Revista ACB: Biblioteconomia em Santa Catarina,
Florianópolis, v. 26, n. 4, p. 1-30, set./dez., 2021.

Recebido em: 28/07/2021 **Aceito em:** 03/12/2021

Completude, consistência e correção em bases de dados digitais sobre mortos e desaparecidos políticos na ditadura civil-militar brasileira

João Alexandre Peschanski[1]
Éder Porto Ferreira Alves[2]

Resumo: O artigo compara quatro bases de dados sobre mortos e desaparecidos na ditadura civil-militar brasileira: a da Comissão Especial sobre Mortos e Desaparecidos Políticos, a do site Desaparecidos Políticos, a do portal Memórias da Ditadura e a do relatório da Comissão Nacional da Verdade. Estas são avaliadas de acordo com a completude, a consistência e a correção das informações, a partir de métodos de análise de bases heterogêneas. Os resultados da comparação indicam variações na qualidade das bases analisadas, tanto se vistas umas em relação às outras quanto se avaliadas as propriedades internas das bases. A comparação envolveu a transferência dos dados nessas bases para o Wikidata, o que fez com que as informações heterogêneas, após curadoria computacional, formassem uma base semântica, mais completa, consistente e precisa.

Palavras-chave: Ditadura brasileira. Mortos e desaparecidos políticos. Bases de dados digitais. Wikidata. Humanidades digitais.

1 INTRODUÇÃO

A curadoria, a estruturação e a catalogação de dados sobre as violações de direitos humanos têm sido um esforço central em projetos de memória sobre a ditadura civil-militar brasileira. Esse esforço iniciou-se clandestinamente, no fim dos anos 1970, com destaque para a coleta e a sistematização de documentos realizadas no contexto do projeto *Brasil: Nunca Mais*. Com o advento das tecnologias digitais, dados sobre as violações de direitos humanos no Brasil entre 1964 e 1985 tornaram-se livremente acessíveis na internet.

Destacam-se quatro bases de dados digitais sobre mortos e desaparecidos políticos na ditadura civil-militar brasileira. A primeira foi criada pela Comissão Especial sobre Mortos e Desaparecidos Políticos, instituída em 1995 pela Secretaria de Direitos Humanos da Presidência da República (BRASIL, 2019). Em 2019, essa base, mantida pelo governo

[1] Doutor em Sociologia pela University of Wisconsin-Madison, professor na Faculdade Cásper Líbero. São Paulo, Brasil. E-mail: japeschanski@casperlibero.edu.br. ORCID: https://orcid.org/0000-0002-2352-1787

[2] Bacharel em Matemática Aplicada pela Universidade de São Paulo, gerente de produtos e recursos no Wiki Movimento Brasil. São Paulo, Brasil. E-mail: eder.porto@wmnobrasil.org. ORCID: https://orcid.org/0000-0001-8723-1431

Revista ACB: Biblioteconomia em Santa Catarina, Florianópolis (Brasil) - ISSN 1414-0594

federal, saiu do ar e os dados nela disponíveis estavam apenas acessíveis a partir de mecanismos de arquivamento de páginas digitais, como o *Wayback Machine*. A segunda base de dados está hospedada no site Desaparecidos Políticos, criado em 2010 pelo Centro de Documentação Eremias Delizoicov e pela Comissão de Familiares dos Mortos e Desaparecidos Políticos (CENTRO DE DOCUMENTAÇÃO EREMIAS DELIZOICOV; DOSSIÊ MORTOS E DESAPARECIDOS POLÍTICOS NO BRASIL, 2019). A terceira base de dados foi produzida pelo Instituto Vladimir Herzog, em 2014, no contexto de um projeto digital mais amplo sobre o período de exceção brasileiro intitulado Memórias da Ditadura (INSTITUTO VLADIMIR HERZOG, 2020). A quarta base de dados é a versão estruturada do relatório final da Comissão Nacional da Verdade, em especial o terceiro volume, de dezembro de 2014 (BRASIL, 2014). No decorrer deste trabalho, as bases de dados digitais são respectivamente identificadas como CEMDP, DP, MD e CNV. Construídas de modo independente, essas bases agregam dados distintos, a partir de metodologias heterogêneas, o que faz com que sua integração se torne problemática.

O objetivo geral deste artigo é comparar características das quatro bases de dados digitais sobre mortos e desaparecidos políticos na ditadura civil-militar brasileira. De nosso conhecimento, não existem trabalhos anteriores que tenham debruçado-se sobre o problema da comparação dos dados nessas iniciativas. Aqui são computadas integralmente as declarações presentes nas bases e avaliadas a partir das seguintes dimensões: completude, consistência e correção. Tais dimensões, que orientam a metodologia comparativa adotada neste artigo, são correntes na literatura (ANGELES; MACKINNON, 2004; SIDI et al., 2012). Na comparação, leva-se em conta apenas a seção estruturada das fichas sobre cada um dos mortos e desaparecidos políticos disponíveis em cada uma das bases; portanto, não são levadas em conta seções narrativas e analíticas, que normalmente acompanham essas fichas.

A apresentação de um projeto de estruturação e consolidação em uma base semântica unificada das informações presentes na CEMDP, DP, MD e CNV é um objetivo complementar deste artigo. A plataforma adotada foi o Wikidata. A unificação dos dados permitiu estabelecer um parâmetro mais geral de controle para a comparação entre as quatro bases digitais estudadas e transcluir os dados computados para a Wikipédia, potencializando a difusão das informações coletadas. A transclusão foi realizada com o robô Listeria (PREDEFINIÇÃO:LISTA DO WIKIDATA, 2018), na página em português *Lista de mortos e*

Revista ACB: Biblioteconomia em Santa Catarina, Florianópolis (Brasil) - ISSN 1414-0594

desaparecidos políticos na ditadura militar brasileira (LISTA DE MORTOS E DESAPARECIDOS POLÍTICOS NA DITADURA MILITAR BRASILEIRA, 2020).

Os objetos, os métodos e a proposta prática deste artigo inserem-se no campo das Humanidades Digitais. De certo modo, as bases CEMDP, DP e MD são objetos originalmente digitais (BERRY, 2011), na medida em que disponibilizam eletronicamente informações relativamente estruturadas, que se tornam a partir daí, individual ou agregadamente, uma referência de pesquisa on-line. O caso da CNV é diferente, na medida em que a versão eletrônica utilizada é uma reprodução digital do relatório impresso. Os parâmetros metodológicos aplicados na convergência das bases transferem aprendizados da ciência da computação e da pesquisa focada em dados para o campo das humanidades (HALL, 2013; NOIRET, 2015). A estruturação semântica dos dados sobre mortos e desaparecidos políticos no Wikidata insere uma ferramenta digital na análise dos dados coletados e, mais do que isso, produz um objeto computacional novo a partir desses dados. A adoção do Wikidata remete a duas linhas contemporâneas de experimentação nas humanidades digitais: o uso de ferramentas propriamente wiki em análises *nascidas digitais* (MARTINS; CARMO, 2019) e a experimentação do Wikidata como ferramenta para a produção automática de conteúdos legíveis por robôs e humanos (AZZELLINI; PESCHANSKI; PAIXÃO, 2019).

O artigo não se propõe a investigar detalhadamente as metodologias de cada base analisada e, de modo específico, a conceituação aplicada aos mortos e desaparecidos em cada uma delas. Essa conceituação é objeto de debate no Brasil (AYDOS; FIGUEIREDO, 2013; SARTI, 2014; AZEVEDO, 2016) e em outros contextos, como na Argentina (VECCHIOLI, 2001), Espanha (DRULIOLLE, 2015; GATTI, 2017) e Chile (BERNASCONI; RUIZ; LIRA, 2018). De todo modo, a análise aqui realizada permite um leque amplo de desdobramentos: por exemplo, reconhecer as diferenças entre as bases organizadas pelo Estado e por movimentos sociais, as variações entre as próprias bases estatais (no caso da CEMDP e da CNV) e os critérios presentes em todos os bancos de dados na definição de um indivíduo como morto ou desaparecido político.

