





## EVIDÊNCIAS, CÓDIGOS E CLASSIFICAÇÕES: O OFÍCIO DO HISTORIADOR E O MUNDO DIGITAL

Evidences, codes and classifications: the historian's craft and the digital world

**Alexandre Fortes<sup>a</sup>**

 <https://orcid.org/0000-0002-3728-2318>  
E-mail: fortes.ufrj@gmail.com

**Leandro Guimarães Marques Alvim<sup>b</sup>**

 <https://orcid.org/0000-0002-1611-7559>  
E-mail: alvim.lgm@gmail.com

<sup>a</sup> Universidade Federal Rural do Rio de Janeiro, Instituto Multidisciplinar, Departamento de História, Nova Iguaçu, RJ, Brasil.

<sup>b</sup> Universidade Federal Rural do Rio de Janeiro, Instituto Multidisciplinar, Departamento de Ciência da Computação, Nova Iguaçu, RJ, Brasil.

**DOSSIÊ**

**História digital e global: novos horizontes para a investigação histórica**

## RESUMO

O artigo examina o impacto da difusão global das tecnologias digitais sobre o ofício do historiador. Parte da análise sobre a relação entre a prática da profissão e natureza do conhecimento histórico formuladas por alguns dos maiores historiadores do século XX. Examina a natureza social da linguagem e seu papel na constituição das evidências e fontes históricas, articulando essa análise com os avanços tecnológicos do “processamento de linguagem natural”. Discute conceitos de diversos ramos das ciências sociais relevantes para a compreensão do processo de desenvolvimento do conhecimento humano e o papel da codificação de informações na elaboração de narrativas e na pesquisa histórica. Por fim, apresenta um panorama das principais metodologias no campo da inteligência artificial atualmente aplicadas à pesquisa histórica.

## PALAVRAS-CHAVE

História Digital. Inteligência artificial. Teoria e Metodologia da História.

## ABSTRACT

This article examines the impact of the global diffusion of digital technologies on the historian’s craft. It is based on an analysis of the linkage between the praxis of the profession and the nature of historical knowledge as stated by some of the greatest historians of the 20<sup>th</sup> Century. It examines the social nature of language and its role in the constitution of evidences and historical sources, and draws connections between this analysis and the technological advances of “natural language processing”. It discusses concepts from various branches of the social sciences that are relevant for understanding the developmental process of human knowledge and the role of information codification in the construction of historical narratives. Finally, it presents an overview of the main methodologies in the field of artificial intelligence currently applied to historical research.

## KEYWORDS

Digital History. Artificial intelligence. Theory and Methodology of History.

**H**á hoje uma crescente percepção de que a revolução tecnológica em curso nas últimas décadas vem transformando aceleradamente todos os aspectos da vida humana, e que, a despeito das imensas desigualdades e disparidades entre países e regiões, esse processo ocorre em escala global. O exercício das mais variadas profissões e o processo de produção e difusão de conhecimento nas mais diversas áreas vêm sendo igualmente afetados de formas cada vez mais profundas. Em geral, a incorporação da tecnologia digital ao trabalho dos pesquisadores ocorre de maneira mais ou menos casual, à medida que os equipamentos eletrônicos se tornam mais acessíveis, os *softwares* se tornam mais conhecidos e amigáveis e as plataformas se tornam componente obrigatório das mais diversas atividades inerentes à vida acadêmica. A tomada de consciência de que os fundamentos básicos do ofício precisam ser revistos e atualizados diante do novo contexto, entretanto, ocorre de forma mais lenta e desigual.

Seria difícil identificar algum aspecto da atividade relativa à produção do conhecimento histórico que não tenha sido significativamente alterado nos últimos quinze ou vinte anos. A grande maioria dos historiadores atuais escreve em processadores de texto e usa, ao menos eventualmente, imagens e planilhas digitais e *softwares* para preparar apresentações. Crescentemente, gerenciadores de referências bibliográficas, bancos de dados, ferramentas de georreferenciamento e programas de suporte à análise qualitativa passam a ser também incorporados como importantes ferramentas de trabalho. As instituições de ensino e pesquisa adotam cada vez mais sistemas integrados de gestão acadêmica e administrativa. A participação em simpósios, a publicação de artigos, a divulgação dos currículos, a obtenção de financiamento, a gestão de programas de pós-graduação e a participação em associações científicas são todos mediados por plataformas online. Identificadores permanentes de autores e publicações, como International Standard Book Number (ISBN), International Standard Serial Number (ISSN), Digital Object Identifier (DOI) e Open Researcher and Contributor ID (ORCID) tornam-se cada vez mais familiares. Fontes históricas dos mais diversos tipos, sobre os mais variados temas e períodos, são hoje disponibilizadas de forma massiva na internet, enquanto fichas e cadernos de notas são substituídos por câmeras digitais e *scanners* quando a ida ao arquivo físico ainda é necessária. Historiadores interagem em âmbito global entre si, com alunos e com o público em geral trocando e-mails, postando em redes sociais – algumas delas criadas especificamente para uso de acadêmicos – e realizando videoconferências. A construção de sítios eletrônicos e *blogs*, a divulgação de vídeos didáticos e *podcasts* passam a ser instrumentos cada vez mais relevantes de divulgação científica e de história pública.<sup>1</sup>

A atuação nesse ambiente digital que permeia crescentemente a produção e circulação do conhecimento histórico desafia os pesquisadores a irem além de uma compreensão instrumental e consumista das novas tecnologias. Cada vez mais é necessário posicionar-se diante dos grandes enfrentamentos políticos relativos a questões como propriedade intelectual e políticas de informação científica. Afinal de contas, esses embates contrapõem interesses diversos, e muitas vezes antagônicos,

<sup>1</sup> Um panorama muito mais detalhado dessas transformações e de experiências de exploração ativa do potencial das tecnologias digitais pelos historiadores pode ser visto em trabalhos como de Cohen e Rosenzweig (2006), Kelly (2013), e Body e Larson (2014).

e o seu resultado determinará cada vez mais as condições de produção da ciência mundial. De um lado, encontram-se as grandes corporações multinacionais que oligopolizam não apenas o mercado de *softwares*, mas também o de publicações acadêmicas e das requintadas ferramentas de análise bibliométrica, que possuem grande impacto sobre a avaliação da produção científica e sobre as políticas de financiamento à pesquisa. De outro, o movimento global pelo desenvolvimento dos *softwares* livres, do acesso aberto e da ciência aberta, com a adoção de repositórios institucionais abrangendo produção científica e bases de dados. Ainda em um terceiro campo, a ação direta de ativistas como a cazaque Alexandra Elbakyan, condenada e perseguida por muitos, mas também reverenciada por um número ainda maior de pesquisadores, particularmente no Hemisfério Sul, que encontram na atividade pirata do *Science Hub* um importante instrumento para diminuir o *gap* que os afasta das condições de trabalho dos países ricos.

Na perspectiva da história do trabalho, seria difícil imaginar que uma atividade profissional atravessasse um contexto marcado por mudanças tão profundas nas técnicas e instrumentos de produção, nos mecanismos que definem sua inserção no processo de circulação de artefatos culturais, com o correspondente enfrentamento de uma agenda política completamente nova, sem passar por uma profunda metamorfose.

O exercício que propomos nesse artigo passa, em um primeiro momento, por resgatar reflexões sobre a natureza do trabalho do historiador produzidas por mestres que, em momentos distintos do século XX, realçaram seu caráter artesanal. Posteriormente, tomando como fio condutor a visão da História como ciência da informação, analisaremos os processos de codificação e decodificação como aspectos inerentes ao ofício do historiador, buscando continuidades e discontinuidades no contexto contemporâneo. Apresentaremos ainda um painel de diversas metodologias computacionais aplicadas à documentação histórica atualmente em desenvolvimento. Por fim, trataremos da necessidade de rever procedimentos tradicionais à luz da atual confluência interdisciplinar que afeta os processos de produção e circulação do conhecimento histórico, visando indicar possibilidades de desenvolvimentos teórico-metodológicos inovadores.

