Alfonso-Goldfarb, Ana Maria; Goldfarb, José Luiz; Ferraz, Márcia Helena Mendes; Souza, Odécio. Bases conceituais e tecnológicas para a modelagem de megadados em Humanidades Digitais: um estudo de caso em *corpora* textuais de história da ciência. *TECCOGS* – Revista Digital de Tecnologias Cognitivas, n. 21, jan./jun. 2020, p. 25-43.

Recebido em: 15 abr. 2020
Aprovado em: 30 abr. 2020

dx.doi.org/10.23925/1984-3585.2020i21p25-43

Bases conceituais e tecnológicas para a modelagem de megadados em Humanidades Digitais:

um estudo de caso em *corpora* textuais de história da ciência

Ana Maria Alfonso-Goldfarb[1]

José Luiz Goldfarb[2]

Márcia Helena Mendes Ferraz[3]

Odécio Souza[4]

Resumo: O intenso trabalho de digitalização realizado desde os anos 90 disponibilizou vastíssimos *corpora* textuais, inclusive para acesso remoto. Mais recentemente, novos recursos informáticos vêm fornecendo possibilidades inéditas de estudos teóricos, contemplando, também, a modelagem de megadados para organização e classificação dos conhecimentos. O presente estudo de caso se inscreve num marco maior de cooperação internacional, destinado a tornar mais efetiva a pesquisa e o trabalho documental em história da ciência. Especificamente, seu objetivo é elaborar ferramentas que permitam a localização e reconhecimento de conceitos comuns a grupos de textos pertences a grandes bases de dados em e para a história da ciência, assim como a sua mudança em função do tempo, tendo em vista sua indexação e classificação.

Palavras-chave: História da ciência. Humanidades Digitais. Organização e classificação do conhecimento. *Corpora* textuais. Modelagem e mineração de dados. Linguística computacional.

1 cv Lattes: lattes.cnpq.br/7427854657719431.

2 cv Lattes: lattes.cnpq.br/1023793876897710. E-mail: jlgoldfarb@dialdata.com.br.

3 cv Lattes: lattes.cnpq.br/5677877981039661.

4 cv Lattes: lattes.cnpq.br/4189566610252580. E-mail: odeciosouza@gmail.com.

Conceptual and technological grounds for big data modelling in digital humanities: textual corpora in history of science as case study

Abstract: Extensive digitization projects conducted since the 1990s made large textual corpora easily available, even to remote access. More recently new computing tools began to be developed that afford novel approaches to theoretical studies and big data modeling likely to allow for more accurate standards for the organization and classification of knowledge. The present case study is integrated within a larger international collaboration aiming at making work with documents relevant for history of science more effective. More in particular, the aim of the present case study is to develop tools to locate and recognize concepts common to texts included in large databases specific for history of science research, as well as their change over time, for the purpose of indexing and classification.

Keywords: History of science. Digital humanities. Organization and classification of knowledge. Textual corpora. Data mining and modeling. Computational linguistics.

Qualificação do problema

Como se sabe, na última década e de modo crescente, a tecnologia digital vem alterando a forma de planejar e executar o trabalho feito por pesquisadores. De fato, conforme Berry (2012), cada vez mais, a atividade de pesquisa é mediada pela tecnologia digital, a tal ponto que o próprio conceito de "pesquisa" vem mudando e, por consequência, afetando as epistemologias e ontologias que subjazem aos programas de pesquisa.

No caso das Ciências Humanas e Sociais – de particular interesse para nosso caso –, originalmente, desenvolveu-se um campo conhecido como "computação nas humanidades" (*humanities computing*). Este, porém, consistia na aplicação, pura e simples, de técnicas de computação aos próprios materiais e objetos gerados ou utilizados pelas Humanidades, visando reproduzir digitalmente grandes projetos editoriais. Todavia, não demorou muito para que as possibilidades geradas pelas novas tecnologias ultrapassassem essa perspectiva (Berry, 2012; Caraco, 2012; Hayles, 2012; McCarty, 2005; Svensson, 2009, 2010), dando início ao que hoje se conhece como "Humanidades Digitais" (HD). Como esperado, tal mudança implicou uma profunda transformação conceitual, resultando num novo campo do saber. Vale lembrar que a escolha do termo HD aponta muito mais para similaridades entre metodologias do que entre objetos, textos ou simples tecnologias (Kirschenbaum, 2010). Assim, o termo é geralmente utilizado como uma espécie de guarda-chuva que abriga "um vasto conjunto de práticas para criar, aplicar, interpretar, interrogar e hackear tecnologias de informação, tanto as mais antigas quanto as mais novas" (Presner, 2010). Caberia, portanto, perguntar se, de fato, estamos tratando com uma única, ou com várias linhagens de HD (Fitzpatrick, 2011).

Nesse sentido, desde o início e conforme já indicado, foram percebidas diferenças na aplicabilidade das tecnologias digitais aos diversos campos do saber. Nada semelhante ao sucesso quase imediato e espetacular nas ciências exatas e naturais se deu entre as ciências humanas e sociais. Muito embora, no caso das pesquisas em ciências sociais a assimilação tenha sido mais fácil e breve do que entre as ciências humanas,

propriamente ditas, e de forma especial nas pesquisas em História (Baur, 2009; Weldon, 2015; Aronova *et al.*, 2017). De acordo com a socióloga N. Baur (2009), os motivos para essas dificuldades seriam vários: 1) diferente dos sociólogos, os historiadores tradicionalmente têm uma percepção aguda de que a interpretação de seus dados depende, em boa parte, da perspectiva de análise tomada, bem como das especificidades próprias a diferentes períodos históricos e localizações geográficas. Ou, em outras palavras, o trabalho com dados históricos depende de algumas regras áureas para evitar interpretações normativas e enviesadas; 2) enquanto os sociólogos focam sua atenção em dados primários, resultantes, em geral, da pesquisa de campo, bem como em sua análise secundária (de maneira semelhante às ciências naturais), os historiadores buscam seus dados em bases documentais, com frequência resultantes de intricados processos históricos e, portanto, difíceis de extrair e analisar; 3) enquanto os historiadores colocam grande atenção em verificar todo e cada um de seus dados, os sociólogos visam à obtenção de amostragens de dados que sejam abrangentes.