Iniciamos este artigo com uma apresentação das bases de dados sobre mortos e desaparecidos políticos na ditadura civil-militar brasileira. Depois, apresentamos a estratégia de estruturação semântica das informações transpostas da CEMDP, DP, MD e CNV para o Wikidata. Na sequência, comparamos as bases de dados de acordo com sua completude, consistência e correção, incluindo na avaliação as informações no Wikidata.

Revista ACB: Biblioteconomia em Santa Catarina, Florianópolis (Brasil) - ISSN 1414-0594

2 INICIATIVAS DE MEMÓRIA DIGITAL

Iniciativas sobre violações de direitos humanos têm utilizado recursos digitais para difundir informações e documentos que coletam. Um propósito é inserir o que é apurado na disputa por narrativas sobre regimes de exceção, como a ditadura civil-militar no Brasil; assim, recursos digitais assumem importância no que foi chamado de *guerras de memória* (PEREIRA, 2015). Os recursos digitais são também formas de garantir mais visibilidade, perenidade e continuidade ao que foi coletado e documentado, participando assim de uma produção social da informação e memória (FROTA, 2019). Nesta seção, são apresentadas com uma revisão de literatura e uma descrição sumária o contexto e o formato da produção de quatro bases sobre mortos e desaparecidos políticos brasileiros entre 1964 e 1985 e a organização das fichas individuais nessas bases.

A Comissão Especial sobre Mortos e Desaparecidos Políticos foi criada a partir da Lei nº 9.140, de 4 de dezembro de 1995, inicialmente vinculada à Secretaria de Direitos Humanos da Presidência da República e realocada no governo de Jair Bolsonaro ao Ministério da Mulher, da Família e dos Direitos Humanos. A CEMDP foi instituída com o propósito político de reconhecer a responsabilidade do Estado no desaparecimento de opositores aos regimes de exceção, em especial na ditadura civil-militar de 1964 a 1985, e a partir da coleta de documentos estabelecer bases para o pagamento de eventuais indenizações associadas a esses desaparecimentos (BRASIL, 1995). A listagem de mortos e desaparecidos na CEMDP assume um caráter de justiça reparadora e foi realizada a partir de levantamentos realizados por advogados e familiares, a partir de depoimentos de presos políticos e agentes do Estado, além de pesquisa documental (ROTTA, 2008). A licença de conteúdo utilizada não foi informada.

As páginas eletrônicas individuais da CEMDP, chamadas fichas descritivas, consistem de três blocos de informação. Num primeiro quadro, são listados: o nome da pessoa morta ou desaparecida, o nome de seu pai e de sua mãe e a idade da pessoa quando desaparecida. Um segundo quadro, intitulado Identificação, traz dados sobre o dossiê e o procedimento administrativo do caso em investigação, além de mais informações biográficas estruturadas e uma seção narrativa de biografia da pessoa morta ou desaparecida. Numa segunda aba associada à mesma ficha descritiva, há informações sobre a situação do

Revista ACB: Biblioteconomia em Santa Catarina, Florianópolis (Brasil) - ISSN 1414-0594

procedimento administrativo. A ordem das informações nas fichas descritivas modifica-se na versão para impressão.

O site Desaparecidos Políticos foi criado em 2010 pelo Centro de Documentação Eremias Delizoicov e a Comissão de Familiares dos Mortos e Desaparecidos Políticos, com o objetivo de difusão sobre os crimes do Estado na ditadura civil-militar. As informações nas fichas da *Lista de nomes*, com mortos e desaparecidos políticos, transpõem as notas biográficas presentes no *Dossiê dos mortos e desaparecidos políticos a partir de 1964* (COMISSÃO DE FAMILIARES DE MORTOS E DESAPARECIDOS POLÍTICOS et al., 1995), uma versão revista e ampliada de compilações anteriores sobre violações de direitos humanos entre 1964 e 1985. Segundo a apresentação da DP, informações da obra de 1995 foram revisadas e completadas. A licença de conteúdo adotada foi o *Copyright* - todos os direitos reservados.

As fichas pessoais da base do DP estão divididas em cinco partes: dados pessoais, dados da militância, dados da repressão, biografia e documentos. As três primeiras são estruturadas com informações de identificação (por exemplo, cidade e data de nascimento e atividade), atuação política (organização, nome falso e dados de morte/desaparecimento) e repressão (agente, órgão). A parte de biografia permanece vazia, normalmente. Na parte de documentos, há listas de materiais coletados sobre a vida e a morte/desaparecimento das pessoas investigadas, incluindo relatórios de direitos humanos, artigos de jornal e fotografias. Os documentos listados não estão normalmente acessíveis digitalmente, mantidos sob a guarda das organizações que criaram o DP.

O portal Memórias da Ditadura foi lançado em 5 de dezembro de 2014 com o propósito de ampliar e aprofundar o conhecimento do *grande público* sobre o período da ditadura civil-militar brasileira, especialmente pensado para atividades pedagógicas, e se autodeclarou “o maior acervo online sobre a história da ditadura no Brasil” (LAITANO, 2019). Foi produzido pelo Instituto Vladimir Herzog, numa parceria com o Programa das Nações Unidas para o Desenvolvimento e a Secretaria de Direitos Humanos da Presidência da República. O portal foi definido como uma ação de *protagonismo da memória*, ao dar relevância às pessoas mortas e desaparecidas na ditadura (KIELING, 2016). Além de uma lista de mortos e desaparecidos na ditadura, o portal traz documentos, subportais sobre temas específicos do período de exceção, produtos multimidiáticos (mapas, linhas do tempo,

Revista ACB: Biblioteconomia em Santa Catarina, Florianópolis (Brasil) - ISSN 1414-0594

mosaicos fotográficos) (MARTINS; MIGOWSKI, 2015). O portal adota a licença *Creative Commons Zero Universal*.

As páginas individuais do MD estão divididas em três seções. Na primeira, há, além de fotografias, dados biográficos estruturados, incluindo informações sobre o nascimento, a atuação profissional e política e a morte ou desaparecimento. Na segunda, há um relato descritivo e analítico das circunstâncias da morte, à qual segue, na terceira seção, um resumo da conclusão da Comissão Nacional da Verdade sobre o caso específico tratado.

O relatório final da Comissão Nacional da Verdade, em especial o terceiro volume, foi publicado em dezembro de 2014 e tornou-se o principal documento oficial sobre mortos e desaparecidos políticos na ditadura civil-militar brasileira. Sancionada em 2011, a CNV foi uma entidade de apuração e documentação dos crimes contra os direitos humanos cometidos pelo Estado brasileiro nos governos de exceção, incluindo o período 1964-1985. O relatório tornou-se um objeto importante de estudo acadêmico, destacando, por exemplo, seu impacto na política de memória e transição democrática no Brasil (FERNANDES, 2015), na articulação de organizações de direitos humanos (CANABARRO, 2014) e uma nova compreensão do papel do Estado como agente de repressão (SCHINKE; CASTRO, 2016) e de grupos e entidades da sociedade civil como apoiadores dessa repressão (COSTA; SILVA, 2018). Houve também diversos trabalhos sobre limitações da CNV, incluindo sobre o acervo utilizado (PEDRETTI, 2017) e disputas de poder internas (SALGADO, 2017). A licença adotada foi o *Copyright* - todos os direitos reservados, mas a publicação oficial pelo governo e a guarda do relatório pelo Arquivo Nacional justificaram sua difusão com licença livre.