## **CLASSIFICAÇÃO EM BUSCA DA INTELIGIBILIDADE PROGRESSIVA: A OFICINA DO HISTORIADOR**

Dois dos mais influentes historiadores do século XX registraram suas visões sobre os grandes debates teórico-metodológicos relativos à disciplina histórica em obras marcadas pela analogia entre a pesquisa histórica e o trabalho artesanal. Para Marc Bloch (2001) e E. P. Thompson (1981), ao burilar as fontes, materiais de caráter “objetivo e determinante”, os historiadores forjam um conhecimento validado pelo diálogo com a realidade, voltado à orientação da ação humana no presente. Bloch destaca que a produção historiográfica não se destina à produção de um saber normativo, mas sim à geração de uma classificação racional das informações em busca de uma progressiva inteligibilidade do processo histórico (BLOCH, 2001, p. 129).

Ambos destacam que a utilização de diferentes tipos de fontes em uma mesma investigação possibilita uma maior aproximação da complexidade do real, mas também que a transformação das fontes (epistemologicamente inertes) em conhecimento histórico é mediada pela formulação de problemas (BLOCH, 2001) ou de perguntas



que conduzam o interrogatório das evidências (THOMPSON, 1981). Essas reflexões sobre a natureza do ofício e do conhecimento que ele produz nos levam a dois tipos de indagações em relação ao impacto da revolução digital na pesquisa histórica.

Em primeiro lugar, trata-se de analisar o potencial da massiva ampliação do universo de fontes potencialmente acessíveis e das ferramentas tecnológicas capazes de auxiliar (e até mesmo automatizar) a “classificação racional de informações” na produção de análises de qualidade superior no que diz respeito à “inteligibilidade do processo histórico”.

Em segundo lugar, cabe refletir sobre os historiadores como profissionais treinados em uma disciplina dedicada a esse sofisticado processo de observação e análise capaz de gerar, a partir de vestígios oriundos de múltiplas temporalidades, novos e substantivos conhecimentos. Qual é a contribuição do ofício para o enfrentamento dos dilemas enfrentados pela sociedade em uma era marcada por fenômenos como *Big Data*, inteligência artificial (IA) e *Fake News*? Em que medida as habilidades intelectuais nutridas pela pesquisa histórica podem dialogar com a ciência da computação na geração de novas tecnologias orientadas a potencializar a ação humana diante dos desafios do presente e do futuro?

Para Jo Guldi e David Armitage, os historiadores, ao se apropriarem das ferramentas analíticas geradas pela tecnologia digital, podem desempenhar papel fundamental no enfrentamento da “sobrecarga de informação” que ameaça a capacidade de pensamento de longo prazo no mundo contemporâneo. Em comentários que ecoam, em um contexto profundamente alterado, os de Bloch nos anos 1940 e de Thompson nos anos 1970, Guldi e Armitage (2014, p. 88-117) caracterizam o ofício como particularmente apto ao exercício crítico do papel de curadoria que envolve problematizar simultaneamente múltiplas bases de dados.

## LINGUAGEM “NATURAL”, CODIFICAÇÃO E DIGITALIZAÇÃO EM PERSPECTIVA HISTÓRICA

As transformações sociais desencadeadas pela revolução digital colocam a necessidade de examinarmos as novas formas assumidas pelo trabalho do historiador na “classificação racional das informações obtidas nas fontes em busca de uma progressiva inteligibilidade do processo histórico”. A primeira questão a ser considerada é que a classificação de informações não está presente apenas na leitura e análise das fontes, mas na sua própria constituição.

Considerando a distinção de Bloch (2001) entre evidências diretas e indiretas, o papel da linguagem no segundo tipo é bastante evidente, uma vez que ele se constitui de relatos ou descrições de acontecimentos humanos, que só podem ser elaborados a partir da combinação de palavras em estruturas narrativas organizadas por regras gramaticais formais e informais. Sobre essa camada textual básica, em certos casos, diferentes tipos de codificação podem ser inseridos nos processos de manipulação pelos quais essas fontes podem passar até serem depositadas em um arquivo ou abandonadas ao esquecimento: anotações e comentários marginais, códigos de classificação temática, metadados de catalogação, etc. O historiador, por sua vez, ao ler a documentação, criará sua própria lógica de seriação e classificação, selecionará trechos de seu interesse e os identificará nas suas fichas físicas ou virtuais,

associando-os a diversos tipos de marcadores, tais como conceitos, datas, indivíduos, organizações, acontecimentos, entre outros.

As evidências diretas, por outro lado, muitas vezes não contêm linguagem humana em si mesmas, mas só poderão se constituir em fontes à medida que os pesquisadores as representem por meio de códigos linguísticos. Fragmentos encontrados em uma escavação arqueológica, por exemplo, precisam ser descritos, quantificados e analisados em relação à sua distribuição espacial (vertical e horizontal) nos respectivos sítios, antes de serem confrontados com o conhecimento previamente existente sobre a sociedade que os originou e com as novas hipóteses dos pesquisadores que os localizaram a fim de que venham a ser posteriormente considerados indicadores potenciais de novos conhecimentos históricos.

De que modo as operações cognitivas referentes à linguagem realizadas pelo historiador são afetadas pelo desenvolvimento da tecnologia digital? A disseminação da computação nas últimas décadas coroa um processo muito mais longo que conecta o desenvolvimento de tecnologias eletroeletrônicas com o de sistemas de codificação que reduzem a linguagem humana a mínimos denominadores comuns, de modo a ampliar o alcance da comunicação em escala global e do processamento de massas com número cada vez maior de dados.

Analisando o contexto de meados do século XIX, Hobsbawm (2012) destaca que mesmo décadas antes da segunda revolução industrial, que generalizou o uso da eletricidade e do petróleo como novas bases da matriz energética mundial, a revolução nos meios de transporte (ferrovias, linhas regulares internacionais de navios a vapor) e comunicação (modernos sistemas de correio, imprensa diária, telégrafo e posteriormente telefone) gerava um grau inédito de integração quase imediata de processos históricos ocorridos nas mais diversas partes do mundo. Considerando-se o papel do código Morse como precursor das linguagens digitais, é significativo que o autor o aponte como indicador do impacto global das proezas tecnológicas do século XIX: “Em 1871, o resultado do Derby era enviado de Londres para Calcutá em nada menos que 5 minutos, apesar de que a notícia era consideravelmente menos excitante que o feito em si” (HOBBSAWM, 2012, p. 74).

Evidentemente essa relação entre o universo linguístico das sociedades humanas e os impulsos elétricos capazes de transmitir comandos e informações processáveis à máquina vem atravessando estágios sucessivamente mais complexos desde então. Como veremos posteriormente, desde a década de 1950, o “processamento de linguagem natural”, ou seja, a interação entre as linguagens dos seres humanos e aquela utilizada na programação dos computadores, se estabeleceu como um campo de desafios estratégicos, associado ao desenvolvimento da inteligência artificial.

O uso do conceito de “linguagem natural” para diferenciação da “linguagem formal”, entretanto, corre o risco de, inadvertidamente, induzir a uma percepção equivocada, à medida que tende a obscurecer o aspecto social da linguagem humana. No início do século XX, diversos pioneiros dos estudos antropológicos levaram o debate sobre o caráter das sociedades humanas a um novo patamar teórico precisamente ao distanciarem sociedade (e linguagem) de qualquer conexão com o “natural”. O distanciamento do determinismo biológico tornou-se cada vez mais um elemento central na compreensão do papel da cultura na história humana, o que pode ser exemplificado pelas formulações de Alfred Kroeber sobre o conceito de “superorgânico” (KROEBER, 1975; ŠKORIĆ, 2016).

