Borgatti e Halgin (2014) detalham que redes de afiliação são bipartidas porque seus nós podem ser particionados em duas classes, de modo que as ligações só ocorrem entre classes, que pode ser interpretada como proximidade por co-afiliação, mas que também pode confundir proximidade social e similaridade, exigindo cautela interpretativa (Borgatti; Halgin, 2014). Em termos operacionais, Banerjee, Jenamani e Pratihari (2017) formalizam a projeção de um modo em grafos bipartidos conectando dois nós do mesmo conjunto quando compartilham vizinhos no outro conjunto; a projeção pode ser ponderada quando se considera a força baseada em vizinhos compartilhados, e suas propriedades topológicas tornam-se relevantes para análises posteriores. Os autores sustentam que, ao usar projeções bipartidas, é recomendável explicitar escolhas de ponderação, limiares e normalizações, pois elas alteram densidade, distribuição de graus e, conseqüentemente, o resultado de algoritmos de comunidade.

Por fim, a etapa de inspeção visual não deve ser apresentada como um adorno, mas como parte do processo de interpretação e validação exploratória. Eppler e Burkhard (2004) definem visualização de conhecimento como o uso de representações visuais para melhorar a criação e a transferência de conhecimento entre pessoas, indo além do transporte de fatos ao apoiar a comunicação de *insights*, perspectivas e experiências de modo que o receptor consiga reconstruir o entendimento. Burkhard (2005) reforça a distinção entre visualização de informação e visualização de conhecimento, argumentando que representações complementares podem favorecer transferência e criação de conhecimento entre indivíduos com diferentes formações. Em redes, isso se conecta diretamente à prática descrita por de Nooy, Mrvar e Batagelj (2018) onde na análise exploratória, a exploração envolve visualização e manipulação de redes concretas e tem como ganho a compreensão intuitiva de conceitos e padrões estruturais, diferindo da lógica estritamente inferencial. Logo, ao apresentar resultados, a visualização pode ser enquadrada como um instrumento para formular e refinar hipóteses sobre comunidades, pontes, *outliers* e componentes, desde que acompanhada por evidências quantitativas, tais como métricas, pesos, coerência temática, e por um relato transparente das escolhas de modelagem.

No caso de acervos culturais, esse enquadramento é particularmente relevante porque a documentação e a curadoria do dado não visam apenas descrever itens, mas também sustentar interpretações sobre práticas, temas, agentes e dinâmicas de circulação cultural. Nessa direção, a descoberta de conhecimento orientada a redes permite tratar registros e metadados como evidências relacionais, por exemplo, entre agentes, objetos, eventos e categorias, articulando exigências de qualidade e de gestão ao longo do ciclo de vida, que determinam consistência, comparabilidade e recuperabilidade, com a extração de padrões estruturais, como agrupamentos e comunidades, que sintetizam regularidades do acervo e tornam os resultados comunicáveis para públicos não especializados (Fayyad; Piatetsky-

Shapiro; Smyth, 1996; Sant'Ana, 2016; Wang; Strong, 1996; Newman, 2010). Assim, a análise relacional funciona como ponte entre a organização/representação da informação e a produção de conhecimento aplicado à gestão e ao uso social do patrimônio cultural.

METODOLOGIA

A pesquisa é de natureza aplicada, de abordagem quali-quantitativa, e de caráter exploratório e descritivo, envolvendo bases de dados do acervo cultural do Funcultura, disponibilizado em repositório digital institucional. Essas bases são exploradas e seus dados tratados para a geração de gráficos estatísticos e redes de informação tripartida, bipartida e monopartida. Os procedimentos metodológicos foram conduzidos em duas etapas, a saber: (i) modelagem e criação das bases de dados; e (ii) análise e exploração dos dados.

A primeira etapa metodológica pode ser considerada a espinha dorsal para a obtenção de dados curados para transformação em fontes de informações qualificadas e destinadas à plataforma de agregação MEDIATECA Capixaba, bem como matéria-prima para gestores da área da cultura em produzir informações estratégicas à tomada de decisões ao Fundo.

Assim sendo, a etapa de Modelagem do Esquema de Metadados foi assistida por meio da técnica de criação de Matriz de Restrições (Baker; Nilsson; Johnston, 2008), em que se integra e prevê um conjunto de registros de metadados que são instâncias válidas para um dado perfil de aplicação, neste caso, o acervo cultural do Funcultura.

A partir do ambiente de desenvolvimento da tecnologia Tainacan¹, pôde-se criar o esquema de dados à luz da Matriz de Restrições. Nessa perspectiva, a produção das bases de dados do Funcultura contou com o uso de padrões de documentação (Lemos; Martins; Macedo, 2024) na organização e representação dos documentos do Fundo, destacando o uso de vocabulários controlados (taxonomias e arquivos de autoridade) e do padrão de metadados Dublin Core como um modelo de referência, visando obter informações normalizadas e consistentes para a extração de insights valiosos à instituição.