Em síntese, seria possível dizer que, diferentemente da abordagem de análise histórica, a sociológica tende a deixar de lado fatores individuais e contingentes, passíveis de produzir dados incomuns ou fora de esquadro que poderiam interferir no conjunto mais geral e modular pretendido (Weldon, 2015; Alfonso-Goldfarb *et al.*, 2018). Todavia, nos últimos anos, as pesquisas históricas têm buscado um nicho adequado e próprio nas HD. No que diz respeito ao presente estudo de caso, seria importante indicar que, já há algum tempo, pesquisadores em História e Filosofia da Ciência (HFC), assim como em estudos em ciência e tecnologia (ECT) têm buscado conceitualizar as várias ciências e suas histórias como um sistema complexo que inclui uma diversidade de fatores. Tais fatores abrangem desde redes sociais e institucionais, padrões de financiamento, desenvolvimentos técnicos e tecnológicos, até as variadas mudanças nos pressupostos epistemológicos, passando por sua lógica intrínseca e outras questões próprias ao conhecimento científico, tal como refletidas nos *corpora* textuais.[5] Até o momento, é possível dizer que vários desses fatores têm se mostrado passíveis de modelagem e análise digital.

5 *Corpus* textual (plural, *corpora*), na linguística, é um conjunto amplo e estruturado de textos, utilizado para a realização de análises estatísticas e testar hipóteses. Assim, essa expressão passou a ser utilizada para nomear o substrato de trabalho nas Humanidades Digitais; grosso modo corresponde ao que nas disciplinas históricas é denominado *corpus* documental. Para maior detalhe, vide Wynne (2005).

Contudo, esse encontro, entre a nova tecnologia digital e as Humanidades, não ocorreu sem uma certa tensão. O motivo fundamental é que nas HD, em geral, e nas áreas históricas, em particular, a geração de conjuntos de dados (*datasets*) não antecede nem é independente do próprio trabalho de pesquisa (SCHREIBMAN *et al.*, 2008, p. xxv). Assim, van Zundert (2012, p. 173-174) alerta que as ferramentas informáticas "também devem assegurar que as heurísticas e hermenêuticas já existentes sejam apropriadamente traduzidas nos seus equivalentes digitais, especialmente nas áreas nas quais a heterogeneidade dos dados e abordagens multifacetadas não são consideradas ruído a ser reduzido, mas propriedades essenciais da área".

Não por acaso, a maioria dos projetos ainda mantêm o foco em aspectos mais pertinentes às ciências sociais, tais como redes sociais, financeiras, tecnológicas e institucionais. Muito embora, o escopo de aplicação de tecnologias digitais à pesquisa em história da ciência já consegue abranger diversas possibilidades, incluindo a análise estatística e semântica de seus *corpora* textuais, de sorte que a criação de acervos digitais organizados nesse campo – tanto com base na vida e na obra de cientistas individuais, como Isaac Newton (WALSH & HOOPER, 2012; ILIFFE, 2004; PASTORINO *et al.*, 2008), Henri Poincaré (*Poincaré Project*) e Charles Darwin (*Darwin Correspondence Project, Darwin Manuscript Project*), quanto de instituições, como *History of the Marine Biological Laboratory* ou *History of the Max Planck Society* – vêm estabelecendo uma linha bem definida de estudos. A organização e catalogação de tais acervos, especialmente de correspondências e atas de sessões institucionais, já teve como efeito colateral a produção de uma grande riqueza de metadados relacionais, que permitiram descrever redes de interação entre atores, sua produção coligada e instituições de referência mais significativas. Outros bons exemplos nessa direção são os projetos *Six Degrees of Francis Bacon* e *Registres de l'Académie*, cujo tipo de *dataset* gerado já se assemelha ao obtido pelas Ciências Sociais.

Esse intenso trabalho de digitalização de documentos *em* e *para* a história da ciência, realizado desde os anos 90, vem fornecendo também novas possibilidades de estudos teóricos sobre a organização e classificação dos conhecimentos. Hoje considerados de grande relevância para o desenvolvimento de campos em formação como as HD, esses estudos são o ponto focal do presente estudo de caso, conforme será visto a seguir.