Se a linguagem, elemento estruturante da vida social humana, constitui-se de sistemas de signos arbitrários estruturados apenas pela sua diferenciação recíproca, como concluiu Saussure (2007), é importante nos questionarmos sobre a natureza do processo que possibilita transformar “evidências” em “fontes”. Entendemos que essa reflexão pode contribuir tanto para a identificação da contribuição das ciências humanas no desenvolvimento de novas aplicações tecnológicas quanto para a análise dos múltiplos impactos sociais decorrentes da disseminação das tecnologias digitais.

## DA PERCEPÇÃO À NARRATIVA, PASSANDO PELO PENSAMENTO

Haveria alguma relação entre a “classificação racional de informações”, inerente ao trabalho dos historiadores, e o “processamento de linguagem natural (ou social)” que os cientistas da computação vêm desenvolvendo há mais de sessenta anos? Até que ponto a sistematização e explicitação da experiência metodológica e analítica inerente à prática do ofício do historiador pode contribuir para a abertura de caminhos de desenvolvimento de ferramentas de inteligência artificial? De que modo esse processo pode vir a gerar novos impactos sobre o futuro da pesquisa histórica?

Os avanços tecnológicos aplicados às diversas facetas da pesquisa histórica são registrados com frequência cada vez maior. Enquanto redigimos este artigo, deparamo-nos com a notícia de que “a inteligência artificial *DeepMind* derrotou seres humanos na decifração de tabuletas gregas antigas danificadas”. A matéria, para alívio da comunidade historiográfica, contém a reconfortante frase “humanos ainda serão necessários para juntar as peças a olho e então decifrá-las” (LI, 2019).

Podemos especular sobre o que distingue as tarefas em que a inteligência artificial já se mostra capaz de superar os seres humanos no campo da pesquisa histórica e aquelas nas quais “o olhar” do historiador ainda se faz necessário. Essa reflexão nos levará a tratar da relação entre cérebro humano e realidade externa na produção de conhecimento.

Ao analisar o processo de desenvolvimento cognitivo ao longo da infância, Jean Piaget (1973) se viu na necessidade de superar tanto as concepções filosóficas aprioristas, que consideram o conhecimento inerente aos conceitos desenvolvidos no cérebro do indivíduo, quanto as empiristas, segundo as quais as propriedades do real se impõem à mente humana pela simples observação do meio. O psicólogo suíço formulou o conceito de “epistemologia genética” (PIAGET, 1973) para descrever o processo pelo qual, por meio da interação com o meio físico e social, o ser humano desenvolve de forma integrada a consciência de si, da realidade objetiva na qual se insere e dos outros.

Podemos identificar ecos dessa visão, que articula de forma retroalimentar o desenvolvimento cerebral à exploração progressiva da realidade, em trabalhos recentes que se encontram na vanguarda do desenvolvimento de interfaces homem-máquina, alguns deles desenvolvidos por cientistas brasileiros de formação acentuadamente interdisciplinar. Uma matéria jornalística sobre as pesquisas de Miguel Nicolelis, que resultaram no desenvolvimento de um exoesqueleto controlado pelo cérebro de pacientes tetraplégicos, destaca que o neurocientista partiu da observação de que os seres humanos têm “no cérebro um modelo de seu próprio corpo”. Se, de um lado, os sentidos constroem nossa percepção das “fronteiras físicas” entre corpo e

“mundo exterior”, de outro, Nicoletis percebeu que o “esquema corpóreo projetado pelo cérebro” é “capaz de incorporar uma série de ferramentas que os homens usam rotineiramente. Um tenista experiente, por exemplo, pode assimilar sua raquete como uma extensão do próprio corpo; uma violinista, seu violino; e um cirurgião, seu bisturi” (ROSA, 2013).

Foi precisamente a análise da relação entre o cérebro de um profissional especializado e seu instrumento de trabalho que levou Eduardo Miranda a desenvolver a tecnologia que possibilitou a uma violinista com graves lesões cerebrais causadas por um acidente de carro há mais de trinta anos voltar a “tocar”, por meio da combinação entre sensores acoplados à sua cabeça e notas musicais exibidas em uma tela (BBC, 2017).

Qual é a relevância dessas reflexões sobre a relação entre percepção do mundo externo, desenvolvimento cognitivo e interação cérebro-instrumentos para a abordagem do processo de produção do conhecimento histórico?

Conforme vimos acima, Bloch (2001) e Thompson (1981) coincidem tanto na afirmação do caráter “objetivo e determinante” do processo histórico quanto na impossibilidade de obtermos um conhecimento pleno e direto sobre ele. A “interrogação das evidências” em busca de respostas para problemas formulados pela análise da realidade social e das diversas interpretações já existentes sobre os temas abordados se constitui no cerne da prática profissional do historiador. As metáforas artesanais utilizadas por ambos na análise das várias dimensões do processo de construção do conhecimento histórico indicam a interatividade que o caracteriza, marcado, como destaca Thompson, por um diálogo no qual o conhecimento produzido é testado em confronto com a realidade. Essa forma específica de “epistemologia genética” faz do estudo da história um exemplo particularmente intenso de expansão da visão de mundo do ser humano para além do seu ego individual em direção à compreensão da objetividade do mundo externo e da complexa relação com “os outros”. Daí a sua relevância para a formação de cidadãos em uma sociedade plural.

Estabelecendo um paralelo com as pesquisas de Nicoletis (ROSA, 2013) e Miranda (BBC, 2017), poderíamos supor também que a modelagem das ondas cerebrais dos praticantes da investigação histórica passa gradualmente a incorporar seus instrumentos de trabalho. Mas ao invés de uma bola, um violino ou um bisturi, os historiadores manipulam evidências, transformando-as em “informações racionalmente classificadas” e narrativas construídas a partir delas.

Podemos vislumbrar os impactos dessa prática sobre a mente do historiador em um texto clássico do pioneiro da micro-história italiana Carlo Ginzburg. Em *Sinais, raízes de um paradigma indiciário*, ele analisa a emergência, a partir do século XIX, de um “modelo interpretativo centrado sobre os resíduos, sobre os dados marginais considerados reveladores”, que perpassa medicina, crítica de arte, psicanálise e literatura de detetives, entre outros campos do conhecimento. Ginzburg destaca que esse paradigma se ancora no desenvolvimento pela prática investigativa, de uma “intuição alta”, presente aos ofícios de “conhecedor ou diagnosticador”, algo que não se aprende “limitando-se a por em prática regras preexistentes” e que envolve “faro, golpe de vista, intuição” (GINZBURG, 2002, p. 149, 179).

De fato, Bloch já advertira que a análise histórica, para além da aplicação de metodologias e técnicas especializadas, envolve uma sensibilidade peculiar, adquirida na prática do ofício: “o fresador usa instrumentos mecânicos de precisão; o luthier



guia-se, antes de tudo, pela sensibilidade do ouvido e dos dedos. [...] Será possível negar que haja, como o tato das mãos, um das palavras?” (BLOCH, 2001, p. 55).

A “intuição alta” identificada por Ginzburg, entretanto, só se aplica ao processo interpretativo após um amplo e disciplinado processo de coleta e classificação de informações, na linha do que também já fora indicado por Bloch. Se o crítico de arte italiano Morelli ou o detetive Sherlock Holmes, personagem de Sir Arthur Conan Doyle, poderiam identificar a autoria de um quadro ou o parentesco entre uma vítima e uma testemunha pelos detalhes relativos ao formato de orelhas, esse detalhe só poderia se tornar significativo após a identificação de que as orelhas estão entre as partes do corpo que apresentam maiores variações e o estabelecimento de uma metodologia comparativa sobre os diferentes elementos que as constituem. Do mesmo modo, o historiador torna-se cada vez mais capaz de identificar detalhes discrepantes que podem ser pistas significativas em diversos tipos de fontes à medida que a experiência lhe possibilita identificar padrões e tipologias, caso contrário, permanecerá perdido num oceano de detalhes e curiosidades encontrado nas fontes e não realizará sua missão.

Chegarão as máquinas a desenvolver esse tipo de “alta intuição”? Em que tipo de operação cognitiva a inteligência artificial já pode, hoje em dia, auxiliar a pesquisa histórica?