Os três volumes do relatório da CNV foram disponibilizados em versão eletrônica, em *Portable Document Format* (PDF), no site da própria comissão e outros meios. O terceiro volume, sobre mortos e desaparecidos, tem 1.996 páginas, com fichas individuais para cada caso apreciado pela CNV. As fichas são compostas de um cabeçalho biográfico estruturado, com dados sobre o nascimento, a atuação política e profissional e a morte ou desaparecimento, uma seção biográfica narrativa, uma avaliação das investigações sobre o caso até a criação da CNV, os resultados da investigação da comissão sobre a morte ou o desaparecimento, a identificação dos agentes de repressão responsáveis pela morte ou pelo desaparecimento e as fontes documentais usadas pela CNV, além de uma parte de conclusões e recomendações da CNV. Não houve a criação de um portal oficial sobre mortos e

Revista ACB: Biblioteconomia em Santa Catarina, Florianópolis (Brasil) - ISSN 1414-0594

desaparecidos pela CNV; para a análise do relatório, foi realizada uma transposição do documento para uma tabela.

3 CONVERGÊNCIA DIGITAL E WIKIDATA

As informações estruturadas das bases CEMDP, DP, MD e CNV foram transpostas integralmente para o projeto Wikidata. O objetivo dessa transposição foi estabelecer um parâmetro geral, interoperável, para unificar as informações das fontes diversas e assim mais facilmente compará-las. Nesta seção, apresentamos o Wikidata, a estratégia de transposição das bases para esse projeto e a modelagem dos itens de mortos e desaparecidos na ditadura civil-militar.

O Wikidata é um repositório colaborativo de dados estruturados e ligados entre si. Nesse projeto, todas as unidades de informação são normalmente classificadas como itens (denominados no Wikidata por QID, *Wikidata Q Identifier*) e descritas a partir de propriedades às quais são atribuídos valores. Esses valores podem ser acompanhados de qualificadores, que tornam a informação computada mais adequada. As classificações e descrições são multilíngues (KAFFEE et al., 2017). A maior parte das informações contidas no Wikidata é conectada entre si, tornando este um dos mais ambiciosos experimentos de web semântica em funcionamento (LUZ; CONEGLIAN; SANTAREM SEGUNDO, 2019). Funciona sobre a tecnologia e a metodologia wiki e adota uma licença livre *Creative Commons 0 1.0 Universal* (CC0 1.0). O repositório é aberto e qualquer pessoa pode editá-lo, incluindo a ontologia.

A transposição das informações sobre mortos e desaparecidos na ditadura civil-militar brasileira do Wikidata para a Wikipédia agrega a esforços sistemáticos de melhorar conteúdo sobre esse período da história na enciclopédia eletrônica. Pelo menos um projeto de educação foi realizado para melhorar conteúdo sobre mortos e desaparecidos na Wikipédia em português (MORAES et al., 2016). Wikiprojetos, isto é, iniciativas organizadas pela comunidade de wikipedistas, eventualmente com apoios externos, buscaram coordenar forças-tarefas de revisão e criação de artigos, especialmente o Wikiprojeto Inter-wikis do Arquivo Nacional (WIKIPÉDIA:GLAM/ARQUIVO NACIONAL, 2019). Criado em 14 de maio de 2014, o artigo *Lista de mortos e desaparecidos políticos na ditadura militar brasileira* passou de uma simples coleção de nomes (LISTA DE MORTOS E DESAPARECIDOS POLÍTICOS

Revista ACB: Biblioteconomia em Santa Catarina, Florianópolis (Brasil) - ISSN 1414-0594

NA DITADURA MILITAR BRASILEIRA, 2014) para uma transposição de informações estruturadas a partir do Wikidata (LISTA DE MORTOS E DESAPARECIDOS POLÍTICOS NA DITADURA MILITAR BRASILEIRA, 2020), inserindo para cada um dos casos listados referências específicas às páginas CEMDP, DP, MD e CNV, entre outras bases e fontes de informação. Na Wikipédia em português, a lista teve 186 mil acessos de 2015 a 2020 (início da série histórica registrada) (VISUALIZAÇÕES DA PÁGINA, 2020).

Um item no Wikidata é designado por um QID, a letra Q seguida de uma sequência numérica, e tem quatro partes. A primeira é conhecida como caixa de descrição, onde estão colocadas informações mínimas de identificação do item: rótulo (no caso, nome), descrição e outros nomes (*alias*). Essas seções podem ser inseridas em todos os idiomas disponíveis no Wikidata e não são clicáveis. A segunda parte é intitulada Declarações, onde há uma espécie de formulário, com propriedades e declarações. As propriedades são as categorias mais gerais de informações e as declarações são os valores que descrevem o item de acordo com a propriedade especificada. As declarações podem ser qualificadas, isto é, descritas com mais precisão a partir da utilização de subpropriedades chamadas *qualificadores* e referenciadas, com a indicação de proveniência do valor computado. Tanto propriedades quanto declarações são geralmente estruturadas, isto é, têm elas mesmas uma ficha no Wikidata. A terceira seção traz propriedades especiais, conhecidas como *identificadores*, que geralmente funciona como uma ligação de correspondência entre o item no Wikidata e páginas estruturadas para esse mesmo item em outras bases de dados. A última seção lista a ocorrência de páginas ou categorias atreladas a esse item nos vários projetos Wikimedia em suas variadas vertentes idiomáticas.

A transposição das informações nas bases da CEMDP, DP, MD e CNV para o Wikidata foi realizada em pelo menos três movimentos. O primeiro movimento é orgânico à atuação editorial voluntária no Wikidata, não relacionado direta e intencionalmente ao projeto de unificação sistemática dos dados das quatro bases. Trata-se de um movimento contínuo, na medida em que o Wikidata é um projeto colaborativo, permanentemente atualizado. Nesse movimento, houve a inclusão de informações provenientes das páginas e documentos nas bases sobre mortos e desaparecidos na ditadura, na medida em que estas são consideradas referências confiáveis. O segundo movimento, agora diretamente relacionado à proposta de espelhar sistematicamente as informações na CEMDP, DP, MD e CNV com o Wikidata, foi a raspagem de informações das bases, com o uso das bibliotecas de Python, particularmente a

Revista ACB: Biblioteconomia em Santa Catarina, Florianópolis (Brasil) - ISSN 1414-0594

Beautiful Soup (MITCHELL, 2018) e a *PyPDF2* (KULKARNI; SHIVANANDA, 2019), então transpostas com o robô de migração de dados QuickStatements (PELLISSIER TANON et al., 2016). A fusão de dados deu-se por processos de reconciliação, quando eram inconsistências simples. A curadoria necessária para a solução de dados conflitantes e a checagem de informações foram realizadas num terceiro movimento, aplicando uma metodologia de *crowdsourcing* (ESTELLÉS-AROLAS; GONZÁLEZ-LADRÓN-DE-GUEVARA, 2012): no contexto de um projeto optativo, estudantes de Jornalismo da Faculdade Cásper Líbero atuaram na verificação e melhoria dos conteúdos inseridos automaticamente a partir das bases, principalmente avaliando criticamente a relevância das declarações quando eram discrepantes, por exemplo, olhando documentos primários dos mortos e desaparecidos disponibilizados pela CNV para dados de filiação e nascimento (PESCHANSKI, 2019).

Para tornar mais eficiente a conexão entre o Wikidata e as várias bases de dados das quais informações foram extraídas foram criadas propriedades específicas sobre essas bases, chamadas identificadores. Assim, a MD tornou-se a P6673, a DP, a P6674 e a CEMDP, a P6692. No caso da CNV, por não haver ligação digital possível, tendo em vista que o documento usado como base é a versão digital do relatório impresso, foi usada a propriedade *descrito pela fonte* (P1343), inserindo a página correspondente ao caso no documento da CNV.