Na segunda etapa metodológica, voltada à análise exploratória, os dados oriundos das coleções do acervo do Funcultura, usados na pesquisa, foram organizados em três tabelas relacionadas. A primeira tabela de dados representa as informações dos 876 autores proponentes dos projetos que foram contemplados. A segunda tabela de dados contém as informações dos 1370 projetos inscritos que foram contemplados bem como o edital ao qual participaram, sendo conectada à primeira tabela por meio de um identificador único do autor proponente. A terceira tabela de dados possui informações sobre 262 produtos vinculados aos projetos contemplados, sendo conectada à segunda tabela por meio de um identificador único do projeto inscrito. Todos os dados da pesquisa, inclusive os arquivos obtidos nos processamentos dos resultados estão disponíveis em repositório para acesso público².

¹ Tainacan é o *software* usado na implementação das bases de dados na rede de provedores da MEDIATECA Capixaba. Disponível em: <https://tainacan.org/>

² Repositório de dados da pesquisa disponível em: <https://drive.google.com/drive/folders/1AQUF007OGKcxl1097FmYHGJy0v3Lyeku?usp=sharing>.

Com apoio do software OpenRefine³, os dados das três tabelas passaram por um processo de preparação e de transformação com foco na qualidade e padronização. Essa etapa envolveu ações, tais como a limpeza de dados com símbolos irrelevantes, eliminação de registros em branco, padronização de nomes de variáveis, e a criação de variáveis a partir da extração de dados existentes para representar, com menor granularidade, informações sobre o ano do edital, a categoria cultural e o eixo temático ao qual os projetos se enquadram. Ao todo foram organizados, por meio de um vocabulário controlado do tipo taxonomia, oito categorias culturais e 49 eixos temáticos dentre os 1370 projetos contemplados. A distribuição da quantidade de projetos nessa taxonomia pode ser vista no **GRÁFICO 1**, na seção de resultados.

Do ponto de vista de qualidade de dados, essas rotinas não foram meramente operacionais, pois elas definem condições de adequação ao uso analítico (*fitness for use*), na medida em que completude, consistência e padronização impactam a fidedignidade das ligações que foram representadas no grafo e, conseqüentemente, a estabilidade de métricas e agrupamentos (Wang; Strong, 1996).

Com foco na identificação de agrupamentos coesos entre os autores contemplados, por meio de similaridades nas categorias culturais e eixos temáticos de seus respectivos projetos, criou-se um mapeamento dos dados das tabelas para o formato de uma rede de informação tripartida contendo os seguintes nós: autores contemplados, editais de inscrição, e eixos temáticos. O referido mapeamento, **QUADRO 1**, foi processado no software OpenRefine e gerou um arquivo no formato GML⁴ para ser interpretado pelo software Gephi⁵.

No próprio software Gephi, a partir da rede tripartida gerada pelo mapeamento do **QUADRO 1**, foram aplicadas projeções bipartidas para a obtenção de outras redes simplificadas, especialmente uma rede monopartida de autores contemplados.

3 OpenRefine é considerado um *Interactive Data Transformation*, sendo uma ferramenta de código aberto utilizada para a limpeza e transformação de volumes massivos de dados estruturados. Disponível em <https://openrefine.org/>.

4 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/.

5 Gephi é um *software* de código aberto utilizado para visualização, análise e manipulação de redes e grafos. Disponível em <https://gephi.org/>.

QUADRO 1 – Mapeamento de dados para o formato de rede executado no software OpenRefine

```
graph [
comment "Rede de contemplados por eixo e chamada do edital"
directed 0

node [ id {{jsonize(cells.idContemplado.value)}}
categoria "idContemplado"
grupo "contemplado"
eixo {{jsonize(cells.eixo.value)}}
chamada {{jsonize(cells.chamadaEdital.value)}} ]

node [ id {{jsonize(cells.eixo.value)}}
categoria "eixo"
grupo "eixo-chamada" ]

node [ id {{jsonize(cells.chamadaEdital.value)}}
categoria "chamada edital"
grupo "eixo-chamada" ]

edge [ source {{jsonize(cells.idContemplado.value)}} target {{jsonize(cells.
eixo.value)}} ]

edge [ source {{jsonize(cells.idContemplado.value)}} target {{jsonize(cells.
chamadaEdital.value)}} ]
]
```

Fonte: dados da pesquisa (2024).