## INTELIGÊNCIA ARTIFICIAL E PROCESSAMENTO DE LINGUAGEM NATURAL

Antes de relacionar a inteligência artificial e a História Digital, primeiramente precisamos compreender como a ciência da computação define inteligência artificial e processamento de linguagem natural e como essas definições evoluíram historicamente. Essa não é uma tarefa trivial. Russel e Norvig (2004) organizaram o quadro abaixo, sintetizando as definições de diversos autores com base em duas dimensões principais: *pensamento/raciocínio*; e *comportamento*.

Quadro 1 – Definições de IA

Sistemas que pensam como seres humanos	Sistemas que pensam racionalmente
“O novo e interessante esforço para fazer os computadores pensarem [...] máquinas com mentes, no sentido total e literal” (HAUGELAND, 1985).	“O estudo das faculdades mentais pelo uso de modelos computacionais” (CHARNIAK; MC DERMOTT, 1985).
“Automatização de atividades que associamos ao pensamento humano, atividades como tomada de decisões, a resolução de problemas, o aprendizado [...]” (BELLMAN, 1978).	“O estudo das computações que tornam possível perceber, raciocinar e agir” (WINSTON; WINSTON, 1992).
Sistemas que atuam como seres humanos	Sistemas que atuam racionalmente
“A arte de criar máquinas que executam funções que exigem inteligência quando executadas por pessoas” (KURZWEIL, 1990).	“A inteligência computacional é o estudo do projeto de agentes inteligentes” (POOLE; MACKWORTH; GOEBEL, 1998).

Fonte: Elaboração dos autores a partir da bibliografia citada.



As abordagens da primeira coluna da tabela medem o sucesso da automação relativamente à sua fidelidade frente ao desempenho humano. Já as da segunda coluna medem o desempenho com relação à racionalidade, definida como a realização de uma tarefa de forma perfeita com as informações disponíveis, sem simular o comportamento ou o modo de pensar característico dos seres humanos. A abordagem centrada nos seres humanos, por outro lado, possui caráter empírico, envolvendo hipóteses e diálogo interdisciplinar, podendo abarcar ciências biomédicas, humanas e exatas. Já as abordagens racionalistas concentram-se na grande área de exatas, em especial nos campos da matemática, computação e engenharias (RUSSEL; NORVIG, 2004).

É interessante analisar as contribuições de diversas disciplinas para o campo da inteligência artificial. Os filósofos tornaram a área concebível, formulando as ideias de que a mente é, em alguns aspectos, semelhante a uma máquina, de que ela opera sobre o conhecimento codificado em alguma linguagem interna e que o pensamento pode ser usado para escolher as ações que deverão ser executadas. Os matemáticos forneceram ferramentas para trabalhar com declarações de certezas lógicas, bem como declarações de incertezas, como as probabilísticas. Definiram a base para a compreensão da computação e do raciocínio sobre algoritmos. Os economistas formalizaram o problema da tomada de decisões que maximizem o resultado esperado. Certas linhas da psicologia adotam a ideia de que os seres humanos podem ser considerados máquinas de processamento de informações. Os linguistas mostraram que o uso da linguagem se ajusta a este modelo. Já os engenheiros de computação forneceram os artefatos que tornam possíveis as aplicações de IA, gerando os grandes avanços de *hardware* que os *softwares* demandam (RUSSEL; NORVIG, 2004). É possível afirmar, portanto, que o campo da IA obteve um vigoroso desenvolvimento a partir de contribuição de várias disciplinas para sua consolidação. No campo da História, entretanto, este processo ainda é incipiente.

A expansão global da internet e a difusão do uso de *Big Data* também contribuíram para grandes saltos no campo de pesquisa denominado Processamento de Linguagem Natural (NLP), que, no entanto, possui uma história bem anterior dividida em quatro períodos: 1950-1970; 1970-1980; 1980-1990; e 1990 até a atualidade. O primeiro período começou com muito entusiasmo e grandes pretensões. O foco então se concentrava na tarefa de tradução automática denominada *Mechanical Translation* (MT). Um dos primeiros trabalhos desse tipo foi a geração de uma amostra de tradução automática de um texto da língua russa para a inglesa. É importante ressaltar que, naquele período, as máquinas eram rudimentares do ponto de vista computacional. O processamento de uma única frase demorava em torno de sete minutos. Um dos pontos de maior relevância dessa fase foi a realização do congresso internacional *The Teddington International Conference on Machine Translation of Languages and Applied Language Analysis*, em 1961, com trabalhos em vários campos da linguística, como sintaxe, semântica, morfologia, interpretação, geração de texto e teoria formal. A área se internacionalizou rapidamente. A União Soviética, os Estados Unidos, países da Europa e Japão logo se tornaram atuantes nesse campo (JONES, 1994).

Já no segundo período, parte da comunidade de NLP se funde à comunidade de IA, com o objetivo de pesquisar representações do conhecimento para manipulação e construção de novos significados. Um dos primeiros trabalhos nessa fase foi a construção de uma base de conhecimento formal para um sistema de perguntas e

respostas, com base em primitivas redes semânticas. Esse trabalho se fundia com a representação do processo de conhecimento por meio da lógica formal. Alguns grandes projetos nesta linha foram desenvolvidos, como foi o caso do *ARPA Speech Understanding Research* (SUR). Foi uma fase que teve como principal característica o caráter prático, com o desenvolvimento de sistemas voltados principalmente para a comunicação com foco no usuário (JONES, 1994).

O terceiro período, por outro lado, foi marcado por uma nova abordagem, denominada de gramaticológica. Com a dificuldade de se avançar na construção de sistemas de perguntas e respostas como os que marcaram a fase anterior, houve um maior desenvolvimento de teorias da gramática entre os linguistas e da representação de conhecimento pelos pesquisadores de IA. Os linguistas elaboraram, por exemplo, a gramática funcional e categórica, tendo como princípio base, em geral, a computabilidade para algoritmos de *parsing* (análise sintática). Nesta mesma época, crescia o paradigma de programação declarativa na computação, com o uso da linguagem de programação *Prolog*. Dessa forma, surgiu a demanda de que o texto fosse tratado por sua sintaxe, possibilitando sua conversão em formas lógicas. Nesta fase, as principais ferramentas desenvolvidas foram *parsers*, dicionários léxicos e gramáticas, com um aumento na oferta de sistemas comerciais relativos a perguntas e respostas associados a bases de dados (JONES, 1994).

O quarto período, finalmente, é marcado pela ênfase na construção de léxicos e abordagens estatísticas. Com o aumento da disponibilidade de textos a partir da expansão da *World Wide Web*, tarefas variadas começam a florescer e ganhar espaço no mercado, tais como: geração de resumos de texto; extração da informação; recuperação da informação através de motores de busca; transcrição automática de áudios; e tradução automática. As abordagens vinculadas à engenharia ganham mais espaço do que a tradicional concepção associada à linguística. Gradualmente, porém, a construção de *corpora* para dar suporte à resolução de diversas tarefas também começou a se difundir. Congressos com *shared task* (tarefas específicas a serem resolvidas pela comunidade de participantes) ganham espaço e ajudam na evolução do campo. Métodos estatísticos e de Aprendizado de Máquina também se expandem e se integram à NLP. Na década atual, tarefas como: tradução; identificação de classes gramaticais; reconhecimento de entidades nomeadas; e análise de sentimentos já estão bastante difundidas, sendo inclusive incorporadas em alguns serviços *online*. Apesar dos grandes avanços ao longo dessas várias décadas, muitos problemas ainda restam irresolutos.

## CIÊNCIA DA COMPUTAÇÃO E HISTÓRIA DIGITAL

Nesta seção apresentamos alguns trabalhos acadêmicos considerados relevantes para a aplicação da ciência da computação à área da História Digital. Dividimos os trabalhos pelos temas de pesquisa aos quais eles se relacionam na área de processamento de linguagem natural e inteligência artificial, tentando indicar, em cada caso, o seu potencial agregador para a pesquisa histórica. Os exemplos selecionados dizem respeito aos seguintes temas: 1) identificação de autoria, 2) modelagem de tópicos e 3) extração da informação.



