Metodologias transdisciplinares em história das ciências: Mineração de dados em documentos históricos

Transdisciplinary methodologies in the history of science: Data mining in historical documents

Daniel Maia

Programa de Pós-Graduação em História das Ciências e das Técnicas e Epistemologia (HCTE), Universidade Federal do Rio de Janeiro (UFRJ)

danielmaia@labrec.pro.br

orcid.org/0000-0002-8458-2655

Regina Maria Macedo Costa Dantas

Programa de Pós-Graduação em História das Ciências e das Técnicas e Epistemologia (HCTE), Universidade Federal do Rio de Janeiro (UFRJ)

regina@hcte.ufrj.br

orcid.org/0000-0001-9782-2008

Resumo. O presente artigo faz parte de um trabalho em desenvolvimento contínuo referente à constituição de uma rede de agenciamentos informacionais, extraída de documentos históricos, notadamente periódicos científicos especializados. É o recorte da dissertação de mestrado apresentada anteriormente pelo autor, tratando de modelagem e análise computacional-imagética de dados, tendo como base a Teoria dos Grafos e a Teoria de Redes, tema de crescente interesse na Ciência da Informação e abordagem recente no campo da Metodologia da História. Visa fortalecer a área de História das Ciências e as conexões transdisciplinares como parte de uma renovação das políticas científicas em procedimentos de inovação como elementos indispensáveis das políticas de Estado para as Ciências e as Tecnologias.

Palavras-chave: História das Ciências no Brasil. Transdisciplinaridade. Teoria de Redes, Humanidades Digitais

Abstract. *This article is part of a work in continuous development regarding the constitution of a network of informational agencies, extracted from historical documents, notably specialized scientific journals. It is the excerpt of the master's thesis previously presented by the author, dealing with data modeling and computational-imagery analysis, based on Graph Theory and Network Theory, a topic of growing interest in Information Science and a recent approach in the field of Methodology of History. It aims to strengthen the area of History of Science and transdisciplinary connections as part of a renewal of scientific policies in innovation procedures as indispensable elements of State policies for Science and Technology in Brazil.*

Keywords: *History of Science in Brasil. Transdisciplinarity. Network Theory, Digital Humanities*

Recebido: 01/10/2017 Aceito: 27/10/10 Publicado: 07/11/2017

Introdução

O desenvolvimento da microeletrônica e as potencialidades que os computadores trouxeram para as ciências nos proporciona hoje um novo horizonte de aplicações sociotécnicas, até mesmo para a ciência histórica, o que requer maior atenção de historiadores profissionais na aquisição de habilidades além do seu campo de atuação, provocando também uma mudança nas mentalidades institucionalizadas das ciências e dos cientistas. Concentrados na formalização e afirmação disciplinar da História desde o século XIX, os historiadores estabeleceram fronteiras epistemológicas mais seguras para analisar o processo de desenvolvimento das civilizações no tempo. No entanto, com o desenvolvimento científico e tecnológico, mudanças profundas surgiram no seio das sociedades que se encontravam num processo de modernização. Passados mais de cem anos, tendo adentrado em um novo milênio, a comunidade científica se vê na necessidade de reformular processos metodológicos no que diz respeito às pesquisas documentais, em particular no âmbito da História.

A partir de uma abordagem quali-quantitativa, tendo o documento como base discursiva, é possível transcrever a linguagem escrita em uma outra, híbrida, indexável e dinâmica, produzindo uma nova categoria de fonte, capaz de revelar novas relações, que de outro modo não seria possível, ou mesmo custoso para o pesquisador de identificar. Um modo de fazer historiográfico que considera um conjunto de fontes que perpassa diversos domínios teórico-metodológicos, e ao tratar a documentação como uma série de dados, tem-se no software de computador um instrumento indispensável na construção de análises mais sofisticadas que utilizam teorias matemáticas para a visualização de dados históricos.

Assim, o estudo de caso realizado na dissertação de mestrado se utilizou do periódico *Archivos do Museu Nacional,* disponível digitalmente e parte do acervo da

Seção de Memória e Arquivo (SEMEAR) do Museu Nacional/UFRJ, como fonte primária para a extração e classificação de dados a serem utilizados na modelagem por software para a criação de um novo tipo de fonte documental: um grafo. O software Gephi[1] foi desenvolvido para visualização e exploração de dados, podendo o pesquisador interagir com a representação visual, manipular as estruturas e suas propriedades para revelar padrões ou relações que de outro modo seria difícil de se estabelecer.