O trabalho recorreu a técnicas de análise de redes sociais (ARS) como estratégia de descoberta de conhecimento, tratando relações entre autores, eixos temáticos e editais como estruturas analisáveis, especialmente adequadas a dados de afiliação (Borgatti; Halgin, 2014). Em termos do processo de KDD, a passagem das tabelas para grafos corresponde a uma transformação que permite aplicar medidas e algoritmos sobre padrões de conexão; no caso de redes bipartidas, projeções podem ser usadas para produzir grafos monopartidos e viabilizar rotinas como detecção de comunidades, desde que se explicitem critérios de ponderação e filtragem (Newman, 2010; Banerjee; Jenamani; Pratihari, 2017). Aliada à ARS, a inspeção visual foi empregada como parte da interpretação exploratória, pois direciona novas perguntas e valida hipóteses a partir de padrões emergentes (Nooy; Mrvar; Batagelj, 2018), em sintonia com a visualização do conhecimento como meio de comunicar insights para públicos com diferentes repertórios técnicos (Eppler; Burkhard, 2004; Burkhard, 2005). Dessa forma, foram usados os softwares Gephi, Looker Studio⁶ e uma planilha eletrônica nessa etapa de análise e interpretação, cujos artefatos visuais e resultados são apresentados na próxima seção.

⁶ Looker Studio é um *software* gratuito para *data visualization* e criação de relatórios e painéis interativos a partir de dados. Disponível em: <https://lookerstudio.google.com/>.

RESULTADOS E DISCUSSÃO

O levantamento de requisitos (**QUADRO 2**) foi fundamental para a organização dos dados do acervo em três coleções (bases de dados): projeto, contemplado (autor) e produto. Cada coleção foi organizada com as suas principais entidades que demandaram metadados para as suas descrições e representatividades nas bases de dados.

QUADRO 2 – Especificação de requisitos das principais entidades das bases de dados do Fundo

Entidade	Descrição	Coleção Tainacan
Dados gerais	Fornece informações administrativas sobre os lançamentos anuais dos editais, suas respectivas categorias e eixos.	Funcultura - Projeto
Projeto contemplado	Fornece informações sobre o projeto contemplado, incluindo título, proponente contemplado, o valor recebido, contatos do agente cultural e seu relatório final de prestação de contas.	
Produto [outra coleção a ser criada no Tainacan, e representada via tipo de metadado relacionamento]	Fornece informações sobre as diferentes ações entregues pelo proponente em forma de resultados do projeto contemplado. Logo, um edital pode se desdobrar em vários produtos. O produto, por sua vez, é o objetivo de um edital, ou seja, o resultado esperado pelo proponente ao submeter uma proposta de projeto.	
Contemplado	Pessoas envolvidas com o desenvolvimento do produto participante de um edital e projeto.	Funcultura – Contemplado
Categoria de Metadados	Com título, descrição e tipo de produto (lista de seleção).	Funcultura – Produto
Acervo	Fornece informações sobre um conjunto de objetos, históricos ou não, que necessitam estar reunidos para ocorrer sua compreensão.	
Audiovisual	Fornece informações sobre um produto que necessariamente apresente imagem e som de forma engajada.	
Documento técnico	Fornece informações registradas em papel ou meio eletrônico que possa ser usado para consulta, estudo ou validação técnica.	
Evento	Fornece informações sobre um acontecimento que reúna pessoas, presencialmente ou virtualmente, para determinado fim.	
Livro	Fornece informações sobre uma publicação física ou digital, composto por páginas, que necessariamente deve ter ISBN ou ISSN.	
Mestres da Cultura Popular	Fornece informações biográficas de personalidades da cultura capixaba.	
Música	Fornece informações de uma prática cultural que combine sons e ritmos.	
Site	Fornece informações sobre um sítio eletrônico criado como ou para o produto, acessado através de endereço HTTP ou HTTPS, disponível ou não, na web.	

Fonte: dados da pesquisa (2024).

A Matriz de Restrições com a relação dos metadados e seus respectivos perfis de aplicação foi elaborada a partir desses requisitos, evidenciando as decisões de modelagem para cada elemento presente na matriz, tornando-o, assim, uma representação da informação e do conhecimento diferenciada para as bases de dados do Funcultura, em que refletiram em metadados normalizados, padronizados e qualificados à luz dos vocabulários controlados propostos na presente pesquisa.