## Identificação de autoria

A atribuição de autoria é um problema que vem sendo estudado ao longo de várias décadas. No presente, essa abordagem se relaciona fortemente às ferramentas de combate ao plágio, o qual ocorre em várias áreas: jornalismo, artes, música, atividade acadêmica, etc. Entretanto, a identificação de autoria por estilo pode ser utilizada até mesmo para a resolução de crimes, tal como na investigação mais longa e mais cara da história do FBI, o caso do terrorista Theodore John Kaczynski, mais conhecido como *Unabomber*. Kaczynski escreveu um manifesto marcado pela crítica social e pelo posicionamento anti-industrialização, e o mesmo estilo de escrita foi identificado em uma carta enviada a seu irmão, que o denunciou. As semelhanças estilísticas analisadas com métodos computacionais contribuíram para confirmar a autoria dos crimes cometidos pelo *Unabomber*.

No campo da história, a aplicação de métodos de identificação de autoria pode acarretar novas hipóteses acerca de fatos já bastante estudados, o que pode ser de grande relevância. Um exemplo clássico é o problema de Identificação Autoria no texto dos *Documentos Federalistas*, escritos entre 1787-1788 por Alexander Hamilton, John Jay e James Madison, com o objetivo de convencer os cidadãos do Estado de Nova York a ratificar a constituição dos EUA. Esses 77 ensaios curtos, contendo de 900 a 3500 palavras, apareceram em jornais, assinados com pseudônimo *Publius*. Em 1778, eles foram juntados a oito artigos adicionais sobre o mesmo assunto e publicados em forma de livro. Hamilton morreu em um duelo em 1804, mas deixou uma lista em que apontava a autoria individual de cada artigo, publicada três anos depois em um periódico da Filadélfia. Entretanto, Madison, após encerrar seu mandato como o quarto presidente dos Estados Unidos, reabriu o debate sobre essa atribuição de autoria. Em 1818, ele afirmou que escrevera um conjunto de artigos que Hamilton havia atribuído a si próprio. Desde então, foi se estabelecendo entre os críticos especializados o consenso em relação a alguns pontos: John Jay é o único autor de cinco dos 85 artigos (os de número 2, 3, 4, 5 e 64); Hamilton é o único autor de 43 artigos; Madison é o único autor de 14. A atribuição de autoria dos 12 artigos restantes, os chamados “documentos contestados” (49-58, 62-63), entretanto, permaneceu polêmica após o falecimento dos autoproclamados autores, até porque os estilos de escrita de ambos eram bastante semelhantes (HOLMES; FORSYTH, 1995). Em 1941, Frederick Williams e Frederick Mosteller contabilizaram o tamanho médio das frases de ambos nos artigos não contestados, chegando a médias de 34,55 e 34,59 palavras por frase respectivamente para Hamilton e Madison, com desvios padrão de 19,2 e 20,3. Desse modo, não era possível chegar a conclusões taxativas sobre a disputa de autoria.

Até 1964, quando Mosteller e Wallace publicaram a primeira edição de seu livro, *Inference and Authorship disputed* as opiniões permaneceram divididas, com cada lado esgrimindo diversos argumentos históricos e estilísticos. O caminho adotado pelos autores para chegar a conclusões mais robustas passou pela análise das *marker words*, palavras utilizadas com frequências significativamente distintas por Hamilton e Madison. *While* e *whilst*, por exemplo, ocorriam com frequências bem distintas entre os dois autores, enquanto *on* e *upon*, eram usadas por ambos com frequência semelhante. Mosteller e Wallace utilizaram-se do Teorema de Bayes, que descreve a probabilidade de um evento ocorrer a partir de um conhecimento prévio, e fizeram inferências com 30 *marker words*. Para isto, partiram dos textos não contestados de

ambos: 94.000 palavras de Hamilton e 114.000 palavras de Madison. Adicionalmente, para encontrar o conjunto final de *marker words* a serem utilizados na inferência, fizeram uma pré-seleção das palavras com maior potencial discriminativo entre os autores em disputa. Ao final do estudo, os autores concluíram que os doze artigos em disputas pertenciam a Madison e com isso estabeleceram uma metodologia que possibilitou avanços posteriores na aplicação a diversos outros casos semelhante (HOLMES; FORSYTH, 1995).

## Modelagem de tópicos

A modelagem de tópicos consiste em encontrar tópicos abstratos, dentro de uma coleção de documentos previamente selecionada, gerando a identificação de grupos de documentos. Essa metodologia é muito útil quando se deseja encontrar estruturas semânticas latentes, que o observador humano teria dificuldade de identificar devido à vastidão da documentação. Trata-se, portanto, de um problema de natureza combinatória: localizar um subconjunto de palavras que representa um subconjunto de documentos.

Nas duas últimas décadas foram publicados diversos trabalhos acadêmicos relevantes a respeito da modelagem de tópicos. Griffiths e Steyvers (2004) aplicaram a técnica a resumos do *Proceedings of the National Academy of Sciences of the United States of America*, um periódico multidisciplinar, para identificar tópicos em ascensão e queda entre o período de 1991-2011. Em outra experiência bastante frutífera, Nelson desenvolveu uma plataforma de análise das mudanças nos assuntos tratados pelo jornal *Richmond Times – Dispatch* durante a Guerra Civil Americana, com o objetivo de entender as mudanças sociais e políticas em Richmond no período (NELSON, 2019). Já Yin *et al.* (2011) utilizam a modelagem de tópicos para analisar a distribuição geográfica de conjuntos documentais por regiões.

Um caso que nos parece digno de uma análise mais detalhada é o de um relatório técnico (ALLEN *et al.*, 2014) produzido por membros no *History Lab* em parceria com a Universidade Columbia. Justificando a relevância do projeto, os autores comentam que, em 2013, funcionários do governo americano e empresas privadas decidiram classificar informações como sigilosas mais de 80 milhões de vezes, um aumento de quase quatro vezes em relação aos 23 milhões de sigilos impostos cinco anos antes (ISOO, 2014, p. 1). No mesmo período, o *Public Interest Declassification Board* da presidência americana estimava que apenas uma das agências de inteligência do país produzia um *petabyte* de dados anualmente, e que seriam necessários dois milhões de arquivistas trabalhando em período integral para revisar a desclassificação desses documentos (PIDB, 2012, p. 17). Mas, ao contrário de ampliar os investimentos nesse trabalho, na verdade o governo estava gastando menos da metade do dinheiro em desclassificação em relação à quinze anos antes. Nesse contexto de insuficiência de recursos arquivísticos, registra-se a destruição de 95 a 97% dos documentos do Departamento de Estado. Isso inclui, por exemplo, todos seus telegramas diplomáticos relacionados à pesquisa científica patrocinada pelo governo que não possuam referência cruzada com outro assunto considerado de maior significado histórico (NEWMAN; BLOCK, 2006).

A equipe do projeto do *History Lab* destaca que o crescimento do sigilo oficial preocupa humanistas, cientistas e cidadãos. Historiadores como Peter Galison

argumentam que enquanto a ciência visa buscar e proteger informações, a classificação sigilosa age na contramão, tornando impossível conhecer as informações mantidas fora do alcance do público. Avaliações oficiais já constataram reiteradamente que o vasto e esmagadoramente complexo sistema criado para guardar segredos, na verdade, torna mais difícil identificar e proteger informações realmente valiosas (LANGBART; FISCHER; ROBERSON, 2007). Os grupos de monitoramento privados encontraram muitos exemplos paradoxais, como a existência de informações técnicas de programas de armas químicas e biológicas nas estantes abertas à pesquisa pública do *National Archives and Records Administration* (ALLEN *et al.*, 2014, p. 21). Mas a regra geral é que cada vez menos documentos são desclassificados no momento correto, conforme definido pela legislação.