As capacidades computacionais do referido software traduzem o esforço epistemológico dessa proposta de modelagem, visto que é dotado de algoritmos capazes de criar layouts e métricas a partir da transcrição dos dados trabalhados ao longo da pesquisa. A representação espacial da informação então pode ser visualizada de uma forma inovadora enquanto proposta teórico-metodológica, produzindo, como dito anteriormente, uma fonte de pesquisa híbrida e dinâmica, algo novo e necessário diante dos problemas ainda enfrentados por profissionais que lidam com documentação não-indexada[2]. O uso do Gephi é bastante apropriado, visto que o modo operacional por trás de suas funções tem como base a teoria de grafos, o que produz uma leitura sobre as redes na historiografia e metodologia da História ainda pouco trabalhada no campo.

A teoria dos grafos é um ramo da matemática que estuda as relações entre os objetos de um determinado conjunto. Tais objetos são denominados *nós,* e as relações entre eles, *arestas*. Estruturas que podem ser representadas por grafos estão em toda parte e muitos problemas de interesse prático podem ser formulados como questões sobre certos grafos. Na pesquisa em questão estabeleceu-se uma analogia conceitual entre as características dos elementos de um grafo e os elementos textuais dos documentos analisados. Tais elementos abrangiam referências à pessoas, lugares ou mesmo objetos, graus de influência ou acontecimentos.

Tal abordagem demonstra as potencialidades da pesquisa transdisciplinar, como também revela novos desafios e problemas. A Teoria de Redes[3] tem também nos fornecido análises cada vez mais precisas sobre as relações entre entidades, aqui especialmente as relações sociais entre indivíduos e instituições dentro de uma cultura científica em desenvolvimento. A formação de agrupamentos ou *clusters* traduz as crises paradigmáticas dos campos do conhecimento, a dinâmica da transmissão de conhecimento e a formação e desaparecimento de comunidades científicas e seus agentes.

Mapear o conhecimento científico parece ocupar, já há algum tempo, o campo intelectual de uma pequena parcela de especialistas (BERNAL, 1946, LEYDESDORFF, 2001). Ainda assim, é de fato o desvelamento de novos resultados e perspectivas inovadoras. Ademais do fato que a linguagem que usamos para descrever

1 https://gephi.org

2 Por documentação não-indexada entende-se aqui como elementos textuais de documento escaneado que não são possíveis de selecionar ou pesquisar de forma automatizada.

3 Redes enquanto construção social, elaborada pelo sociólogo Manuel Castells, um dos precursores de uma abordagem interdisciplinar sobre as Redes Sociais.

as ciências é repleta de metáforas espaciais, como "campo" e "área" de pesquisa, quando tentamos verdadeiramente criar um mapa das ciências, logo percebemos que os procedimentos usados para fazer mapas de conotação espacial estão se tornando insuficientes.

Devemos lidar com as associações e relações abstratas entre entidades como ideias científicas, especialidades, campos ou disciplinas cujas próprias existências podem estar abertas à crítica. Chaomei Chen lança a pergunta em *Mapping Scientific Frontiers* (CHEN, 2013, p. vi): faz sentido procurarmos uma representação espacial de tais entidades abstratas, ou mesmo hipotéticas? Nossos cérebros estão programados para considerar o que é relacional e projetar isso no espaço real? Considerando o desenvolvimento tecnológico na criação de interfaces que estendem as capacidades de representação da cognição humana, ao projetar abstrações num espaço virtual cheio de pixels, podemos dizer que sim. Ainda segundo Chen, a apreensão do mapeamento científico pode parecer difícil devido a três estágios conceituais requeridos para dar sentido ao processo como um todo. Primeiro, uma unidade de análise deve ser escolhida para abranger as partículas elementares do universo científico. Segundo, uma medida de associação entre as unidades deve ser definida. Terceiro, um meio deve ser encontrado para retratar as unidades e suas relações num espaço dimensional perceptível, geralmente duas dimensões.

Críticos da cientometria (HARRIS, 2006; STYHRE, 2003; WILSON, 2002) alegam que o foco na literatura científica como sua fonte primária limita severamente os dados cujos estudos de ciências podem ser baseados. Por outro lado, a crescente disponibilidade de textos completos de artigos científicos, em formatos que podem ser lidos por computador, abre muitos novos tipos de dados para análise os quais, quando usados em conjunto com os bancos de dados online padronizados vão muito além do que tem sido possível usando apenas os índices padronizados.