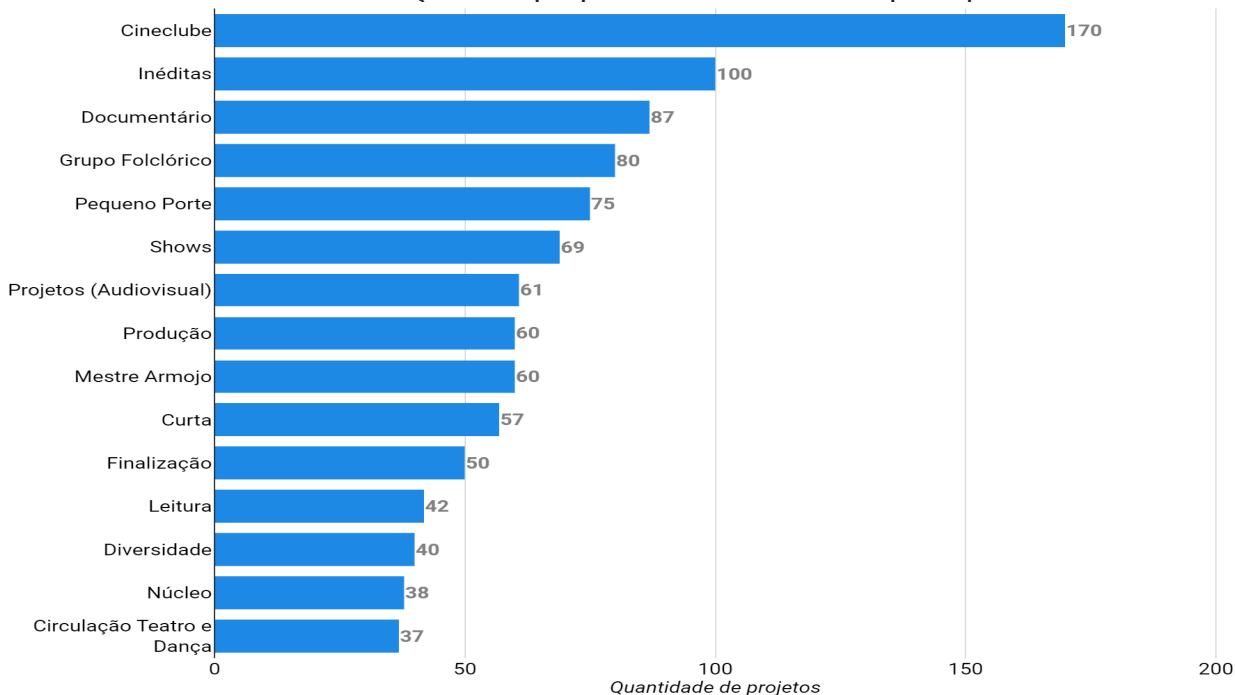
As taxonomias criadas para o acervo digital do Funcultura são em nove: autor de livro; categoria e eixo; estilo musical; local de atuação dos mestres; local de gravação; localidade; localidade do evento; mestre da cultura popular; e tipo de evento. Para além das taxonomias, vários arquivos de autoridade foram modelados para uso nos valores de dados de alguns metadados, as cardinalidades indicadas, os principais filtros atribuídos, a obrigatoriedade ou não na entrada de dados, as orientações de uso do metadado na perspectiva do usuário, entre outras configurações, garantindo, assim, a qualidade na catalogação dos dados e, consequentemente, na recuperação da informação.

Além de apoiar a recuperação, taxonomias e arquivos de autoridade contribuem diretamente para a qualidade orientada ao usuário, na medida em que reduzem ambiguidade e variabilidade na representação e tornam os dados mais compreensíveis e comparáveis para análises posteriores (Wang; Strong, 1996). Em termos institucionais, esse investimento em consistência e padronização é relevante porque erros, inconsistências e degradação de valores ao longo do tempo podem comprometer aplicações analíticas e decisórias, com impactos concretos na gestão (Eckerson, 2002). Além disso, o tratamento de registros ausentes e inconsistentes foi importante, pois a sua falta poderia afetar inferências e interpretações sobre o fenômeno estudado (McKnight *et al.*, 2007).

Finalmente, uma vez as bases de dados criadas com padrões de documentação fundamentados em princípios teórico-metodológicos da Ciência da Informação, a exemplo da modelagem de taxonomias apropriadas ao contexto de uso, tornou-se possível obter informações relevantes à Instituição a partir de dados qualificados, como pode-se observar nos gráficos e nas redes de informação projetados a seguir.

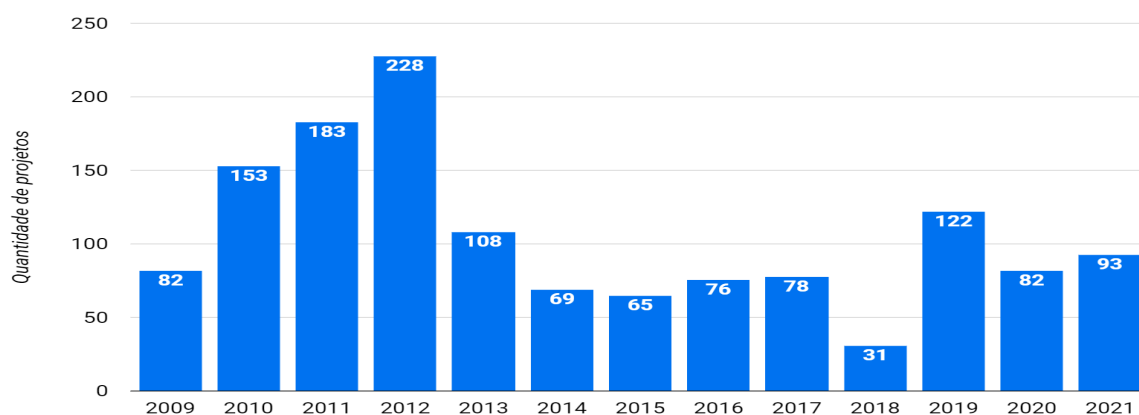
Os resultados de um breve levantamento estatístico sobre a quantidade de projetos distribuídos para os eixos temáticos do Fundo são mostrados no **GRÁFICO 1**. Destaca-se também a distribuição da quantidade de projetos ao longo dos anos de publicação dos editais, apresentada no **GRÁFICO 2**.