Essas reflexões realçam a relevância dos trabalhos de projetos como o *Topic Modeling Official Secrecy* (ALLEN *et al.*, 2014). Os autores, a partir de uma coleção de 1.1 milhão de telegramas desclassificados pelo Departamento de Estado americano entre os anos de 1973-1976, conseguiram construir um modelo que verifica se um dado telegrama atualmente aberto à consulta foi originalmente classificado como sigiloso. Adicionalmente, pela aplicação da modelagem de tópicos a partir da técnica *Latent Dirichlet Allocation* (LDA), conseguiram identificar tópicos mais sensíveis e, portanto, com maior probabilidade de permanecerem sigilosos por várias décadas.

Em um primeiro estudo, os pesquisadores analisaram uma coleção de documentos desclassificados de relações internacionais dos Estados Unidos da América (EUA), apontados por historiadores especializados como os mais representativos de cada época em relação a determinados temas. A cronologia foi subdividida nos períodos 1952-1960 (a era de Eisenhower) e 1961-1968 (as eras de Kennedy e Johnson). A partir da aplicação do LDA em cada grupo de documentos, os autores identificaram 20 tópicos para cada um dos dois grupos. Para o período de Eisenhower, descobriram que os documentos com maior chance de serem mantidos sob sigilo eram os relacionados aos termos: “óleo”, “dia”, “homem”, “vezes”, “empresas”, “Arábia”, “construção”. Analisando os documentos associados a esses tópicos, identificaram que eles dizem respeito, de um lado, às empresas de petróleo americanas diante do surgimento da Organização dos Países Exportadores de Petróleo (OPEP) e, de outro, às operações da Agência de Inteligência Norte-americana (CIA) tais como as relativas à derrubada do governo democrático da Guatemala em 1954, que incluíam a sabotagem aos suprimentos de petróleo do país centro-americano. Comparando os documentos associados àquele primeiro grupo de termos a outros vinculados aos tópicos: “estrangeiro”, “troca”, “banco”, “departamento”, “exportação”, “mercado” e “garantia”, verifica-se que a probabilidade destes últimos serem mantidos sob sigilo é 24 vezes inferior.

O mesmo experimento foi aplicado para as eras de Kennedy e Johnson. Os tópicos com maiores chances de serem classificados naquele período foram: “origem”, “armas”, “área”, “missão”, “informação”, “oficiais” e “base”. A análise dos documentos indicou que eles estão associados, por exemplo, ao bombardeio de Laos, um país neutro na guerra do Vietnã. A presença desses tópicos aumenta a probabilidade de um documento ser mantido sob sigilo em 13 vezes em relação aos documentos mais fortemente associados a: “ajuda”, “milhões”, “assistência”, “econômico”, “países”, “Japão” e “estrangeiros”. Isto mostra como os documentos relativos à ajuda externa e ao comércio são consistentemente menos propensos a sofrer censura de trechos sensíveis.

Essa breve análise de alguns exemplos significativos demonstra como, apesar de ser ainda um trabalho incipiente e exploratório, há um grande potencial de técnicas de modelagem de tópicos para tarefas desta natureza. Como o processo manual de desclassificação de documentos custa bilhões de dólares ao governo americano, a utilização de ferramentas como estas pode reduzir os custos e viabilizar que, mesmo diante de um aumento exponencial da massa documental, seja possível assegurar condições para o seu estudo por parte de pesquisadores profissionais, assim como de cidadãos democraticamente engajados.

## Extração da informação

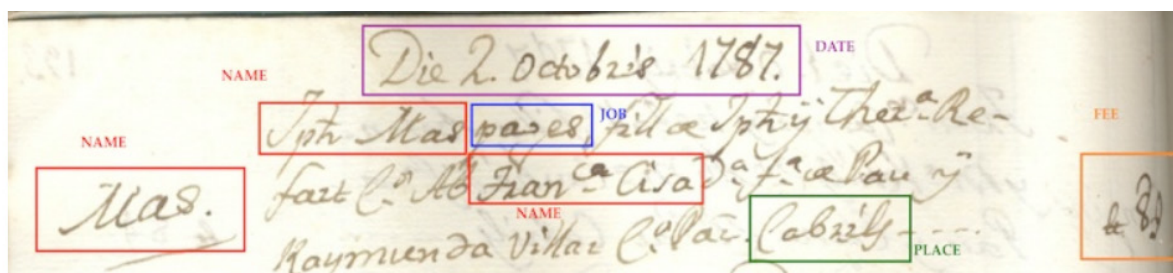
O termo extração de informação (EI) aplica-se à atividade de extrair automaticamente informações pré-especificadas em textos na linguagem natural. Por exemplo, em acervos digitalizados de jornais de negócios que contenham anúncios de “eventos de sucessão de gerenciamento” (aposentadorias, nomeações, promoções, etc.), é possível, por meio dessa técnica, extrair os nomes dos participantes (empresas e indivíduos), o cargo envolvido em cada um desses eventos, o motivo da abertura vaga e assim por diante. A EI também pode ser utilizada para preencher bancos de dados a partir de fontes de informação estruturadas, de fontes de informação não estruturadas ou até mesmo de textos livres. Esse banco de dados é então utilizado para pesquisas e análises convencionais ou submetido a técnicas de mineração de dados, visando gerar resumos ou índices dos textos de origem (GAIZAUSKAS; WILKS, 1998).

A maioria das fontes valiosas para a pesquisa histórica, mesmo aquelas já digitalizadas, jamais passou por qualquer processamento que permita a recuperação da informação e o cruzamento dos dados nelas contidos. A *International Conference on Document Analysis and Recognition* (ICDAR) tem como objetivo incentivar o avanço de sistemas capazes de extrair informações relevantes de documentos históricos. O evento inclui uma série de competições relativas à atividades específicas, como por exemplo: Reconhecimento de *Layout* de Documento Histórico Manuscrito; Análise de *Scripts* Manuscritos; etc. Um trabalho recente, cujos autores participaram da competição denominada *Information Extraction in Historical Handwritten Records* (IEHHR), aplicou a extração da informação em 125 páginas contendo um total de 1221 registros manuscritos de casamento (39.527 imagens de palavras) dos arquivos da Catedral de Barcelona entre o período de 1617 e 1619. Conforme ilustrado na figura 1, esses registros contêm informações como nomes de pessoas, parentescos, lugares, cargos, entre outras.

A tarefa realizada na competição consistiu em encontrar palavras relevantes em registros de casamentos, transcrevê-las, rotulá-las com uma categoria semântica e definir a que pessoas elas se relacionam. Palavras relevantes poderiam pertencer a cinco categorias: “Nome”, “Sobrenome”, “Ocupação”, “Localização” e “Estado Civil”. Essas categorias semânticas estão associadas a sete tipos de relações: “Esposa”, “Marido”, “Pai da Esposa”, “Mãe da Esposa”, “Pai do Marido”, “Mãe do Marido” e “Outra Pessoa”. A competição apresenta duas possibilidades em relação às tarefas a serem resolvidas: na básica, o sistema tem que prover a transcrição e a categoria semântica (por exemplo: sobrenome, localização, etc.); na completa, o sistema deve inferir também as relações entre pessoas (por exemplo: entre marido e esposa). Em

ambos os casos (básica e completa), os autores chegaram a taxas de acerto de cerca de 94% (TOLEDO; CARBONELL; FORNÉS; LLADÓS, 2019).

Figura 1 – Exemplo de registro anotado da base Esposalles



Fonte: Esposalles Database<sup>2</sup>

O leitor pode se questionar sobre a utilidade de se classificar imagens em categorias previamente definidas. Ocorre que, uma vez que os modelos são aprendidos nesse conjunto inicial de imagens anotadas, é possível aplicá-los em novas imagens, mesmo sem realizar Reconhecimento Óptico de Caracteres (OCR), e gerar novas informações classificadas nas categorias utilizadas nas anotações. Esse é um exemplo de um estudo que abre espaço para a construção de sistemas de recuperação da informação para historiadores e pesquisadores de áreas afins que trabalham com fontes que, com as tecnologias até recentemente disponíveis, permaneciam de difícil processamento.