Os agenciamentos do processo cartográfico das ciências, entre as várias escolhas para unidades de análise como palavras, referências, autores, periódicos, e os meios de associá-los como palavras relacionais, citações cruzadas ou diretas e coautoria parecem se reduzir aos tipos de estruturas e níveis relacionais que queremos observar.

Outra questão importante é a interpretação dos mapeamentos. É sabido que os agenciamentos em rede que são representados em mapas são hiperdimensionais, e que a projeção em duas dimensões é inevitavelmente uma aproximação que pode colocar duas unidades pouco relacionadas muito próximas. Isso requer a necessidade de prestar bastante atenção aos agenciamentos propriamente ditos, os links, o que faz surgir a solução bidimensional em primeiro lugar, ao tentar visualizá-los como rede neural.

Apenas sabendo o que os agenciamentos significam teremos uma melhor compreensão do que o mapeamento representa. Isso envolve olhar mais profundamente para o contexto do plano de forças atuando no espaço desses agenciamentos, e procurar novos modos de representação e categorização dessas relações, se tem função causal, lógica, social, hipotética ou metafórica. Em última instância, tais análises visuais têm

como objetivo servir de suporte para decisões estratégicas, particularmente no que diz respeito às políticas científicas.

A particularidade do mapeamento científico é perceber o horizonte constantemente mutável, em que, ano após ano, uma nova leva de artigos publicados, provocam mudanças estruturais e fazem surgir novas áreas, evoluem outras tantas, e algumas acabam por sucumbir. O acaso é inerente às ciências, assim como seu mapeamento. Não se sabe se as descobertas podem ser previstas, se existem antecedentes reconhecíveis ou condições, se podem ser programadas para acontecer antecipadamente. Mas, como as descobertas ficam aparentes nos mapas após sua ocorrência, também há a possibilidade de estudar os mapeamentos de períodos antecedentes e procurar por suas estruturas fundacionais. É nesse sentido que se abre espaço para a abordagem da teoria dos grafos enquanto abstração matemática para a elaboração do modelo computacional-imagético proposto na pesquisa.

Para ilustrar o problema proposto no presente artigo, apresento três situações possíveis na extração de dados de documentos históricos. Considerando a utilização de softwares que podem ou não automatizar o processo de construção de mapas relacionais, isso definirá a precisão do modelo e da própria análise das informações.

Agardh (G. H.)
Baillon (Henrique).
Barboza du Bocage. (J. V.)
Beaurepaire Rohan (Henrique de)
Beneden (Ed. Van).
Benthan (Jorge).
Bom Retiro (Visconde do)
Braun (Alexandre).
Bureau (Eduardo).
(Candolle (Affonso de).
Coelho d'Almeida (Thomas J.)
Darwin (Carlos).
Decaisne (José).
Delpino (F.)
Duchartre (Pedro).
...

Figura 1. Possibilidades e restrições para mineração de dados textuais em 3 estapas .
Fonte: Próprio Autor.

Como demonstrado na Figura 1, o processamento de informações em documentação histórica pode sofrer inconsistências dependendo da técnica aplicada. Na primeira coluna, temos o documento em estado bruto, apenas escaneado; na segunda coluna, o documento passa por um processo de reconhecimento óptico de caracteres (OCR), o que ainda não garante cem por cento de precisão, mas já permite a seleção de textos; e por fim, um texto que já se origina em seu formato digital possibilita a transcrição direta de informações na construção de um modelo computacional-imagético.

A problemática da representação de agenciamentos, quando estes se manifestam como associações subjetivas, na esfera do social e do mental, requer a construção de metáforas visuais que traduzam o espaço de fluxo de modo a validar o não-dito.

"Atributos visuais de configurações topológicas e geométricas tem de transformar o conhecimento intangível e invisível em algo concreto e significativo" (CHEN, 2013).

O processo de transcrição do documento histórico textual em uma fonte híbrida, constituída de camadas de informação digital e visual, é o modo como esse método transdisciplinar de modelagem pode construir um novo entendimento sobre o conhecimento. Torna legível as infinitas associações visuais subjetivas, onde, em "isolamento" o texto acaba por permanecer numa linearidade inescapável. Enquanto estudo histórico das instituições científicas, podemos compreender mais facilmente a dinâmica entre os sujeitos, onde o processo cognitivo classifica os caminhos, os clusters, as relações, como um organograma fluido. Quanto mais associações são feitas no espaço de fluxo, melhor nossa compreensão sobre sua dinâmica e suas transformações estruturais.