Foco de nossas pesquisas em Humanidades Digitais (HD), História e Filosofia da Ciência (HFC) e Estudos de Ciência e Tecnologia (ECT)

O foco de nossas pesquisas tem como base trabalhos realizados ou em desenvolvimento no CESIMA (Centro Simão Mathias de Estudos em História da Ciência, PUC-SP), centro de pesquisa fundado em 1994. Apenas para recordar, de forma breve, nosso centro de pesquisa foi criado, explicitamente, dentro de uma perspectiva digital, logo no início dos questionamentos sobre a aplicabilidade dessa nova vertente aos complexos campos das Humanidades. O objetivo imediato era disponibilizar, em suporte digital, um *corpus* documental significativo para pesquisas em história da ciência e campos relacionados. Desde o início, a biblioteca digital do CESIMA esteve aberta, não só a membros de nosso centro, mas a outros estudiosos brasileiros ou latino-americanos que requeressem seu uso para fins de pesquisa. Vale lembrar que, antes da possibilidade recente de acesso pela rede mundial, a consulta aos documentos originais com frequência demandava o deslocamento dos pesquisadores até outros centros, muitas vezes distantes, como os europeus, norte-americanos ou mezzo-orientais (GOLDFARB & SOUZA, 2018). Ou ainda, aqueles centros que, tanto no país, quanto no exterior, não tinham seus acervos microfilmados ou digitalizados, trabalho este iniciado com sucesso por membros de nossa equipe. Assim, a primeira fase do trabalho consistiu na digitalização de um vasto *corpus* textual, que, na atualidade, compreende mais de 30.000 títulos. A segunda fase, já em vias de conclusão, consiste na disponibilização completa do acervo para acesso remoto.

Contudo, ao organizar a base digital de dados, nos deparamos com uma série de questões, relacionadas à classificação e à catalogação dos documentos, que dificultavam bastante a sua localização pelos usuários. Questões que, na verdade, não são de fácil solução, pois no caso da história da ciência têm origem em sua própria natureza, estabelecida através das interfaces entre um variado leque de áreas (ALFONSO-GOLDFARB, 2008; ALFONSO-GOLDFARB *et al.*, 2013; GARBER, 2002, 2009; RHEINBERGER, 2007, 2010; WELDON, 2009; WITHROW, 1964, 1976). Por sua vez, tal singularidade pode levar ao fenômeno de *overtagging*, em buscas digitais, em que o volume e emaranhado de termos obtidos é tal que não se produzem resultados significativos (HANRAHAN & SCHNÖPF, 2013). Com a finalidade de enfrentar essa e outras questões, referentes à organização e classificação do conhecimento, o CESIMA deu início a uma série de pesquisas e atividades envolvendo parcerias nacionais e internacionais. Dentre estas, vale destacar as mantidas há muitos anos com o *Grupo de Pesquisa em Tec-*

nologia Aplicada à Educação (GTech.Edu/UFRGS), produtor do minerador de conceitos brasileiro *Sobek*, o *Digital History and Philosophy of Science Consortium* (dHPS), o *Committee of Bibliography and Documentation, Division of History of Science and Technology, International Union of History and Philosophy of Science* (CBD/DHST/IUHPS) e a *Isis Current Bibliography* (ICB) o principal indexador de nosso campo de estudos.

Nesse contexto, inicialmente, foi desenvolvido, já em 1994, um projeto piloto para criação de biblioteca digital que, então inédita no meio acadêmico nacional, teve apoio da FAPESP. O fortalecimento e ampliação dessa iniciativa contou por quase duas décadas (1999-2018) com três grandes Projetos Temáticos da FAPESP. De forma mais específica e, desta vez, com apoio do CNPq, entre os anos de 2010-2012, foi desenvolvido o projeto "Novas perspectivas de classificação e abordagem em história da ciência: aspectos teórico-metodológicos e técnicos para elaboração de instrumentos adequados de busca". No decorrer do mesmo, foram realizadas revisões meticulosas e, ao mesmo tempo, abrangentes das assim chamadas "árvores do conhecimento", ou sistemas de hierarquização das ciências. Ao mesmo tempo, foram também realizados encontros programados, onde se deram amplas discussões com destacados especialistas em classificação, bibliografia, biblioteconomia, arquivística, além de especialistas em ciências da informação e em tecnologia da informação (TI). Os resultados, comunicados em diversos encontros internacionais e publicações especializadas, nos levaram a propor o desenvolvimento, para uma base digital em história da ciência, da chamada classificação *Colon*. Tal classificação foi elaborada nos anos 30, do século XX, pelo matemático indiano S.R. Ranganathan e aplicada com grande sucesso nas Ciências Exatas e Naturais, a partir da década de 1950, pelo *Classification Research Group* (CRG) do Reino Unido (FRICKÉ, 2012, p. 212 *et seq.*; SPITERI, 1998).

O sistema de Ranganathan apresenta notáveis vantagens: 1) decompõe (faz a análise) a informação em unidades básicas ("facetas", correspondendo a: tempo, espaço, materialidade, energia/movimento e personalidade/características individuais) que, na sequência, podem ser recombinadas (faz a síntese), dando origem a categorias gerais, incluindo as multi-, inter- e transdisciplinares; 2) a abordagem "de baixo para cima" (ou seja, de elementos inerentes ao próprio documento às categorias gerais às quais pertence) que acompanha a atuação real do usuário, especialmente no mundo digital, no qual a sequenciação temporal (*strings*) substituiu a categorização espacial (prateleiras tangíveis em bibliotecas); 3) as facetas permitem representar conceitos, em vez de termos (RANGANATHAN, 2012).

Todavia, é preciso lembrar que, pelo menos, desde a década de 1960 entende-se que a história da ciência lida com conceitos e suas ressignificações, em vez de termos (Canguilhem, 1977, p. 11-27). Assim, as tentativas iniciais logo detectaram dois obstáculos de difícil solução: 1) um dos passos na programação exige, necessariamente, decisão humana (escolha de palavras-chave mais significativas ou que melhor descrevam um grupo de documentos) e, portanto, não poderia ser automatizado; 2) levando em conta, como indicado há pouco, que a história da ciência lida com conceitos e que estes mudam no decorrer do tempo, seria essencial incluir esse componente de variação nos demais itens da classificação. Portanto, seria necessário elaborar uma modelização do tempo, algo que até o momento é de difícil realização, como demonstrado em publicações recentes de, por exemplo, Weldon (2015); Alfonso-Goldfarb *et al.* (2018); Gibson *et al.* (2019).