Vale notar que as informações no Wikidata superam muitas vezes o disponível nas bases de dados, na medida em que informações extraídas para os itens provêm de fontes múltiplas. Há também vários casos em que constam outros identificadores.

4 DIMENSÕES COMPARATIVAS

A comparação entre as bases da CEMDP, DP, MD e CNV é realizada a partir de três dimensões: completude, consistência e correção. São dimensões ditas internas, isto é, no nível da modelagem e operacionalização, e relacionadas aos dados. Abaixo, são apresentadas definições para essas dimensões (ANGELES; MACKINNON, 2004; SIDI et al., 2012):

- *Completude*: a medida em que os dados existem e, quando existem, são descritos sem que lhes faltem atributos e valores. A métrica de completude adotada para casos absolutos é: número de itens sobre o total de itens.

Revista ACB: Biblioteconomia em Santa Catarina, Florianópolis (Brasil) - ISSN 1414-0594

- *Consistência*: a medida em que os dados são apresentados no mesmo formato. A métrica de consistência adotada é: número de consistências sobre o total de verificações de consistência.
- *Correção*: a medida em que os dados são confiáveis. A métrica de correção adotada é: número de valores corretos sobre o total de verificações de correção.

A análise de completude na comparação das bases de mortos e desaparecidos na ditadura civil-militar brasileira leva em conta o número de casos agregados na base e o conjunto de propriedades associadas a cada caso na base. A análise de consistência verifica a modelagem das informações, especialmente de acordo com procedimentos de reconciliação e etiquetagem. Na dimensão de correção, analisa-se a confiabilidade dos dados inseridos nas bases.

É importante salientar que nenhuma das bases analisadas é propriamente semântica. Apesar de haver seções estruturadas nas fichas, estas foram preenchidas sem normalização, tornando-as pouco eficazes para buscas. Aliás, não há ferramenta de busca que permita, por exemplo, agregar informações sobre diferentes casos. Há também, nas seções narrativas, não levadas em conta para a comparação, confusões entre dados biográficos, considerações mais analíticas sobre as condições da morte ou desaparecimento, informações sobre a investigação e detalhes administrativos, como a data do recolhimento da documentação física para o Arquivo Nacional.

A verificação dos valores das propriedades das quatro bases em relação ao Wikidata foi realizada através de comandos automatizados e curadoria humana.

4.1 COMPLETUDE

Nesta seção, as quatro bases sobre mortos e desaparecidos na ditadura civil-militar brasileira são descritas e avaliadas de acordo com o número de casos que agregam e, para os casos, o número de declarações e atributos que são computados. O gráfico 1 e o diagrama 1 ilustram, respectivamente, o número de casos agregados em cada base e a distribuição e as sobreposições de casos agregados nas bases. A tabela 1 compara as declarações e atributos nas bases.

A lista de fichas descritivas da CEMDP constituía-se oficialmente de 362 nomes, mas apenas 349 possuíam uma página correspondente na base de dados. Desses 349 nomes, 1

possuía uma ficha descritiva vazia, 5 possuíam somente links quebrados para a base de dados e outros 5 correspondiam a pessoas mortas durante o Massacre de Ipatinga, em 7 de outubro de 1963, portanto antes do golpe militar de 1964 e fora do escopo deste artigo. Dessa forma, para a análise da base da CEMDP serão considerados somente os 338 nomes que possuem links ativos, isto é, somente os indivíduos que possuem uma página com uma ficha descritiva com alguma informação.

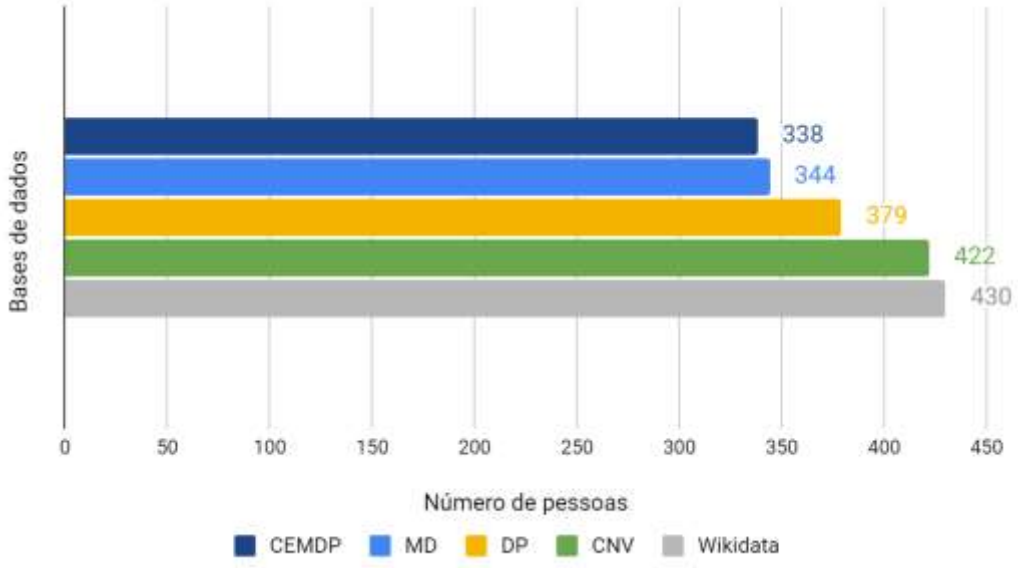
Na página de apresentação do site Desaparecidos Políticos é declarado que a base de dados do site constitui-se de 383 nomes de pessoas mortas ou desaparecidas durante a ditadura militar. Porém, tanto na página da lista desses nomes quanto no sistema de busca do site, há apenas 379 registros. Através das funcionalidades do *Wayback Machine*, foi possível resgatar um *snapshot* da página da lista datando de 28 de junho de 2008 (BRASIL, 2008) e mesmo então a página só contava com 380 nomes, tendo um único nome sido retirado por não ser considerado uma vítima, mas um agente infiltrado do Estado. Portanto, a análise dos dados do site Desaparecidos Políticos considerou 379 nomes.

A base de dados do site Memórias da Ditadura possuía 430 nomes, dos quais 13 eram duplicatas de outros nomes na lista e 12 correspondiam a pessoas mortas antes do golpe militar de 1964. Dos 405 nomes restantes, 31 possuíam na ficha descritiva apenas uma imagem e outros 30 possuíam fichas descritivas vazias. Para os procedimentos adotados na análise que este artigo se propõe a fazer, serão considerados somente os 344 nomes que possuíam alguma informação na ficha descritiva além da imagem.

O relatório final da Comissão da Verdade registra em seu terceiro volume 434 nomes de pessoas mortas e desaparecidas durante a ditadura militar, dos quais 12 correspondem a pessoas mortas antes do golpe de 1964, logo estão fora do escopo deste artigo. Para a análise dos dados estruturados da CNV, serão considerados 422 nomes. O relatório também cita 8 pessoas que não tiveram seus nomes incluídos por não ter sido possível caracterizar a responsabilidade do Estado por essas mortes.