O pensamento visual é uma fonte peculiar de estruturação e processamento de conceitos, onde voltamos nossa atenção para o campo como um todo, depois partindo para os detalhes. No processo da abstração o indivíduo pode reestruturar ou mesmo transformar os conceitos, resultando disso uma experiência concreta agora visível. A modelagem computacional-imagética realizada a partir da extração de informações textuais de cunho histórico nos oferece caminhos para uma análise transdisciplinar como complemento da visualização da informação complexa. Trata-se da constante comunicação entre métodos analíticos e processos cognitivos, que juntos geram insights visuais para revelar novas interpretações. São construções metafóricas, muitas vezes implícitas, que a informação textual isolada não é capaz de revelar.

O uso do software Gephi apresenta, para um pesquisador fora da área da computação, uma curva de aprendizagem bastante acentuada. Ainda assim, o programa apresenta recursos de visualização e manipulação de dados que se adéquam as necessidades do usuário. Uma das características utilizadas no software foi a filtragem de dados e a construção da topologia de rede. Dentre inúmeras possibilidades de filtragem, foi escolhida a Rede Egocentrada com a indicação manual de um nó específico da rede, nesse caso, o Museu Nacional. Quanto a topologia da rede, o software disponibiliza layouts que organizam a distribuição de acordo com o número de arestas e nós interconectados, ou para usar o vocábulo mais presente na pesquisa, de acordo com os agenciamentos estabelecidos entre sujeitos, instituições e a produção científica.

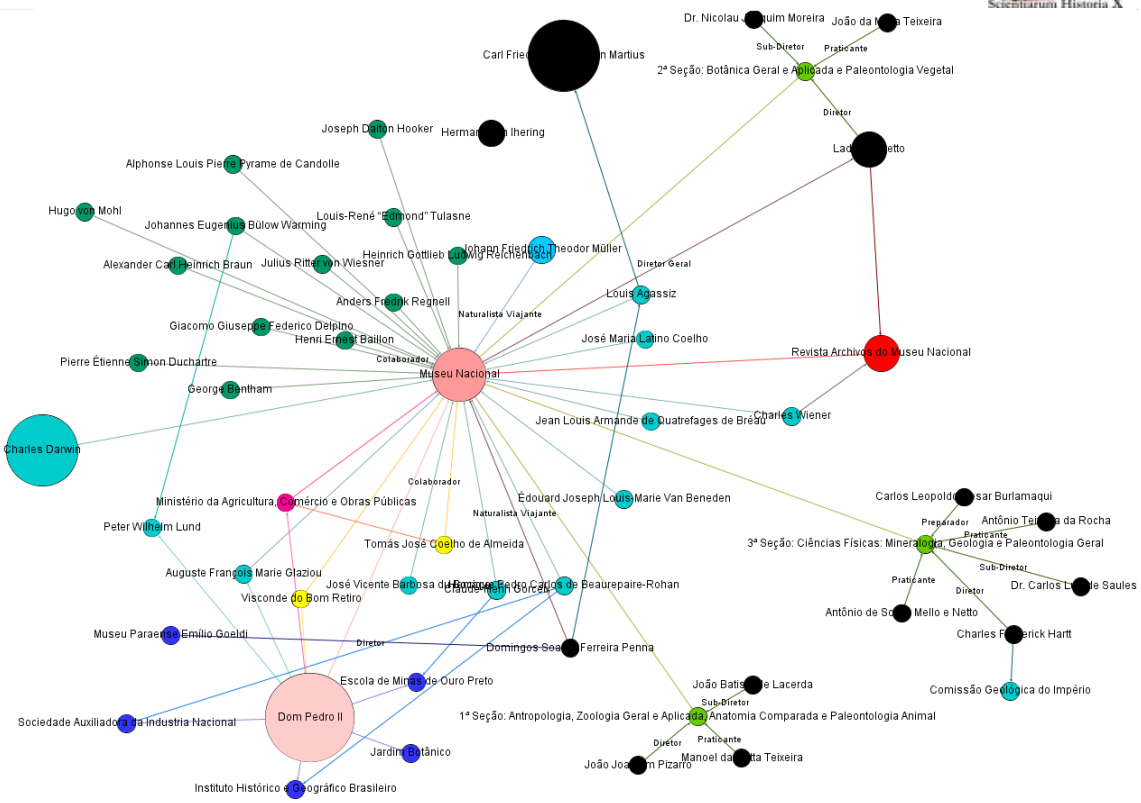

Figura 2. Grafo da Rede de Relações Científicas Institucionais do Museu Nacional.