Conforme indicado antes, porém, existem outras possibilidades que implicam na adequação de modelos e métodos desenvolvidos em outros campos informáticos às questões próprias da história da ciência. Uma de tais estratégias é a modelagem de tópicos, a saber, um conjunto de métodos matemáticos e estatísticos para se inferir conceitos ou tópicos a partir de grandes coleções de textos digitalizados. Tais métodos foram desenvolvidos dentro da área de aprendizagem automática (*machine learning*) com o objetivo básico de otimizar ferramentas de busca *online*. A modelagem de tópicos foi amplamente adotada nas HD, geralmente para explorar o conteúdo temático de grandes *corpora* históricos ou literários. Em HFC/ECT, a modelagem de tópicos pode ser utilizada para mapear as mudanças na representação de tópicos e conceitos através do tempo (Hall *et al.*, 2008). Como exemplos dessas ferramentas modeladoras, temos a *Análise Semântica Latente* (LSA – *Latent Semantic Analysis*) e a *Alocação Latente Dirichlet* (LDA – *Latent Dirichlet Allocation*), aplicadas para mapear, respectivamente, a cronologia dos trabalhos alquímicos/químicos por Newton (Walsh & Hopper, 2011) e as mudanças dos hábitos de leitura de Darwin (Murdock *et al.*, 2015). Em ambos os casos, trata-se de uma análise quantitativa que expressa relações latentes entre textos e conceitos.

De maneira semelhante, o campo da linguística computacional representa uma fonte promissora para a abordagem digital da documentação em história da ciência. A linguística computacional utiliza a aprendizagem automática de maneira estatística para estudar a estrutura e a evolução da linguagem, permitindo a detecção e comparação de estruturas epistêmicas e padrões de linguagem numa determinada comunidade científica. Um caso exemplar é representado pelo trabalho desenvolvido por Pumfrey & Ashcroft (2015) sobre a conceptualização dos trabalhos de laboratório na primeira modernidade.

Assim, para melhor indicar as condições existentes para o desenvolvimento e estabelecimento dessas pesquisas, junto ao CESIMA, na sequência são oferecidos dois itens especificando o trabalho de implementação da biblioteca *CESIMA Digital* e seus vínculos com outros centros de estudos.

a. Biblioteca *CESIMA Digital*

Originalmente dedicada a digitalizar obras antigas ou raras sobre as ciências, em diferentes épocas e suportes (manuscritos, livros, cartas, diários, atas de reuniões, cadernos de laboratório e outros mais em microfilmes, microformas, papel etc.), a base de dados documental do CESIMA passou por uma série de modificações, através da adaptação de equipamentos e softwares diversos, até tornar-se a atual Biblioteca *CESIMA Digital*.

No que tange ao acervo, este foi obtido a partir de autorizações ou aquisições nos mais diversos centros de documentação, bibliotecas, arquivos, nacionais e internacionais. A padronização dessa documentação digitalizada ou já digital foi feita em Adobe/PDF, por ser este considerado um padrão mundialmente utilizado. Além disso, todos os arquivos com origem em material impresso foram submetidos a aplicativo de reconhecimento de caracteres (OCR – *Optical Character Recognition*), de maneira a facilitar as buscas e uso de mineradores. Todavia, a Biblioteca *CESIMA Digital* também engloba um número considerável de arquivos com origem em manuscritos que ainda depende de futuros desenvolvimentos de aplicativos especiais para reconhecimento de seus irregulares e variadíssimos tipos de caracteres.

Atualmente, o acervo completo, passado por OCR ou não, encontra-se no servidor do CESIMA, instalado e configurado no *backbone* na *Divisão de Tecnologia da Informação* (DTI) da PUC-SP. Foi selecionado o software *Dspac*e de gerenciamento e preservação de documentos online, sendo este compatível com um sistema internacional de suporte a bibliotecas e indexação de seus conteúdos. Para sua utilização, dispõe-se de duas unidades de armazenamento (*Storage*) conectadas que se encontram devidamente instaladas e ligadas ao servidor do CESIMA com autorização de acesso pela DTI. Além dessa configuração inicial, foi escolhido o banco de dados gratuito *PostgreSQL*, compatível e adequado ao *Dspace*. Detalhes teóricos e técnicos, mais específicos, podem ser obtidos em Souza (2019), cuja tese de Doutorado teve como foco o processo de formação e disponibilização da Biblioteca *CESIMA Digital*.

Aberta universalmente ao público, em julho de 2018, a Biblioteca *CESIMA Digital*, oferece mais de 1/3 de seu acervo para acesso remoto, enquanto o restante passa por um criterioso, mas acelerado, processo de verificação e catalogação antes de seu *upload* que, se espera, seja concluído brevemente. Embora o catálogo desse grande conjunto de obras seja aberto, para acessá-las *online* ou fazer seu *download* é solicitado aos usuários um cadastramento, a partir do qual recebem uma senha de acesso, após aprovação do comitê científico do CESIMA. Desta forma, tornou-se possível garantir o devido respeito aos direitos autorais e uso científico dos materiais acessados, conforme previamente solicitado pelos centros de documentação em que tiveram origem e a norma internacionalmente aceita. De igual maneira, tal cadastramento tem facilitado saber quem são os usuários, de onde provêm e, de certa forma, suas áreas de interesse. A relevância de tal iniciativa pode ser avaliada por seu alcance. Com menos de dois anos de operação, ainda sem o catálogo completo *online* e sem grande divulgação, a Biblioteca *CESIMA Digital* já conta com dezenas de usuários cadastrados e centenas de consultas, somando 1946 acessos, desde 97 diferentes territórios, em diferentes partes do mundo. No mapa abaixo, as áreas em azul indicam os territórios originários das consultas, enquanto a tabela subsequente oferece sua localização específica e seus números, até os primeiros meses de 2020.