Revista ACB: Biblioteconomia em Santa Catarina, Florianópolis (Brasil) - ISSN 1414-0594

Gráfico 1 - Comparação do número de mortos e desaparecidos na ditadura civil-militar brasileira (1964-1985), em cinco bases de dados



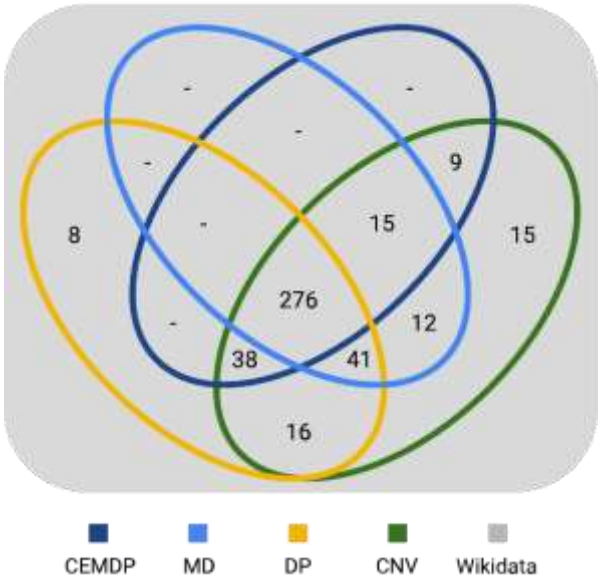
Fonte: elaborado pelos autores (2020)

O diagrama 1 mostra a distribuição das pessoas mortas e desaparecidas em cada base de dados analisada. Cada base de dados (CEMDP, MD, DP e CNV) é representada por uma elipse colorida e o somatório dos números dentro de uma elipse é equivalente ao total de casos no escopo deste artigo que estão registrados na base que a elipse representa. As áreas em que as elipses se sobrepõem representam os totais de casos comuns às bases representadas naquela sobreposição. Por exemplo, as bases de dados CEMDP e CNV possuem 9 nomes que só aparecem nessas duas bases, assim como a base DP possui 8 nomes que não constam em nenhuma das outras bases. A área do quadrado cinza representa a base de dados do Wikidata, que possui registro de cada uma das pessoas mortas e desaparecidas que estão presentes nas outras bases.

Há variados motivos para a inclusão ou exclusão de casos nas bases. Os quatro repositórios adotaram metodologias distintas, o que leva a diferenças não apenas numéricas, mas também de tipos de casos considerados (o que é evidenciado pelo diagrama de sobreposição). A avaliação da completude também leva em conta o número de declarações associadas a cada caso.

Revista ACB: Biblioteconomia em Santa Catarina, Florianópolis (Brasil) - ISSN 1414-0594

Diagrama 1 - Distribuição do número de mortos e desaparecidos na ditadura civil-militar brasileira (1964-1985), em cinco bases de dados



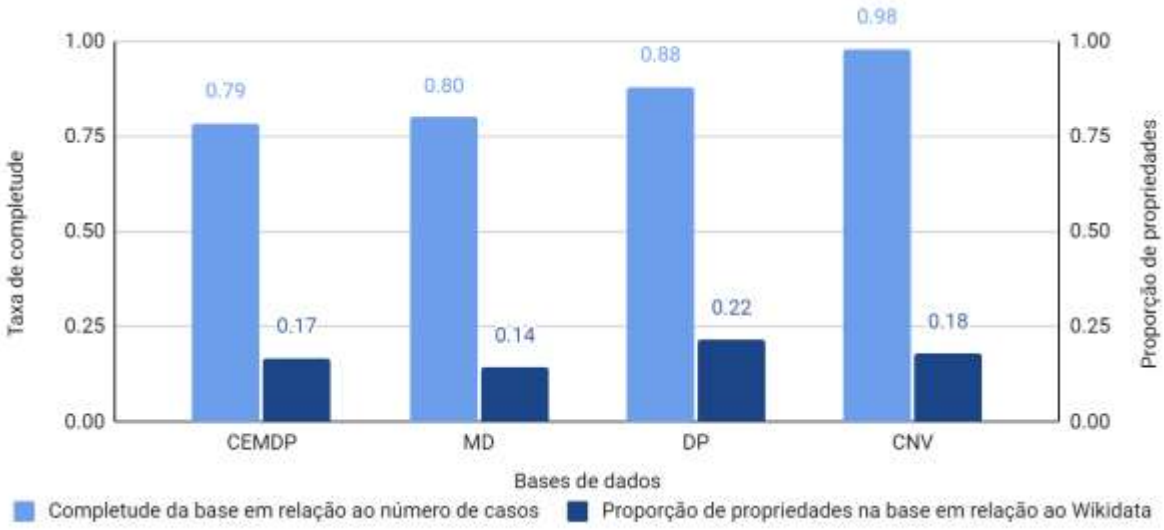
Fonte: elaborado pelos autores (2020)

As bases de dados analisadas organizam suas informações utilizando as propriedades por elas definidas. Essa singularidade na modelagem pode levar a erros de interpretação da informação apresentada se a propriedade não estiver bem definida e dificulta a comparação com propriedades de outras bases. Um exemplo para esse caso é a propriedade *Prisão* da base DP, que delimita valores tanto para a data quanto para o local de detenção de uma pessoa, porém de forma bem estruturada e consistente, separando claramente data e local, embora no mesmo campo. Para a análise deste artigo, porém, essas propriedades com definições ou valores compostos de mais de uma informação foram separadas ou agrupadas em propriedades bem definidas e sem sobreposição de valores. Logo, por exemplo, a propriedade mencionada anteriormente foi dividida entre *Data de detenção* e *Local de detenção*. Portanto, daqui em diante tratamos por *propriedades na base* os tipos de informações distintas que obtemos de cada campo de ficha descritiva de uma base.

O gráfico 2 retrata as taxas de completude das bases CEMDP, MD, DP e CNV em relação ao número de casos no Wikidata e a proporção de propriedades nessas bases em relação às 90 propriedades (ALVES, 2020) associadas a esses casos no Wikidata.

Revista ACB: Biblioteconomia em Santa Catarina, Florianópolis (Brasil) - ISSN 1414-0594

Gráfico 2 - Completude em relação ao número de casos no Wikidata e proporção de propriedades na base em relação ao número de propriedades no Wikidata



Fonte: elaborado pelos autores (2020)

A proporção de propriedades nas bases em relação ao Wikidata indica a preocupação na modelagem dos casos analisados. Vê-se que o relatório da CNV, apesar de ter o maior número de casos em comparação com os 430 mortos e desaparecidos inseridos no Wikidata, mantém uma proporção mediana no gráfico no que diz respeito às propriedades estruturadas nas outras bases. As informações nesse relatório estão principalmente num formato narrativo, o que dificulta por exemplo a comparação computacional de características dos casos sob análise, o que também acontece na base MD, com o agravante de declarar menos propriedades em sua ficha. A base da DP é a que mais se preocupa em ter uma ficha descritiva completa, apesar de o número de propriedades ser ainda relativamente parecido com o das outras bases e bem abaixo do número no Wikidata.

A tabela 1 avalia o grau de completude das propriedades nas bases que possuem uma propriedade correspondente no Wikidata. Cada célula da tabela é calculada como o quociente entre o número de casos com a informação preenchida na base e o número de casos com a informação preenchida no Wikidata. Foram considerados preenchidos todos os campos em que alguma informação foi passada, portanto, valores como *não informado*, *desconhecido*, *n/i* etc. também fazem parte da contagem. O que é diferente de campos das bases que apareciam em branco ou mesmo não eram exibidos por estarem vazios. Abaixo há uma descrição matemática da fórmula usada para calcular a taxa de completude

Revista ACB: Biblioteconomia em Santa Catarina, Florianópolis (Brasil) - ISSN 1414-0594

Sejam *p* uma das propriedades da tabela, *B* o conjunto de casos em uma das bases (B ∈ {CEMDP, MD, DP, CNV}) e *W* o conjunto de casos no Wikidata. A taxa de completude TC para cada propriedade em cada base é calculada com a fórmula:

$$TC(B, p) = \frac{|B_p|}{|B \cap W_p|}$$

Onde $|B_p|$ é a cardinalidade do conjunto de casos com a propriedade *p* preenchida dentro da base *B* e $|B \cap W_p|$ é a cardinalidade do conjunto de casos com a propriedade *p* preenchida no Wikidata em interseção com os casos da base *B*.