Fonte: Próprio Autor

A Figura 2 demonstra o resultado da modelagem feita a partir da extração manual dos dados textuais do periódico *Archivos do Museu Nacional*, e assim é possível verificar a formação de padrões de agrupamento, como também o grau de influência de determinados nós da rede, não apenas pelo número de conexões que se atravessam mas também pelo seu papel na topologia da rede.

O exemplo emblemático dessa modelagem foi a interpretação do espaço ocupado por Dom Pedro II (destacado em rosa claro) na Rede: apesar de ser um elemento de grande influência na formulação de políticas científicas do Império ele não era um elemento central da estrutura comunicacional em análise. Já o Museu Nacional, por meio de seu periódico (destacado em vermelho), determina as capacidades comunicacionais entre os demais elementos da rede, visto que era a partir da instituição que se produzia a revista e também por meio dela, se estabelecia o vínculo com pesquisadores brasileiros e estrangeiros, com destaque para Charles Darwin (destacado em verde claro).

Conclusão

Sobre as dificuldades da pesquisa, o primeiro obstáculo observado na análise documental, levando-se em consideração a preocupação com a eficácia da extração dos dados, como a impossibilidade da identificação automática dos termos sensíveis na elaboração de um banco de dados relacionais. No trabalho historiográfico, a análise documental representa uma grande parcela do tempo de dedicação à pesquisa. Ainda assim, tendo hoje disponível tecnologia capaz de reconhecer padrões de escrita, conhecida como OCR (Optical Character Recognition), tal recurso não pôde ser utilizado com eficácia devido o tempo que levaria para corrigir a imprecisão do próprio procedimento, fazendo com que fosse descartado para a pesquisa então realizada.

Isso gerou uma dificuldade na precisão da coleta de dados e a própria construção dos grafos, visto que o software poderia ser alimentado de forma automática com o *output* do processo de reconhecimento de caracteres. Em conformidade com as condições de coleta e extração de dados, a pesquisa prosseguiu também com uma bibliografia complementar de origem eletrônica, a partir de livros e artigos encontrados na Rede Mundial de Computadores. Isso nos traz ao final dos questionamentos da pesquisa no que diz respeito não só a elaboração de novos métodos e sobre a criatividade individual do cientista, mas também sobre as políticas científicas elaboradas no seio das instituições governamentais e das comunidades acadêmicas e os custos financeiros restritos à um modelo de política protecionista que não proporciona acesso à informação, à própria comunidade científica e ao público em geral.

A questão da acessibilidade científica está presente desde antes do surgimento dos periódicos especializados, e sendo a presente pesquisa centrada justamente nesse tipo de fonte como objeto, foi possível perceber como a dinâmica das descobertas científicas se modifica quando estas são disponibilizadas abertamente por meio do agenciamento de redes de sujeitos e instituições.

Financiamento

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Referências

ARCHIVOS DO MUSEU NACIONAL. Museu Nacional: Imprensa Industrial, v. 1, 1876.

BERNAL, J. D. **The Social Function of Science**. London: G. Routledge & Sons Ltd, 1946.

CHEN, C. **Mapping Scientific Frontiers**. 2. ed. Pennsylvania: Springer, 2013.

HARRIS, K. Knowledge management enables the high performance workplace. **Gartnet Inc**, 2006. Acesso em 08 jan 2017. Disponível em: https://www.gartner.com/doc/489448/knowledge-management-enables-highperformance-workplace.

LEYDESDORFF, L. **The Challenge of Scientometrics**: the development, measurement, and self-organization of scientific communications. 2.ed. Florida, USA: Universal Publishers, 2001.

MAIA, D. **Uma Cartografia de Redes Institucionais: método transdisciplinar e modelagem computacional-imagética da História das Ciências no Brasil**. Rio de Janeiro: 2017. Dissertação (Mestrado em História das Ciências, das Técnicas e Epistemologia) – Universidade Federal do Rio de Janeiro.

STYHRE, A. **Understanding knowledge management**: Critical and postmodern perspectives. Copenhagen: Copenhagen Business School Press, 2003.

WILSON, T. D. The nonsense of 'knowledge management'. **Information Research**, 8(1), 2002.