Todavia, lembrando que o *CESIMA Digital* não se trata apenas de uma biblioteca online para consulta, como tantas outras, mas de uma coleção muito especial reunida por pesquisadores – através de suas pesquisas individuais e conjuntas – e oferecida a outros estudiosos de HFC/ECT, as questões teóricas e práticas ali encetadas têm sido de grande interesse para as HD e campos afins. Nesse sentido, o próximo item é dedicado,

Figura 1. No mapa, as áreas em azul indicam os territórios originários das consultas à Biblioteca Digital do CESIMA (Fonte: AUTORES, 2020)

Acessos por País (01jul2018 a 19mar2020) total 1946							
Brazil	829	Russia	8	Pakistan	3	Cape Verde	1
United States	192	Austria	7	Romania	3	Cyprus	1
Argentina	178	Chile	7	Slovenia	3	Dominican Republic	1
United Kingdom	77	Colombia	7	Thailand	3	Estonia	1
Portugal	69	Greece	7	Georgia	2	Ethiopia	1
Spain	48	Ireland	7	Hong Kong	2	Guinea	1
Italy	43	Sweden	7	Iraq	2	Guatemala	1
France	33	Czechia	6	Kenya	2	Croatia	1
Mexico	32	Egypt	6	Lebanon	2	Kuwait	1
Germany	29	Hungary	6	Sri Lanka	2	Kazakhstan	1
China	26	South Africa	6	Mozambique	2	Lithuania	1
Canada	25	Ecuador	5	Norway	2	Madagascar	1
Netherlands	24	Finland	5	Slovakia	2	North Macedonia	1
India	20	Morocco	5	El Salvador	2	Macao	1
Australia	19	Poland	5	Ukraine	2	Niger	1
Japan	17	Taiwan	5	Venezuela	2	Nigeria	1
(not set)	15	Denmark	4	United Arab Emirates	1	Oman	1
Switzerland	12	New Zealand	4	Angola	1	Réunion	1
Israel	12	Peru	4	Bosnia & Herzegovina	1	Serbia	1
Philippines	12	Cameroon	3	Bulgaria	1	Saudi Arabia	1
Turkey	11	Algeria	3	Benin	1	Singapore	1
Indonesia	10	Ghana	3	Belize	1	Uruguay	1
South Korea	10	Iran	3	Côte d'Ivoire	1	Vietnam	1
Belgium	8	Malaysia	3	Costa Rica	1	Yemen	1
						Zimbabwe	1

Figura 2. A tabela mostra os acessos por país das consultas e seus números até os primeiros meses de 2020. (Fonte: AUTORES, 2020)

especialmente, a oferecer um pouco mais de detalhes sobre duas parcerias, dentre aquelas mantidas pelo CESIMA, com foco particular em sua biblioteca digital.

b. Parcerias focadas na Biblioteca *CESIMA Digital*

Entre os parceiros mais relevantes para o *CESIMA Digital*, indicados anteriormente – CBD, ICB, GTech.Edu, dHPS –, os dois últimos merecem especial atenção pelo trabalho conjunto que vem sendo desenvolvido no que concerne as HD e campos coligados. O mais antigo dentre ambos, com cuja equipe já desenvolve pesquisas há anos, é o GTech.Edu, estabelecido na UFRGS e reconhecido nacional e internacionalmente.

Liderada pelo Prof. Eliseo Reategui, essa equipe desenvolveu o SOBEK, ferramenta de acesso aberto que fornece a possibilidade – não só a pesquisadores da UFRGS, mas a outros espalhados pelo planeta – de compor, interpretar e avaliar textos. Segundo indicado na própria página de tal ferramenta, seus fundamentos técnicos e teóricos foram desenvolvidos a partir dos estudos de Schencker (2003), cuja tese de Doutorado fornece os princípios estatísticos a partir dos quais os algoritmos são habilitados para "minerar", ou seja, identificar os termos mais significativos de um texto e, seguidamente, relacioná-los em "árvore". Tal processo, utilizado e ampliado pelo SOBEK, transformaria simples termos em *quasi*-conceitos, uma vez que passa a fornecer sentido a estes, através de seus vínculos hierárquicos e relacionais com o todo do texto, possibilitando, assim, a sua interpretação.