## CONSIDERAÇÕES FINAIS

O avanço das tecnologias digitais, mais do que mudanças no cotidiano de trabalho do historiador ou inovações potenciais nas técnicas de análises de um volume crescente de documentação digitalizada, traz um grande potencial de transformações na natureza e alcance das pesquisas em História e na relação entre os profissionais da área e a sociedade como um todo.

Um exemplo de iniciativa visando à ampliação do escopo das análises e das conclusões possíveis a partir do desenvolvimento de projetos coletivos inovadores é o *Collaborative for Historical Information and Analysis* (CHIA), coordenado por Patrick Manning, consagrado africanista e liderança no campo da História Mundial. Inspirado pelos resultados gerados pelo uso de *Big Data* nas ciências naturais, em campos tão distintos como estudos climáticos, astronomia e biologia, o CHIA consiste em um repositório por meio do qual historiadores de diversas partes do mundo podem compartilhar bases de dados, particularmente aquelas referentes ao período anterior a 1950, em relação ao qual as séries estatísticas disponíveis são extremamente desiguais e descontínuas. A metodologia proposta busca equilibrar duas tarefas complementares: construção da colaboração entre os pesquisadores e construção da tecnologia necessária ao processamento de um volume crescente de massas de dados (MANNING, 2013, p. 1-2). Trata-se de uma experiência de aplicação dos princípios da

<sup>2</sup> Disponível em <http://dag.cvc.uab.es/the-esposalles-database/>. Acesso em 20 out. 2019.



*Open Science*<sup>3</sup> ao campo da História, que demandará mudanças profundas na cultura do ofício para que o seu potencial possa ser plenamente desenvolvido.

Outro âmbito no qual a História Digital pode gerar fortes impactos é o da História Pública, potencializando a transformação da relação entre historiadores profissionais, cidadãos envolvidos em práticas diversas de estudo e preservação do passado e o público em geral. A trajetória do falecido historiador norte-americano Roy Rosenzweig é bastante ilustrativa a esse respeito. Integrante do coletivo de historiadores do trabalho que escreveu, nos anos 1980, a pioneira obra de referência *Who Built America?*, uma revisão abrangente das narrativas tradicionais sobre a história norte-americana com ênfase no papel dos trabalhadores (LEVINE, 1989), Rosenzweig já demonstrava seu interesse na exploração dos potenciais das novas tecnologias para a difusão do conhecimento histórico ao criar, em 1994, uma versão em CD-ROM da mesma obra (ROSENZWEIG *et al.*, 1994). Quatro anos mais tarde, ao coeditar um livro que registrava um amplo painel de “usos populares da história” nos EUA, escreveu um posfácio sintomaticamente intitulado *Everyone a Historian*,<sup>4</sup> no qual defendia que “os profissionais de História precisam trabalhar mais pesado para aprender a escutar e respeitar as muitas formas em que os contadores de história populares percorrem o terreno do passado que é tão presente para todos nós” (ROSENZWEIG; THELEN, 1998, p. 358).

Em 2006, ao coautorar com Daniel Cohen o primeiro trabalho de síntese sobre os potenciais e os perigos da História Digital, Rosenzweig (2006) conclamava os colegas a concentrar esforços para “manter e alargar a espantosamente rica teia de história pública que emergiu na última década”, por meio da adesão à Iniciativa de Budapeste pelo Acesso Aberto e aos esforços de base para “colocar o passado online”, desde a disponibilização de uns poucos documentos até “projetos mais ambiciosos de criação de arquivos públicos abertos”. Concluía enfatizando que “a mais importante arma para construir o futuro digital que queremos é sermos ativos na criação da História Digital no presente” (COHEN; ROSENZWEIG, 2006, p. 13).

Esperamos que este artigo possa oferecer uma contribuição nesse sentido, ao sistematizar reflexões teórico-metodológicas geradas no debate interdisciplinar entre História e Ciência da Computação com o objetivo de estimular a ampliação da exploração do vasto, embora relativamente recente, terreno da História Digital. Conforme buscamos indicar ao longo de todo o texto, torna-se cada vez mais necessário analisar as transformações vividas pelo ofício do historiador no contexto atual e os desafios que elas colocam em relação ao próprio processo de formação dos futuros pesquisadores da área.

## REFERÊNCIAS

ALLEN, David *et al.* Topic Modeling Official Secrecy. *History Lab.*, 2014. Disponível em: [http://history-lab.org/images/presentations/Topic\\_Modeling\\_OS.pdf](http://history-lab.org/images/presentations/Topic_Modeling_OS.pdf). Acesso em: 25 out. 2019.

<sup>3</sup> Para maiores informações sobre o movimento *Open Science*, ver: <https://cos.io>.

<sup>4</sup> O posfácio está disponível em: <http://chnm.gmu.edu/survey/afterroy.html>.



BBC. Violinista que sofreu acidente há 30 anos volta a tocar com tecnologia criada por brasileiro. *BBC*, 03 set. 2017. News Brasil. Disponível em: <https://www.bbc.com/portuguese/geral-41032970>. Acesso em: 25 out. 2019.

BELLMAN, Richard. *An Introduction to Artificial Intelligence: Can Computers Think?*. San Francisco: Boyd & Fraser Pub. Co, 1978.

BLOCH, Marc. *Apologia da história ou o ofício do historiador*. Rio de Janeiro: Jorge Zahar, 2001.

BRUCE, Levine C. *Who Built America? Working People and the Nation's Economy, Politics, Culture, and Society*. American Social History Project. 1. ed. New York: Pantheon Books, 1989.

BOYD, Albert Douglas; LARSON, Mary. *Oral History and Digital Humanities: Voice, Access, and Engagement*. New York: Palgrave Macmillan, 2014.

CHARNIAK, Eugene; MCDERMOTT, Drew V. *Introduction to Artificial Intelligence*. Reading: Addison-Wesley, 1985.

COHEN, Daniel J.; ROSENZWEIG, Roy. *Digital History: A Guide to Gathering, Preserving, and Presenting the Past on the Web*. Philadelphia: University of Pennsylvania Press, 2006.

GAIZAUSKAS, Robert; WILKS, Yorick. Information Extraction: Beyond Document Retrieval, *International Journal of Computational Linguistics & Chinese Language Processing*, Hsinchu, v. 3, n. 2, p. 17-60, Aug. 1998. Disponível em: <https://www.aclweb.org/anthology/O98-4002.pdf>. Acesso em: 25 out. 2019.

GINZBURG, Carlo. *Mitos, emblemas, sinais morfologia e história*. 2. ed. São Paulo: Cia das Letras, 2002.

GRIFFITHS, Thomas L.; STEYVERS, Mark. Finding Scientific Topics. *Proceedings of the National Academy of Sciences of the United States of America*, v. 101, s. 1, p. 5228-5235, Apr. 2004. Disponível em: [https://www.pnas.org/content/pnas/101/suppl\\_1/5228.full.pdf](https://www.pnas.org/content/pnas/101/suppl_1/5228.full.pdf). Acesso em: 25 out. 2019.

GULDI, Jo; ARMITAGE, David. *The history Manifesto*. Cambridge: Cambridge University Press, 2014.

HAUGELAND, John. *Artificial Intelligence: The Very Idea*. Cambridge: MIT Press, 1985.

HOBBSAWM, Eric John Ernest. *A Era do Capital: 1848-1875*. São Paulo: Paz e Terra, 2012.

HOLMES, David I.; FORSYTH, Richard S. The Federalist Revisited: New Directions in Authorship Attribution. *Literary and Linguistic Computing*, Oxford, v. 10, n. 2, p. 111-127, Jan. 1995. Disponível em: <https://doi.org/10.1093/lc/10.2.111>. Acesso em: 25 out. 2019.