Tabela 1 - Completude de propriedades nas bases, em comparação com o Wikidata

Propriedade	**CEMDP**	**MD**	**DP**	**CNV**
Nome	**1**	**1**	**1**	**1**
Imagem	0,8679	0,9836	0,4064	**1**
País de cidadania	0	0	**0,6966**	0
Data de nascimento	**0,9970**	0,9535	0,6201	0,9905
Local de nascimento	**0,9911**	0,9622	0,6807	0,9882
Data de desaparecimento	0,0601	0,5243	0,7463	**0,9688**
Local de desaparecimento	0,7322	0,5135	0,7610	**0,9330**
Idade quando desaparecido	**1**	0	0	0
Data de morte	0,0059	**0,7006**	0,5752	0,4882
Local de morte	0,4467	**0,6890**	0,5805	0,4882
Data de detenção	0	0	**1**	0
Local de detenção	0	0	**1**	0
Codinome	0,3389	0	**0,8667**	0,0051
Mãe	0,9941	0,9593	0	**1**
Pai	0,9882	0,9593	0	**1**
Organização política	0,7899	0,8960	0,7889	**0,9858**
Atuação profissional	0,0444	0,9738	0,7678	**1**
Universidade	0	0	**0,9627**	0
Médico legista	0	0	**0,9732**	0,4966

Revista ACB: Biblioteconomia em Santa Catarina, Florianópolis (Brasil) - ISSN 1414-0594

Agente da repressão	0	0	0,2343	**1**
Órgãos de repressão	-	-	-	-
Status	-	-	-	-
Taxa de completude absoluta	**0,5680**	**0,6929**	**0,6077**	**0,7343**

Fonte: elaborada pelos autores (2020)

As completudes absolutas das bases da CEMDP, MD, DP e CNV são, respectivamente: 0.5680, 0.6929, 0.6077 e 0.7343 e foram calculadas como uma proporção entre o número de propriedades com valores preenchidos na base e o número de propriedades com valores preenchidos no Wikidata em interseção com a respectiva base. Apenas a propriedade Nome tem o mesmo grau de completude nas quatro bases, por sinal idêntico ao Wikidata. As propriedades em que a CEMDP apresenta o maior grau de completude são: Data de nascimento, Local de nascimento e Idade quando desaparecido. Note-se que esta última propriedade, facilmente calculada a partir da data de nascimento e a de desaparecimento, só aparece na CEMDP. As propriedades em que a base do MD apresenta o maior grau de completude são: Data de morte e Local de morte. Isso deve-se à representação dos casos dessa base: mais de 70% dos casos são sobre mortos políticos, enquanto nas bases CEMDP, DP e CNV essa proporção cai para 15.38%, 58.76% e 48.7%, respectivamente. As propriedades em que a base do DP figura com o mais elevado nível de completude são: País de Cidadania, Data de detenção, Local de detenção, Codinome, Universidade e Médico legista. Não está claro como a primeira dessas propriedades foi construída, eventualmente pode ter sido deduzida da informação sobre o local de nascimento. As outras propriedades são praticamente exclusivas da base do DP. Vale notar a taxa de completude da categoria Codinome, que provavelmente dependeu de entrevistas e testemunhos de familiares, amigos ou militantes.

A base da CNV lidera na taxa de completude nas seguintes propriedades: Imagem, Data de desaparecimento, Local de desaparecimento, Pai, Mãe, Organização política, Atuação profissional e Agente da repressão. De certo modo, o relatório da CNV apresenta praticamente uma carteira de identidade digital relativamente completa dos mortos e desaparecidos políticos. Note-se que nas propriedades Imagem, Mãe, Pai, Atuação profissional e Agente da repressão a completude é completa, em comparação com as informações no Wikidata. Em particular, a propriedade Agente da repressão no Wikidata lista

Revista ACB: Biblioteconomia em Santa Catarina, Florianópolis (Brasil) - ISSN 1414-0594

toda a cadeia de comando das pessoas responsáveis pela morte ou desaparecimento das vítimas.

As propriedades Órgãos de repressão e Status, das bases DP e CEMDP, que identificam as instituições que tiveram participação na morte ou desaparecimento das vítimas e se uma pessoa é considerada desaparecida política ou morta política, respectivamente, não têm propriedade correspondente no Wikidata. A base DP possui 222 casos com a propriedade Órgãos de repressão preenchida e todos os casos da base apresentam o Status jurídico das vítimas. A base CEMDP possui 336 casos de 338 com o Status jurídico preenchido. A princípio, a propriedade Status no Wikidata pode ser deduzida a partir das propriedades relacionadas a desaparecimentos na razão com o total de casos. Órgãos de repressão e Status, apesar de listadas na tabela, não foram consideradas para o cálculo de completude neste artigo.

4.2 CONSISTÊNCIA

A comparação das bases sobre mortos e desaparecidos na ditadura civil-militar brasileira leva aqui em conta a consistência das informações compiladas. Essa dimensão comparativa, relativa à representação das informações, é especialmente importante para análises agregadas, para a qual a reconciliação de valores é esperada.

O método de comparação baseia-se na verificação de consistência de cada valor na propriedade em relação à formatação dos demais valores daquela propriedade na base. Isso resulta em medidas de comparação de uniformidade interna de cada base. Apesar de uma mesma informação poder ser apresentada de formas distintas em outras bases, a comparação ocorre em relação aos formatos dos valores da própria base. Por exemplo, se uma das bases apresenta os municípios de nascimento no formato *Município (Sigla do Estado)*, um valor como *Vitória da Conquista (BA)* será consistente, enquanto *Vitória da Conquista-BA* ou *Vitória da Conquista (Bahia)*, não.

Em propriedades das bases que apresentam múltiplos valores preenchidos, a consistência foi checada para cada valor individualmente, de modo que foram ignoradas as inconsistências no uso de separadores de valores adotados pelas bases, como *"ou"*, *"e"*, *";"* e *"/"*, etc. Por exemplo, para um indivíduo que possua *PCB; Vanguarda Armada Revolucionária (VAR-Palmares)* no campo *Organização política*, são checadas as

consistências dos valores *PCB* e *Vanguarda Armada Revolucionária (VAR-Palmares)* separadamente.

A tabela 2 apresenta os valores de consistência das propriedades das bases CEMDP, MD, DP e CNV como uma proporção entre a quantidade de valores consistentes e a quantidade total de valores, definida em Batini et al. (2009).

Tabela 2 - Consistência dos valores apresentados para as propriedades nas bases

Propriedade	**CEMDP**	**MD**	**DP**	**CNV**
Nome	0,9970	0,9884	0,9815	**0,9976**
Imagem	**1**	**1**	**1**	**1**
País de cidadania	-	-	**1**	-
Data de nascimento	**0,9911**	0,9789	0,9660	0,9905
Local de nascimento	0,9284	0,9759	**1**	0,9689
Data de desaparecimento	0,8462	0,9535	**1**	0,9123
Local de desaparecimento	0,8955	0,7	**0,9158**	0,6628
Idade quando desaparecido	**0,9806**	-	-	-
Data de morte	**1**	0,9	**1**	0,9628
Local de morte	0,8947	0,8158	**1**	0,9147
Data de detenção	-	-	**1**	-
Local de detenção	-	-	**1**	-
Codinome	**1**	-	**1**	**1**
Mãe	**0,994**	0,9699	-	0,9623
Pai	**0,994**	0,9699	-	0,9646
Organização política	0,9662	0,9387	**0,9975**	0,9740
Atuação profissional	**1**	**1**	**1**	**1**
Universidade	-	-	**1**	-
Médico legista	-	-	**1**	**1**
Agente da repressão	-	-	**1**	**1**
Órgãos de repressão	-	-	**1**	-
Status	**1**	-	**1**	-
Taxa de consistência absoluta	**0,9762**	**0,9487**	**0,9939**	**0,9734**

Fonte: elaborada pelos autores (2020)

Revista ACB: Biblioteconomia em Santa Catarina, Florianópolis (Brasil) - ISSN 1414-0594

A base CEMDP apresenta as maiores taxas de consistência nas propriedades Imagem, Data de morte, Codinome e Atuação profissional, todas com grau máximo de consistência, e também nas propriedades Data de nascimento, Mãe, Pai e Idade quando desaparecido, sendo a última exclusiva dessa base. As propriedades da CEMDP mais inconsistentes são respectivamente Data de desaparecimento, Local de morte e Local de desaparecimento, sendo a primeira taxa decorrente da baixa quantidade de valores, o que aumenta significativamente o peso de toda e qualquer inconsistência.