Uma nova versão do SOBEK para *desktop*, elaborada em conjunto com o Prof. Reategui e sua equipe, possibilitou a obtenção de um bom número desses *quasi*-conceitos em uma seleção de obras pertencentes ao acervo do *CESIMA Digital*. Com isso, encontra-se em desenvolvimento um trabalho para, a partir desses *quasi*-conceitos, formular computacionalmente o que especialistas em documentação chamam de descritores, ou seja, descrições breves de cada documento associadas às informações convencionais para a sua catalogação (autor, título, ano etc.). Penosos e demorados, pois feitos individualmente por especialistas e/ou documentalistas conhecedores de cada obra em questão, os descritores quase sempre aparecem em catálogos virtuais apenas como cópias dos já existentes em meio físico. Dessa forma, a nova versão do SOBEK, aplicada ao *CESIMA Digital*, deverá oferecer, em futuro não muito distante, possibilidades de busca que vão além das meras informações dos catálogos convencionais. Possibilidades essas que poderão se ampliar imensamente uma vez cruzadas, por tal sistema, às informações contidas em diferentes obras e, assim, verificar seus conjuntos conceituais em comum. Algo que, conforme já mencionado, interessa de forma particular aos estudos em história da ciência, sempre focados em identificar a mudança conceitual de um mesmo termo, em diferentes épocas. Em uma perspectiva mais próxima e um pouco menos ambiciosa, já se encontra em desenvolvimento um trabalho para, a partir dos descritores, acompanhar – com a devida autorização do usuário – quem consultou ou está consultando um determinado documento, quais outros documentos semelhantes consultou, e se deixou alguma observação sobre os mesmos (SOUZA, 2019).

Uma segunda parceria que vem se desenvolvendo com força, embora há menos tempo, está vinculada ao dHPS, por constituir-se num marco referencial para os estudos em HD. Marco este cujo espectro internacional e mais amplo vem, inclusive, beneficiando as parcerias já existentes e gerando outras novas para o *CESIMA Digital*.

Apenas para resumir algo de sua história, o dHPS foi criado em 2011, a partir de projetos em documentação exclusivamente para a história da ciência, financiados pela *Nacional Science Foundation* (NSF) dos EUA. Não demoraria muito para que esse consórcio conseguisse reunir tanto historiadores e filósofos da ciência, quanto especialistas em TI, Ciências da Computação, Biblioteconomia e outras ciências da informação. O propósito original desse colegiado multi e interdisciplinar era desenvolver novas formas de integrar as abordagens tradicionais de pesquisa, em HFS, com ferramentas computacionais e recursos digitais. Atualmente, a missão proposta pelo dHPS é mais abrangente e ambiciosa, conforme

expresso em suas próprias palavras: "desenvolver, dar suporte e promover projetos digitais em HFC, incluindo edições, publicações e as ferramentas de pesquisa necessárias para tanto. Na medida do possível e reconhecendo os desafios e limites o Consórcio está compromissado com produtos de código e acesso abertos. Comprometemo-nos a desenvolver infraestrutura sustentável para ambos, projetos e produtos" (Disponível em: digitalhps.org. Acesso em: 25 jul. 2020).

Estimulado pela própria NSF, entre 2016 e 2017, o dHPS apresentou e foi contemplado, por essa mesma agência, com um projeto para coordenar pesquisas em rede, o chamado *Research Coordination Network* (RCN). Fazem parte desse projeto, que se propõe a fundamentar boa parte dos trabalhos da dHPS, além de membros de diversas universidades dos EUA, também representantes das seguintes instituições: *Chemical Heritage Foundation*/EUA, *American Science Museum*/EUA, *Stevens Institute of Technology*/EUA, *Humboldt-Universität zu Berlin*/Alemanha, *Max Planck Institute*/Alemanha, *University of Melbourne*/Austrália, *University of Cambridge*/Reino Unido, *Université de Nantes*/França. O CESIMA/PUC-SP teve o prazer de também ser convidado a participar desse seleto grupo, colaborando com a elaboração do projeto desde o início e até agora com o seu desenvolvimento.

A partir desse projeto, um dos desafios específicos do dHPS é desenvolver *standards* e soluções tecnológicas para mapear e integrar as diferentes autoridades e ontologias em uso. Primeiramente, espera-se que a boa consecução desse processo deva alcançar o tão sonhado compartilhamento de dados entre projetos. Mas, além disso, a pesquisa necessária para tanto, também deve produzir novas bases conceituais e epistemológicas, no que diz respeito ao manejo e à modelagem de megadados (*big data*), para o trabalho documental em HFC/ECT (Gibson *et al.*, 2019). Nesse contexto, o *CESIMA Digital*, ao não consistir em meras coleções de HFC/ECT, mas ter sido planejado, desde o início, como uma base de dados em que o texto e o contexto das obras devem ser vistos de maneira interligada, foi selecionado como um estudo de caso preferencial para as pesquisas do RCN. De igual forma, foi considerado um ambiente experimental ideal para testar e/ou desenvolver instrumentos de busca que visem detectar conceitos e suas transformações, ao longo do tempo, em textos dedicados às ciências de diferentes épocas.

Considerações finais

A pesquisa em Humanidades, progressivamente inserida na era digital, vem se transformando, ao aderir, não só aos processos tecnológicos, mas teórico-metodológicos produzidos nesse novo contexto. Congregadas cada vez mais no campo recente das Humanidades Digitais (HD), pesquisas históricas, como aquelas em história da ciência e áreas afins, podem agora (ou poderão em breve) se beneficiar das novas ontologias e técnicas informáticas e produzir análises estatísticas e semânticas inéditas em seus *corpora* textuais.