GRÁFICO 1 – Distribuição dos projetos culturais nos 15 principais eixos temáticos



Fonte: dados da pesquisa (2024).

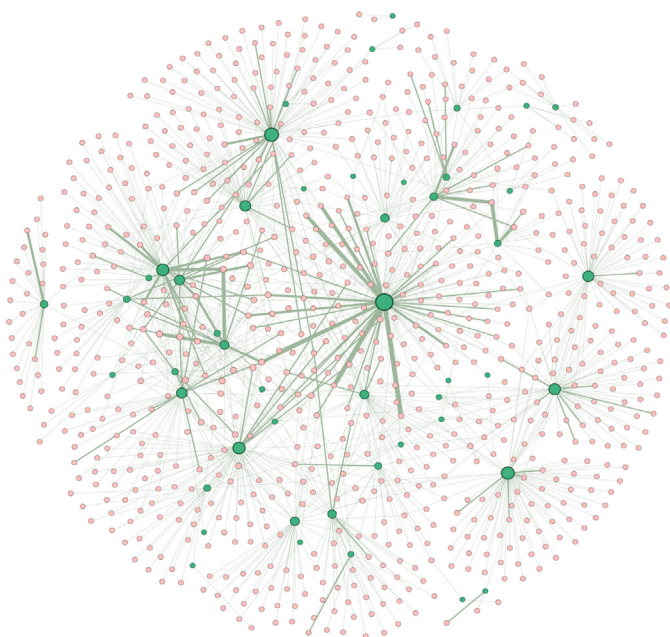
GRÁFICO 2 – Distribuição da quantidade de projetos ao longo dos anos



Fonte: dados da pesquisa (2024).

No que diz respeito às redes de informação, a rede bipartida é composta por nós avermelhados, que representam os autores contemplados, e nós esverdeados, que representam os eixos temáticos, conforme representado na **FIGURA 1**.

FIGURA 1 – Rede bipartida com destaque para eixos com maior incidência de inscrições de autores



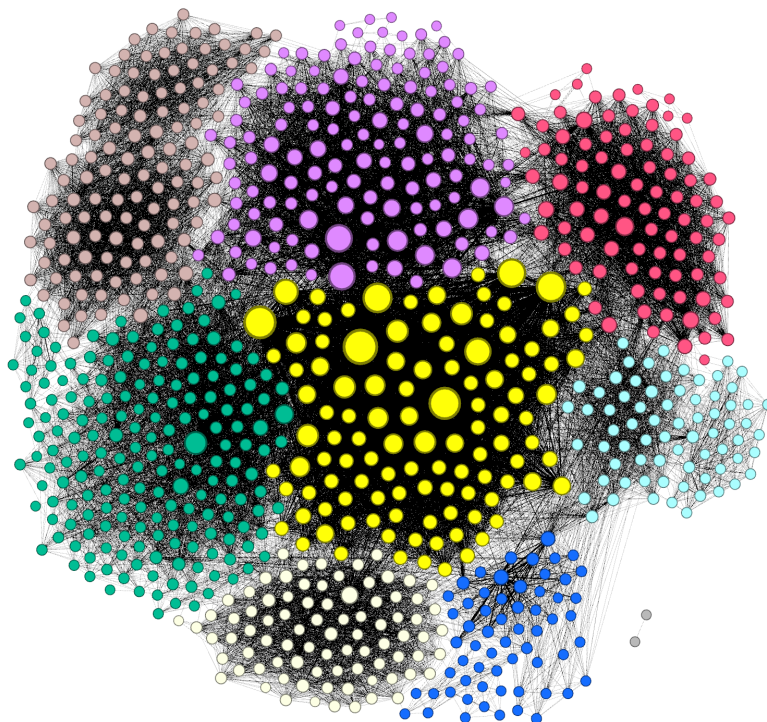
Fonte: elaboração própria (2024).