Já para a base MD, somente as propriedades Imagem e Atuação Profissional apresentam o grau máximo de consistência. Apesar da maioria das taxas estarem acima de 0,9, a MD é a base com a menor taxa absoluta de consistência dentre as quatro bases analisadas.

A base de dados DP possui as maiores taxas de consistência das propriedades. Isso só não é verificado em Nome e Data de nascimento, que apresentam maior inconsistência que as demais bases. País de cidadania, Universidade, Órgãos de repressão, Data de detenção e Local de detenção aparecem somente na DP, tendo todas grau máximo de consistência, assim como todas as demais propriedades, exceto Nome, Data de nascimento, Local de desaparecimento e Organização Política.

A base da CNV possui as maiores taxas de consistência em Imagem, Codinome, Atuação profissional, Médico Legista, Agente da repressão e Nome, apresentando o grau máximo de consistência em todas, exceto na última. As propriedades da CNV com as menores taxas são: Local de desaparecimento e Data de desaparecimento.

Considerando-se as taxas de consistência absoluta, definidas como a proporção entre o número total de valores consistentes e o número total de valores, define-se que a ordem de consistência entre as bases é, da mais consistente à menos consistente: DP, CEMDP, CNV e MD, com taxas de consistência de 0,9939, 0,9762, 0,9734 e 0,9487, respectivamente. Nota-se que as duas primeiras são bases organizadas de forma tabular, ou seja, mais estruturadas, enquanto as duas últimas são organizadas de forma mais textual e, portanto, menos estruturadas.

Os valores das taxas de consistência individuais para cada propriedade não necessariamente têm peso proporcional na taxa de consistência absoluta, pois esta última é o somatório do número total de valores consistentes em todas as propriedades sobre o somatório do número total de valores preenchidos em todas as propriedades. Portanto, uma propriedade

Revista ACB: Biblioteconomia em Santa Catarina, Florianópolis (Brasil) - ISSN 1414-0594

com poucos valores preenchidos pode apresentar uma taxa de consistência alta e contribuir pouco para o somatório (como Data de morte, na CEMDP), enquanto uma propriedade com muitos valores preenchidos pode ter uma taxa de consistência relativamente menor e contribuir bastante para o somatório (como Organização política, na CEMDP). Como exemplo dessa influência, pode-se verificar a taxa de consistência absoluta da base CNV, que tem muita influência dos milhares de valores declarados para agentes da repressão envolvidos nas mortes e desaparecimento das vítimas; sem esses casos, o valor seria 0,956.

4.3 CORREÇÃO

A terceira dimensão de análise deste artigo aborda a correção dos valores inseridos nas bases sobre os mortos e desaparecidos na ditadura civil-militar brasileira, isto é, avalia em que medida os dados são confiáveis e certificados, além de representar corretamente a realidade.

O método de comparação desta dimensão baseia-se na verificação da confiabilidade de que cada valor declarado em uma propriedade representa a realidade daquela informação. Algumas propriedades são mais controversas que outras, principalmente as que permeiam a morte e o desaparecimento das vítimas, pois corriqueiramente há mais de um valor declarado para essas propriedades devido às versões e depoimentos presentes nos processos judiciais. Para esta análise, o Wikidata foi o objeto considerado para representar a realidade.

Propriedades das bases que agregam mais de um campo de informação, como Data e local de nascimento agregando Data de nascimento e Local de nascimento foram separadas antes da análise. Como na análise da dimensão da consistência, as inconsistências decorrentes dos separadores de valores que as bases adotaram foram ignoradas, assim como valores vazios ou que só apresentassem pontos ou outras pontuações em que não fosse possível extrair informação alguma.

A tabela 3 apresenta os valores de correção das propriedades das bases CEMDP, MD, DP e CNV como uma proporção entre a quantidade de valores corretos para a propriedade na base e a quantidade total de declarações feitas para aquela propriedade na base, definida em Batini et al. (2009).

Revista ACB: Biblioteconomia em Santa Catarina, Florianópolis (Brasil) - ISSN 1414-0594

Tabela 3 - Correção dos valores apresentados para as propriedades nas bases

Propriedade	**CEMDP**	**MD**	**DP**	**CNV**
Nome	**1**	0,9884	0,9815	**1**
Imagem	-	-	-	-
País de cidadania	-	-	**1**	-
Data de nascimento	0,9644	**0,9909**	0,8255	0,9833
Local de nascimento	0,9881	**0,9967**	0,9651	0,9952
Data de desaparecimento	0,6923	**0,8760**	0,7744	0,8669
Local de desaparecimento	0,9328	**1**	0,9263	0,9962
Idade quando desaparecido	**1**	-	-	-
Data de morte	**1**	**1**	0,8885	**1**
Local de morte	0,9605	0,9962	**0,9963**	0,9905
Data de detenção	-	-	**1**	-
Local de detenção	-	-	**0,9778**	-
Codinome	**1**	-	**1**	**1**
Mãe	0,9940	**1**	-	**1**
Pai	0,9940	**0,9669**	-	0,9623
Organização política	0,9737	0,9908	**0,9975**	0,9870
Atuação profissional	**1**	**1**	**1**	**1**
Universidade	-	-	**1**	-
Médico legista	-	-	**1**	**1**
Agente da repressão	-	-	**1**	**1**
Órgãos de repressão	-	-	-	-
Status	-	-	-	-
Taxa de correção absoluta	**0,9822**	**0,9879**	**0,9653**	**0,9896**

Fonte: elaborada pelos autores (2020)

A base CEMDP apresenta as maiores taxas de correção nas propriedades Nome, Idade quando desaparecido, Data de morte, Codinome e Atuação profissional, todas com grau máximo de correção. A propriedade da CEMDP com a menor taxa de correção é a Data de desaparecimento, sendo decorrente da baixa quantidade de valores declarados, fazendo com

Revista ACB: Biblioteconomia em Santa Catarina, Florianópolis (Brasil) - ISSN 1414-0594

que um único valor incorreto tenha muito mais impacto na taxa do que em outras propriedades.

Para a base MD, todas as taxas de correção, com exceção às propriedades Nome, Local de morte e Organização política, são as maiores dentre as demais bases. Dessas, Data de nascimento, Local de Nascimento, Data de desaparecimento e Pai são as propriedades sem taxas de correção máximas. A CNV e a MD são as bases com as maiores taxas de correção absolutas dentre as quatro bases analisadas, respectivamente com taxas de 0,9896 e 0,9879.

A base DP apresenta as maiores taxas de correção nas propriedades Local de morte, Codinome, Organização política e Atuação profissional, Médico legista e Agente da repressão, além das propriedades que são exclusivas dessa base: País de cidadania, Data de detenção, Local de detenção e Universidade. As propriedades da DP que apresentam as menores taxas são Data de desaparecimento e Nome.

A base da CNV possui altas taxas de correção para a maior parte das propriedades, alcançando o grau máximo em Nome, Data de morte, Codinome, Mãe, Atuação profissional, Médico legista e Agente da repressão. As restantes variam entre si, e exceto a propriedade Data de desaparecimento, que tem taxas relativamente baixas em todas as bases, as taxas da CNV permanecem sempre acima de 0,96.