De sorte que, ao formar parte de um marco internacional e mais amplo, em que esses estudos têm acontecido em ritmo célere, o *CESIMA Digital* busca constituir uma via de mão dupla, através da qual possa fluir tanto o que for produzido de mais inédito e/ou recente no exterior, quanto as colaborações e/ou formulações originais brasileiras. Assim, espera-se que o desenvolvimento dos projetos dos quais participa coopere na elaboração de novas bases conceituais e tecnológicas referentes ao manejo e à modelagem de megadados para o trabalho documental das Humanidades. Todavia, considerando-se que o processo acima descrito tem como foco principal os estudos históricos, os resultados iniciais deverão contemplá-los e, de forma mais direta, aqueles relacionados à história da ciência e campos coligados. Espera-se ainda que esses desenvolvimentos promovam a vinculação e trabalho conjunto entre membros das equipes inter e multidisciplinares, nacionais e internacionais, que até agora não fazem parte de seu escopo atual de parcerias. Com isso, busca-se também a ampliação de tais redes de pesquisa, além da formação de novos pesquisadores e, não menos importante, um trabalho conjunto e potencializado para tornar efetivo o compartilhamento de dados entre projetos, assim como uma interface mais efetiva e sinérgica entre as pesquisas em Humanidades Digitais (HD) e aquelas em História e Filosofia da Ciência (HFC) e Estudos em Ciência e Tecnologia (ECT).

Referências

ALFONSO-GOLDFARB, A.M.; WAISSE, S.; FERRAZ, M.H.M. From shelves to cyberspace: the organization of knowledge and the complex identity of history of science. *Isis*, v. 104, n. 3, p. 551-560, 2013.

_____. New proposals for organization of knowledge and their role in the development of databases for history of science. *Circumscribere*, v. 21, p. 1-12, 2018.

Alfonso-Goldfarb, A.M. Centenário Simão Mathias: Documentos, métodos e identidade da história da ciência. *Circumscribere*, v. 4, p. 5-9, 2008.

Aronova, E.; von Oertzen, C.; Sepkoski, D. (org.). Data histories. *Osiris*, v. 32, n. 1, p. 1-354, 2017.

Baur, N. Problems of linking theory and data in historical sociology and longitudinal research. *Historical Social Research*, v. 34, n. 1, p. 7-21, 2009.

Berry, D.M. (org.) *Understanding digital humanities*. New York: Palgrave Macmillan, 2012.

Berry, D.M. The computational turn: thinking about the digital humanities. *Culture Machine*, v. 12, 2011.

Canguilhem, G. *Ideologia e racionalidade nas ciências da vida*. Lisboa: Edições 70, 1977.

Caraco, B. Les digital humanities et les bibliothèques. *Bulletin des bibliothèques de France*, n. 2, 2012.

Deegan, M.; McCarty, W. (org.) *Collaborative research in the digital humanities*. Farnham: Ashgate, 2012.

Ferraz, M. H. M.; Alfonso-Goldfarb, A.M.; Waisse, S. Science and History of Science: between Comte and Canguilhem. *Transversal*, v. 4, p. 108-117, 2018.

Fitzpatrick, K. The humanities, done digitally. *The Chronicle of Higher Education*, 8/5/2011. Disponível em: <chronicle.com/article/The-Humanities-Done-Digitally/127382/>. Acesso em: 12 maio 2020.

Frické, M. *Logic and the organization of information*. New York: Springer, 2012.

McCarty, W. *Humanities computing*. New York: Palgrave, 2005.

Garber, D. Philosophia, historia and mathematica: shifting sands in the intellectual geography of the seventeenth century. *Studies in the History and Philosophy of Science*, v. 24, p. 1-7, 2009.

Garber, D. *Storia della scienza, v. 4: La rivoluzzione scientifica*. Roma: Instituto della Enciclopedia Italiana, 2002.

Gardiner, E.; Musto, R. *The digital humanities*: a primer for students and scholars. Cambridge: Cambridge University Press, 2015.

Gibson, A.; Laubichler, M.D.; Maienschein, J. (org.). Dossiê: Computational History and Philosophy of Science. *Isis*, v. 110, n. 3, p. 497-566, 2019.

Goldfarb, J.L.; Souza, O. From the Golem's Jewish Myth to IBM's responsive Watson: where are we going? *Circumscribere*, vol. 21, p. 118-122, 2018.

Hanrahan, E.; Schnöpf, M. Scholarly digital editions: connecting archives and libraries. Conference *New Directions in Digital History of Science*. Berlin: Max Planck Institute of History of Science/Committee of Documentation and Bibliography, IUHPS, 2013.

Hall, D.; Jurafsky, D.; Manning, C.D. Studying the history of ideas using topic models. *Proceedings of the Conference on Empirical Methods in NATURAL Language Processing, EMNLP/2008*, Edimburgo, 25-27 outubro 2008.

Hayles, K. *How we think: digital media and contemporary technogenesis*. Chicago: The University of Chicago Press, 2010.

Hu, W.C.; Kaabouch, N. *Big data management, technologies and applications*. Hershey: Information Science Reference, 2014.

Hu, Y.; Boyd-Graber, J.; Satinoff, B.; Smith, A. Interactive topic modeling. *Machine Learning*, v. 95, p. 423-469, 2014.

Iliffe, R. Digitizing Isaac: the Newton Project and an electronic edition of Newton's papers. *In*: Force, J.E.; Hutton, S. (org.) *Newton and Newtonianism*: new studies. Dordrecht: Springer, p. 23-38, 2004.

Kirschenbaum, M. What is digital humanities and what's doing in English departments? *ADE Bulletin*, n. 150, p. 1-7, 2010.

Montemurro, M.A.; Zanette, D.H. Keywords and co-occurrence patterns in the Voynich manuscript: An information-theoretic analysis. *Plos One* 21/6/2013 DOI: 10.1371/journal.pone.0066344

Murdock, J.; Allen, C.; Dedeo, S. Exploration and exploitation of Victorian science in Darwin's reading notebooks. *Cornell University Library*, 10/12/2015 Disponível em: <arXiv:1509.07175v2>. [cs.CL]. Acesso em: 12 maio 2020.