Essa rede destaca o tamanho dos nós de forma proporcional ao maior grau ponderado. Destaca também a espessura de arestas por aquelas com maior quantidade de conexões coincidentes entre autores com projetos de mesmo eixo temático. Nesse caso, o eixo temático “Cineclube”, na parte central da rede, é destacado, por estar associado a um maior número de autores com projetos contemplados e encontra-se em sintonia com o destaque em ocorrências absolutas que aparece no **GRÁFICO 1**. Além disso, devido a múltiplas conexões de muitos nós avermelhados (autores), a rede revela a existência de um grande número de autores com projetos associados a diversos eixos temáticos.

Do ponto de vista analítico, a **FIGURA 1** exemplifica uma rede de afiliação (bipartida) em que a ligação expressa coocorrência entre um autor e um eixo temático, e não necessariamente um vínculo social direto; por isso, leituras de proximidade e centralidade devem ser interpretadas como similaridade por participação temática (Borgatti; Halgin, 2014). A escolha por representar intensidade via grau ponderado e espessura de arestas torna explícita a força das coafiliações; entretanto, tais medidas são sensíveis a decisões de limpeza, normalização e ponderação, que alteram densidade e distribuição de graus e podem influenciar a percepção de eixos dominantes, como “Cineclube” (Banerjee; Jenamani; Pratihari, 2017).

A rede monopartida de autores contemplados, oriunda de uma projeção bipartida, é mostrada na **FIGURA 2**.

FIGURA 2 – Rede monopartida de autores contemplados



Fonte: elaboração própria (2024).

Para a obtenção dos nove grupos apresentados, destacados em cores diferentes, foi aplicado um algoritmo de clusterização e um algoritmo de distribuição disponíveis no software Gephi. O tamanho dos nós foi apresentado de forma proporcional ao grau ponderado dos mesmos.

A identificação desses grupos deve ser entendida como uma sumarização exploratória da estrutura relacional. Newman (2010) discute que algoritmos de detecção de comunidades procuram decompor redes em subgrupos mais densamente conectados internamente; nesse sentido, os agrupamentos coloridos fornecem uma hipótese interpretável sobre similaridade por coocorrência de eixos. Contudo, como a rede monopartida resulta de projeção bipartida, escolhas de projeção, ponderação e limiar afetam diretamente a densidade e, portanto, o resultado de algoritmos de comunidade (Banerjee; Jenamani; Pratihari, 2017), reforçando a necessidade de avaliar a estabilidade dos achados por inspeção e confronto com estatísticas e conhecimento de domínio, em um ciclo iterativo de interpretação (Fayyad; Piatetsky-Shapiro; Smyth, 1996).

Em consonância com a rede da **FIGURA 2**, a **TABELA 1** apresenta a ocorrência dos eixos temáticos nos oito principais grupos de autores contemplados mostrados na **FIGURA 2**. Para a obtenção de um resultado mais sistematizado, foram considerados apenas os eixos temáticos cujo percentual de ocorrência de contemplados é maior do que 10% sobre o restante dos eixos temáticos do mesmo grupo.

Os agrupamentos resultantes na rede da **FIGURA 2**, analisados por inspeção visual com auxílio da **TABELA 1**, revelam alguns padrões de incidência de autores que participam regularmente de projetos em eixos temáticos, tais como “Documentário”, “Projetos

(Audiovisual)”, “Curta” e “Finalização”, no grupo de cor rosa; “Circulação Teatro e Dança”, “Residência Teatro e Dança”, “Circo”, “Infantil” e “Lona de circo”, no grupo cor azul escuro. Por outro lado, o grupo de cor amarela, destaca o eixo “Cineclube” de forma isolada sugerindo uma interpretação de que seus autores (maiores graus ponderados da rede) tendem a permanecer no mesmo eixo.

TABELA 1 – Distribuição dos eixos temáticos nos oito principais grupos da rede monopartida de autores contemplados (**FIGURA 2**)

Grupo cor amarela		Grupo cor azul escuro	
Cineclube	96%	Circulação Teatro e Dança	26%
Grupo cor branca		Residência Teatro e Dança	20%
Shows	61%	Circo	18%
Setorial (Literatura)	22%	Infantil	18%
Grupo cor rosa		Lona de circo	15%
Documentário	26%	Grupo cor verde	
Projetos (Audiovisual)	26%	Pequeno Porte	26%
Curta	16%	Núcleo	14%
Finalização	13%	Leitura	12%
Grupo cor bege		Diversidade	12%
Grupo Folclórico	54%	Grupo cor azul claro	
Mestre Armojo	43%	Produção	49%
Grupo cor vermelha		Bolsa	37%
Inéditas	64%		

Fonte: dados da pesquisa (2024).