INFORMATION SECURITY OVERSIGHT OFFICE. *2013 Report to the President*. Washington: National Archives, 2014. Disponível em: <https://fas.org/sgp/isoo/2013rpt.pdf>. Acesso em: 25 out. 2019.

JONES, Karen Sparck. Natural Language Processing: A Historical Review. In: ZAMPOLLI, Antonio; CALZOLARI, Nicoletta; PALMER, Martha (ed.). *Current Issues in Computational Linguistics: In Honour of Dom Walker*. New York: Springer, 1994. p. 3-16. Disponível em: [https://doi.org/10.1007/978-0-585-35958-8\\_1](https://doi.org/10.1007/978-0-585-35958-8_1). Acesso em: 25 out. 2019.

KELLY, T. Mills. *Teaching History in the Digital Age*. Ann Arbor: University of Michigan Press, 2013.

KROEBER, Albert Louis. Lo Superorgánico. In: KAHN, Joan S. (org.). *El Concepto de Cultura: Textos Fundamentales*. Barcelona: Anagrama, 1975. p. 47-83.

KURZWEIL, Ray. *The Age of Intelligent Machines*. Cambridge: MIT Press, 1990.

LANGBART, David; FISCHER, William; ROBERSON, Lisa. Appraisal of Records Covered by N1-59-07-3-P. Washington DC: National Archives and Records Administration, 2007.

LI, Gege. DeepMind AI Beats Humans at Deciphering Damaged Ancient Greek Tablets. *New Scientist*. 18 Oct. 2019. Technology. Disponível em: <https://www.newscientist.com/article/2220438-deepmind-ai-beats-humans-at-deciphering-damaged-ancient-greek-tablets/#ixzz62pl5EdOq>. Acesso em: 25 out. 2019.

MANNING, Patrick. *Big Data in History*. New York: Palgrave Macmillan, 2013.

NELSON, Robert K. Introduction. In: NELSON, Robert K. *Mining the Dispatch*. Richmond: Digital Scholarship Lab, 2019. Disponível em: <https://dsl.richmond.edu/dispatch/pages/intro>. Acesso em: 25 out. 2019.

NEWMAN, David J.; BLOCK, Sharon. Probabilistic Topic Decomposition of an Eighteenth-Century American Newspaper. *Journal of the American Society for Information Science and Technology*, Chapel Hill, v. 57, n. 6, p. 753-767, Feb. 2006. Disponível em: <https://doi.org/10.1002/asi.20342>. Acesso em: 25 out. 2019.

PIAGET, Jean. *Epistemologia genética*. 2. ed. Petropolis: Vozes, 1973.

POOLE, David L.; MACKWORTH, Alan K.; GOEBEL, Randy. *Computational Intelligence: A Logical Approach*. New York: Oxford University Press, 1998.

PUBLIC INTEREST DECLASSIFICATION BOARD. *Transforming the Security Classification System*. Washington: National Archives and Records Administration, 2012. Disponível em: <https://www.archives.gov/files/declassification/pidb/recommendations/transforming-classification.pdf>. Acesso em: 25 out. 2019.



ROSA, Guilherme. Pesquisa de Nicolelis mostra como o cérebro integra objetos externos ao corpo. *VEJA*. 23 ago. 2013. Ciência. Disponível em: <https://veja.abril.com.br/ciencia/pesquisa-de-nicolelis-mostra-como-o-cerebro-integra-objetos-externos-ao-corpo>. Acesso em: 25 out. 2019.

ROSENZWEIG, Roy et al. *Who Built America? From the Centennial Celebration of 1876 to the Great War of 1914*. American Social History Project. New York: Voyager, 1994. Macintosh Version.

ROSENZWEIG, Roy; THELEN, David P. *The Presence of the Past: Popular Uses of History in American Life*. New York: Columbia University Press, 1998.

RUSSEL, Stuart; NORVIG, Peter. *Inteligência Artificial*. Rio de Janeiro: Campus, 2004.

SAUSSURE, Ferdinand de. *Curso de linguística geral*. 28. ed. São Paulo: Cultrix, 2007.

ŠKORIĆ, Marko. Alfred Kroeber and the concept of the superorganic. *Issues in Ethnology and Anthropology*, Belgrade, v. 11, n. 1, p. 85-111, Apr. 2016. Disponível em: <https://doi.org/10.21301/EAP.V11I1.4>. Acesso em: 25 out. 2019.

THOMPSON, E. P. *A miséria da teoria ou um planetário de erros: uma crítica ao pensamento de Althusser*. Rio de Janeiro: Zahar, 1981.

TOLEDO, J. Ignacio et al. Information Extraction from Historical Handwritten Document Images with a Context-aware Neural Model. *Pattern Recognition*, York, v. 86, p. 27-36, Aug. 2018.

WINSTON, Patrick Henry. *Artificial Intelligence: Instructor's Manual*. 3. ed. Reading: Addison-Wesley, 1992.

YIN, Zhijun et al. Geographical Topic Discovery and Comparison. In: CONFERENCE ON WORLD WIDE WEB, 20, 2011, New York. Anais eletrônicos [...]. New York: ACM, 2011, p. 247-256. Disponível em: <https://doi.org/10.1145/1963405.1963443>. Acesso em: 25 out. 2019

## NOTAS

---

### AUTORIA

**Alexandre Fortes:** Doutor. Professor Associado, Universidade Federal Rural do Rio de Janeiro, Instituto Multidisciplinar, Departamento de História, Nova Iguaçu, RJ, Brasil.

**Leandro Guimarães Marques Alvim:** Doutor. Professor Adjunto, Universidade Federal Rural do Rio de Janeiro, Instituto Multidisciplinar, Departamento de Ciência da Computação, Nova Iguaçu, RJ, Brasil.

### ENDEREÇO PARA CORRESPONDÊNCIA

Alexandre Fortes. Rua General Glicério, 445, ap. 1204, 22245-120, Rio de Janeiro, RJ, Brasil.



## ORIGEM DO ARTIGO

Artigo original elaborado a partir da colaboração acadêmica entre os dois coautores no âmbito do Mestrado em Humanidades Digitais da UFRRJ.

## AGRADECIMENTOS

Agradecemos a todos os colegas e discentes com os quais as ideias que levaram à elaboração desse artigo foram discutidas ao longo do segundo semestre de 2019.

## FINANCIAMENTO

As atividades acadêmicas a partir das quais o artigo foi elaborado foram financiadas pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), por meio do custeio ao funcionamento do Programa de Pós-Graduação em Humanidades Digitais da UFRRJ, pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), por meio de Bolsa de Produtividade em Pesquisa, e pela Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ), por meio da bolsa Cientista do Nosso Estado.

## CONSENTIMENTO DE USO DE IMAGEM

Não se aplica.

## APROVAÇÃO DE COMITÊ DE ÉTICA EM PESQUISA

Não se aplica.

## CONFLITO DE INTERESSES

Não houve conflito de interesses.

## LICENÇA DE USO

Este artigo está licenciado sob a [Licença Creative Commons CC-BY](#). Com essa licença você pode compartilhar, adaptar, criar para qualquer fim, desde que atribua a autoria da obra.

## PUBLISHER

Universidade Federal de Santa Catarina. Programa de Pós-Graduação em História. Portal de Periódicos UFSC. As ideias expressadas neste artigo são de responsabilidade de seus autores, não representando, necessariamente, a opinião dos editores ou da universidade.

## EDITORES

Flávia Florentino Varella (editora-chefe)  
Rodrigo Bragio Bonaldo

## HISTÓRICO

Recebido em: 29 de outubro de 2019  
Aprovado em: 26 de março de 2020

Como citar: FORTES, Alexandre; ALVIM, Leandro Guimarães Marques. Evidências, códigos e classificações: o ofício do historiador e o mundo digital. *Esboços*, Florianópolis, v. 27, n. 45, p. 207-227, maio/ago. 2020.