Pastorino, C.; Lopez, T.; Walsh, J.A. The digital Index Chemicus: toward a digital tool for studying Isaac Newton's Index Chemicus. *Body, Space & Technology*, v. 7, n. 2, 2008.

Presner, T. Digital humanities 2.0: a report on knowledge. *OpenStax CNX*. 8/6/2010. Disponível em: <cnx.org/contents/2742bb37-7c47-4bee-bb34-0f35bda760f3@6>. Acesso em: 14 maio 2020.

Pumfry, S.; Ashcroft, P. Alchemy, chemistry or chymistry: an analysis of actors' categories in early modern England. *16th shac Postgraduate Workshop*. Oxford, 30 outubro 2015. Programme and Edited Abstracts, p. 5-6.

Ranganathan, S.R. *Colon Classification*. [reimp. sexta ed. 1960]. Bangalore: Ess Ess Publ., 2012.

Rheinberger, H.-J. *On historicizing epistemology*: An essay. Stanford (ca): Stanford University Press, 2010.

Rheinberger, H.-J.*Historische Epistemologie zur Einführung*. Hamburg: Junius, 2007.

Schreibman, S., Siemens, R.; Unsworth, J. (org.). *Companion to digital humanities*. Oxford: Blackwell, 2004.

Schencker, A. *Graph-theoretic Techniques for Web Content Mining*. PhD Dissertation. Tampa: University of South Florida, 2003.

Souza, O. *Cesima Digital*: *uma ferramenta para a história da ciência*, 2019. Tese (Doutorado em História da Ciência). Pontifícia Universidade Católica de São Paulo, São Paulo.

Spiteri, L. A simplified model for facet analysis. *Canadian Journal of Information and Library Science*, v. 23, p. 1-30, 1998.

Svensson, P. Humanities computing as digital humanities. *Digital Humanities Quarterly*, v. 3, n. 3, 2009.

Svensson, P. The landscape of digital humanities. *Digital Humanities Quarterly*, v. 3, n. 3, 2009.

Terras, M; Nyhan, J.; Vanhoutte, E. (org.) *Defining digital humanities*: a reader. Farnham: Ashgate, 2013.

Unsworth, J. What is humanities computing and what is not? *In*: Terras, M; Nyhan, J.; Vanhoutte, E. (org.) *Defining digital humanities*: a reader. Farnham: Ashgate, p. 35-48, 2013.

Walsh, J.A.; Hooper, W.E. The liberty of invention: alchemical discourse and Information Technology standardization". *Literary and Linguistic Computing*, v. 27, p. 55-79, 2012.

Weldon, S.P. Historians and their data. *In*: Alfonso-Goldfarb, A.M. *et al.* (org.) *Crossing oceans: Exchange of products, instruments and procedures in the history of chemistry and related sciences*. Campinas: cle/unicamp, p. 299-322, 2015.

Weldon, S.P. The Isis bibliography from its origins to the present day: one hundred years of evolution of a classification system. *Circumscribere*, v. 6, p. 26-46, 2009.

Withrow, M. A classification scheme for the history of science, technology and medicine. *Isis Cumulative Bibliography*, v. 3, p. 621-623, 1976.

Withrow, M. Classification schemes for the history of science: a comparison. *Journal of Documentation*, v. 20, n. 3, p. 120-136, 1964.

Wynne, M. (org.). *Developing linguistic corpora*: a guide to good practice. Oxford: Oxbow Books, 2005.

Zundert, J. v. "If you build it, will we come? Large scale digital infrastructures as a dead end for digital humanities". *Historical Social Research*, v. 37, n. 3, p. 165-186, 2012.

Consórcios, projetos e ferramentas

Biblioteca *CESIMA Digital*. Disponível em: <cesimadigital.pucsp.br>. Acesso em: 25 jul. 2020.

Darwin Correspondence Project. Disponível em: <darwinproject.ac.uk>. Acesso em: 25 jul. 2020.

Darwin Manuscript Project. Disponível em: <amnh.org/our-research/darwin-manuscripts-project>. Acesso em: 25 jul. 2020.

Digital History And Philosophy Of Science. Disponível em: <digitalhps.org>.

History Of The Marine Biological Laboratory. Disponível em: <history.archives.mbl.edu>. Acesso em: 25 jul. 2020.

History Of The Max Planck Society. Disponível em: <mpiwg-berlin.mpg.de/en/research/projects/DEPT1_458_HistMPS>. Acesso em: 25 jul. 2020.

Poincaré Correspondence Project. Disponível em: <poincare.univ-nancy2.fr>. Acesso em: 25 jul. 2020.

Sobek. Disponível em: <sobek.ufrgs.br>. Acesso em: 25 jul. 2020.

Registres De L'académie. Disponível em: <mpiwg-berlin.mpg.de/en/research/projects/DEPT1_458_HistMPS>. Acesso em: 25 jul. 2020.

Research Coordination Network/National Science Foundation. Disponível em: <digitalhps.org/node/184 nsf.gov/awardsearch/showAward?AWD_ID=1656284&HistoricalAwards=false>. Acesso em: 25 jul. 2020.

Six Degrees Of Francis Bacon. Disponível em: <mpiwg-berlin.mpg.de/en/research/projects/DEPT1_458_HistMPS>. Acesso em: 25 jul. 2020.

Text Encoding Initiative Consortium. Disponível em: <tei-c.org>. Acesso em: 25 jul. 2020.