A análise exploratória de dados (Castro; Ferrari, 2017) é empregada para examinar, resumir e visualizar os dados, com vistas a entender características principais, identificar padrões, detectar *outliers* e formular hipóteses preliminares. No encadeamento do KDD, esse movimento se aproxima da etapa de interpretação/avaliação, em que padrões extraídos são confrontados com o contexto e julgados quanto à sua utilidade e plausibilidade em ciclos iterativos (Fayyad; Piatetsky-Shapiro; Smyth, 1996; Hand; Mannila; Smyth, 2001). Em análise de redes, a exploração visual é parte constitutiva do trabalho investigativo, pois permite navegar estruturas e validar leituras sobre grupos e conexões (Nooy; Mrvar; Batagelj, 2018), podendo ser compreendida como visualização do conhecimento ao apoiar a comunicação de insights e a reconstrução de entendimento por diferentes públicos (Eppler; Burkhard, 2004; Burkhard, 2005). Os Gráficos 1 e 2 e as Redes (Figuras 1 e 2) mostram que a análise exploratória pode ser apoiada pela sumarização em categorias pré-concebidas, viabilizada pelo uso de vocabulário controlado nas bases de dados. Desse modo, é possível perceber que o que distingue dado ou um conjunto de dados de informação é o conhecimento que ela propicia ao tomador de decisões (Hjørland, 2018).

Logo, na base de dados de Projetos, por exemplo, nota-se que o dado “identificação dos projetos”, representado pelo metadado “ID_Projeto”, é classificado numa categoria oriunda da taxonomia Categoria-Eixo. Por meio desta forma de modelagem de dados, foi possível gerar visualizações de informações importantes acerca das categorias culturais e eixos temáticos envolvidos nos editais, projetos e grupos de contemplados. O gestor de cultura da Instituição poderá, a partir de conhecimentos e habilidades técnicas, tomar decisões a partir das informações geradas frente às bases de dados, incluindo elaboração de estudos do perfil dos contemplados participantes de editais; extração de insights valiosos sobre o perfil do público, sua preferência em ações culturais, entre outras. Enfim, essas informações podem ser usadas pelos gestores para melhorar a experiência com os editais e tipos de projetos, além da possibilidade de criar editais personalizados e direcionar melhor recursos públicos e estratégias de captação de produtos e serviços culturais para o Estado.

Do mesmo modo, a partir da relação entre os dados “ano dos projetos” e “identificação dos projetos” (**GRÁFICO 2**), tornou-se possível projetar a distribuição dos projetos ao longo dos períodos na trajetória do Funcultura. Assim, o gestor do Fundo também poderá, a partir desse potencial indicador, tomar decisões de investir em programas e políticas públicas para setores culturais que se sobressaíram nos últimos anos; além de identificar tendências de interesse dos contemplados frente aos produtos desenvolvidos anualmente a partir dos projetos e seus eixos temáticos. Esses insights podem orientar o desenvolvimento de curadorias digitais na plataforma MEDIATECA, com trajetos narrativos temáticos do Fundo que estimulem a participação de artistas capixabas nos editais ofertados pelo governo do Estado.

CONSIDERAÇÕES FINAIS

O trabalho apresentou o percurso de construção e qualificação das bases de dados do Funcultura, articulando padrões de documentação, qualidade e ciclo de vida dos dados com uma estratégia de descoberta de conhecimento orientada a redes. A partir de análise exploratória e de redes de afiliação, foram identificados agrupamentos coesos de autores contemplados por similaridades nos eixos temáticos de seus projetos, produzindo evidências úteis para a compreensão de perfis de participação e tendências temáticas no Fundo.

O estudo reforçou a importância dos padrões de documentação para qualificar os dados, por meio de modelagem de metadados, vocabulários controlados e regras de catalogação que aumentam consistência e interpretabilidade, condição necessária para análises orientadas ao uso.

A análise de redes de informação revelou relações entre autores, editais e temas, evidenciando padrões de participação e distribuição de recursos, enquanto a análise exploratória apoiou a identificação de tendências para subsidiar políticas públicas culturais e curadorias digitais. Nesse contexto, a inspeção visual e a leitura de comunidades funcionam como mecanismos de sumarização e comunicação de padrões, aproximando o momento de interpretação/avaliação do KDD da perspectiva de visualização do conhecimento.

Como limitações, ressalta-se que a qualidade e a completude dos identificadores e descrições condicionam fortemente os resultados: ausências, inconsistências e registros incompletos podem fragmentar conexões e distorcer medidas e comunidades. Adicionalmente, por se tratar de rede de afiliação, a projeção bipartida e as escolhas de ponderação/limiar influenciam densidade e agrupamentos, exigindo cautela interpretativa e documentação de parâmetros. Como trabalhos futuros, recomenda-se testar análises de comunidade diretamente em redes bipartidas, comparar parametrizações e ampliar indicadores descritivos para robustecer a triangulação entre padrões de rede, estatísticas e conhecimento de domínio.

REFERÊNCIAS

- BAKER, T.; NILSSON, M.; JOHNSTON, P. DCMI. **The Singapore framework for dublin core application profiles**. 2008. Disponível em: <http://dublincore.org/documents/singapore-framework/>. Acesso em: 12 fev. 2026.
- BANERJEE, S.; JENAMANI, M.; PRATIHAR, D. K. Properties of a projected network of a bipartite network. **arXiv**, [s. l.], v. 1, July 2017. DOI 10.48550/arXiv.1707.00912.
- BORGATTI, S. P.; HALGIN, D. S. Analyzing affiliation networks. In: SCOTT, J.; CARRINGTON, P. (ed.). **The SAGE handbook of social network analysis**. London: SAGE Publications Ltd, 2014. DOI 10.4135/9781446294413.n28. cap. 28, p. 417-433.
- BURKHARD, R. A. Towards a framework and a model for knowledge visualization: synergies between information and knowledge visualization. In: TERGAN, S.-O.; KELLER, T. (org.). **Knowledge and Information Visualization**. Heidelberg: Springer Berlin Heidelberg, 2005. p. 238-255. (Coleção Lecture Notes in Computer Science, v. 3426). DOI 10.1007/11510154_13.
- CASTRO, L. N.; FERRARI, D. G. **Introdução à mineração de dados**. São Paulo: Saraiva, 2017.
- CRISTOVÃO, H. M.; VALE, L. S. Análise de redes sociais de autores e instituições participantes nos anais do WIDAT 2017 a 2023 por coautoria e palavras-chave. **Ciência da Informação**, Brasília, v. 53, n. 2, p. 248-270, 2024. Edição especial. DOI 10.18225/ci.inf.v53i.6923.
- ECKERSON, W. W. **Data quality and the bottom line**: achieving business success through a commitment to high quality data. Seattle: The Data Warehouse Institute, 2002. Disponível em: <http://download.101com.com/pub/tdwi/Files/DQReport.pdf>. Acesso em: 12 fev. 2026.
- EPPLER, M. J.; BURKHARD, R. A. **Knowledge visualization**: towards a new discipline and its field of application. Lugano: Università della Svizzera italiana, 28 July 2004. (paper 2, v. 2.5). Disponível em: <https://share.google/EzpLXmOlaYgdWiNxt>. Acesso em: 12 fev. 2026.
- ESPÍRITO SANTO. **Análise executiva do FUNCULTURA/ES**. 2. ed. Vitória: Instituto Jones dos Santos Neves, 2023. 61 p.; il. tab. (Relatório de Pesquisa).
- FAYYAD, U.; PIATETSKY-SHAPIO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI Magazine**, [s. l.], v. 17, n. 3, p. 37-54, Fall, 1996.
- GILLILAND, A. J. Setting the stage. In: BACA, M. (ed.). **Introduction to metadata**. 3. ed. Los Angeles: Getty Research Institute, 2016. *E-book*. Disponível em: <https://www.getty.edu/publications/intrometadata/setting-the-stage/>. Acesso em: 12 fev. 2026.

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. 227 p. (Coleção Metodologias de Pesquisa).

HJØRLAND, B. Data (with big data and database semantics). **Knowledge Organization**, [s. l.], v. 45, n. 8, p. 643–652, 2018. DOI 10.5771/0943-7444-2018-8-685.

LE MOS, D. L. S.; MARTINS, D. L.; MACEDO, D. F. Plataforma de agregação midiateca capixaba: projeto cultura digital nas instituições de memória do Espírito Santo. *In*: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 24., 2024, Vitória. **Anais [...]**. Vitória: ANCIB, 2024. Disponível em: <https://enancib.ancib.org/index.php/enancib/xxivenancib/paper/viewFile/2565/1516>. Acesso em: 12 fev. 2026.

MCKNIGHT, P. E.; MCKNIGHT, K. M.; SIDANI, S.; FIGUEREDO, A. J. **Missing data: a gentle introduction**. New York: The Guilford Press, 2007. il.

NEWMAN, M. E. J. **Networks: an introduction**. Oxford; New York: 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. England; New York: Cambridge University Press, 2018.

SANT'ANA, R. C. G. Ciclo de vida dos dados: uma perspectiva a partir da ciência da informação. **Informação & Informação**, Londrina, v. 21, n. 2, p. 116-142, maio/ago. 2016. DOI 10.5433/1981-8920.2016v21n2p116.

WANG, R. Y.; STRONG, D. M. Beyond accuracy: what data quality means to data consumers. **Journal of Management Information Systems**, [s. l.], v. 12, n. 4, p. 5-33, 1996.

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

AGRADECIMENTOS

Fundação de Amparo à Pesquisa e Inovação do Espírito Santo (FAPES). Secretaria da Ciência, Tecnologia, Inovação e Educação Profissional do Estado do Espírito Santo (SETIC).