As propriedades Imagem, Órgãos de repressão e Status não estão representadas na tabela por diferentes motivos. As duas últimas não possuem propriedade equivalente no Wikidata e por isso não podem ser comparadas. A propriedade Imagem, nessa dimensão, não foi avaliada, pois em muitos casos as bases usam fotografias distintas de diferentes períodos da vida das vítimas ou em baixa qualidade.

Considerando-se as taxas de correção absoluta, definidas como a proporção entre o número total de valores corretos e o número total de valores declarados, define-se que a ordem de correção entre as bases é: CNV com 0,9896, seguida de MD com 0,9879, CEMDP e DP, com taxas de correção de 0,9822 e 0,9653 respectivamente. Ao contrário do resultado da análise de consistência das bases, as bases com maior correção são as duas organizadas de forma mais textual.

Revista ACB: Biblioteconomia em Santa Catarina, Florianópolis (Brasil) - ISSN 1414-0594

5 CONSIDERAÇÕES FINAIS

Este artigo comparou a completude, consistência e correção de quatro bases digitais sobre mortos e desaparecidos na ditadura civil-militar brasileira, a da Comissão Especial sobre Mortos e Desaparecidos Políticos, a do site Desaparecidos Políticos, a do projeto Memórias da Ditadura e a do relatório final da Comissão Nacional da Verdade. Entendeu-se completude como uma medida em relação à presença e amplitude dos dados. A consistência foi medida de acordo com a uniformidade de formato dos dados. Correção disse respeito à confiabilidade dos dados.

O objetivo principal foi a comparação da qualidade dessas bases. A partir dos dados coletados, percebeu-se que há variações em cada propriedade analisada numa mesma base. No geral, as taxas de completude da CEMDP, MD, DP e CNV foram respectivamente 0,5680, 0,6929, 0,6077 e 0,7343. No que diz respeito à consistência, as medidas gerais foram, respectivamente para a CEMDP, MD, DP e CNV, 0,9762, 0,9487, 0,9939 e 0,9734. Na medida de correção, o que se reportou foram as seguintes medidas gerais: CEMDP, 0,9822; MD, 0,9879; DP, 0,9653; e CNV, 0,9896.

Um objetivo complementar foi a criação de um sistema em que convergissem as informações das várias fontes, criando um elo entre as bases heterogêneas de mortos e desaparecidos. Utilizou-se o Wikidata, que agora agrega os dados referenciados de todas as bases, dando-lhes uma característica semântica.

A sistematização das informações sobre mortos e desaparecidos na ditadura no Wikidata tem diversos impactos. Um primeiro, relatado no artigo, foi a produção de um material de alta qualidade na Wikipédia em português, ampliando a difusão da informação nas bases. Um segundo impacto potencial é que, com a estruturação dos dados numa plataforma semântica, há possibilidades jamais antes exploradas de visualização das informações.

AGRADECIMENTOS

Agradecemos à equipe de difusão científica do CEPID NeuroMat por ter feito apontamentos importantes durante o seminário no qual esta pesquisa foi apresentada. Agradecemos a Kelly Braghetto a leitura crítica da primeira versão deste artigo.

FINANCIAMENTO

A pesquisa de João Alexandre Peschanski integra o projeto FAPESP 2013/07699-0 e tem apoio do Centro Interdisciplinar de Pesquisa da Faculdade Cásper Líbero.

DISPONIBILIDADE DE DADOS

Todos os dados que baseiam as análises e conclusões deste artigo estão tabelados e disponíveis em https://doi.org/10.5281/zenodo.3969786.

REFERÊNCIAS

ALVES, Éder Porto Ferreira. Propriedades Wikidata dos mortos e desaparecidos da ditadura civil-militar brasileira. **Wikidata Query Service**. [S.l.], 2020. Disponível em: https://w.wiki/Yw5. Acesso em: 25 fev. 2020.

ANGELES, María del Pilar; MACKINNON, Lachlan M. Detection and Resolution of Data Inconsistencies, and Data Integration using Data Quality Criteria. *In*: 5TH INTERNATIONAL CONFERENCE ON THE QUALITY OF INFORMATION AND COMMUNICATIONS TECHNOLOGY. 2004, Porto, Portugal. **Anais** [...]. Porto, Portugal: Universidade Portucalense, 2004. Disponível em: http://ceur-ws.org/Vol-1135/paper11.pdf. Acesso em: 24 jan. 2020.

AYDOS, Valéria; FIGUEIREDO, César Alessandro Sagrillo. A construção social das vítimas da ditadura militar e a sua ressignificação política. **Interseções: Revista de Estudos Interdisciplinares**, Rio de Janeiro, v. 15, n. 2, 2013. Disponível em: https://www.e-publicacoes.uerj.br/index.php/intersecoes/article/view/9521. Acesso em: 20 jun. 2021.

AZEVEDO, Desirée de Lemos. **"A única luta que se perde é aquela que se abandona": Etnografia entre familiares de mortos e desaparecidos políticos no Brasil**. Tese (Doutorado em Antropologia Social) – Instituto de Filosofia e Ciências Humanas, Universidade Estadual de Campinas, Campinas, 2016. Disponível em: https://hdl.handle.net/20.500.12733/1628236. Acesso em: 23 jan. 2020.

AZZELLINI, Érica Camillo; PESCHANSKI, João Alexandre; PAIXÃO, Fernando Jorge da. As potencialidades de narrativas estruturadas para o Jornalismo Computacional: competências jornalísticas na elaboração de textos gerados com bancos de dados. **Texto Livre: Linguagem e Tecnologia**, Belo Horizonte, v. 12, n. 1, pp. 138–152, 2019. Disponível em: https://periodicos.ufmg.br/index.php/textolivre/article/view/16837/13598. Acesso: 23 jan. 2020.

BATINI, Carlo; CAPPIELLO, Cinzia; FRANCALANCI, Chiara; MAURINO, Andrea. Methodologies for Data Quality Assessment and Improvement. **ACM Computing Surveys**, Nova Iorque, Estados Unidos da América, v. 41, n. 3, 2009. Disponível em:

https://doi.org/10.1145/1541880.1541883. Acesso em: 1 abr. 2020.

BERNASCONI, Oriana; RUIZ, Marcela; LIRA, Eizabeth. What defines the victims of human rights violations? The case of the Comité Pro Paz and Vicaría de la Solidaridad in Chile (1973-1992). *In*: DRULIOLLE, V.; BRETT, R. (eds.). **The Politics of Victimhood in Post-conflict Societies: Comparative and Analytical Perspectives**. Londres, Reino Unido: Palgrave Macmillan, p.101-131. (St. Antony's Series)

BERRY, David. The computational turn: thinking about the digital humanities. **Culture Machine**. Coventry, Reino Unido, v. 12, 2011. Disponível em: https://culturemachine.net/wp-content/uploads/2019/01/10-Computational-Turn-440-893-1-PB.pdf. Acesso em: 23 jan. 2020.

BRASIL. **Comissão Especial de Mortos e Desaparecidos Políticos**. Brasília: Secretaria de Direitos Humanos da Presidência da República, 2019. Disponível em: https://web.archive.org/web/20190820183211/http://cemdp.sdh.gov.br. Acesso em: 17 jan. 2020.

BRASIL. **Comissão Nacional da Verdade**. Brasília: Comissão Nacional da Verdade, 2014. Disponível em: http://cnv.memoriasreveladas.gov.br. Acesso em: 17 jan. 2020.

BRASIL. Lei nº 9.140, de 4 de dezembro de 1995. Reconhece como mortas pessoas desaparecidas em razão de participação, ou acusação de participação, em atividades políticas, no período de 2 de setembro de 1961 a 15 de agosto de 1979, e dá outras providências. **Diário Oficial da União**: seção 1, Brasília, ano 132, n. 232, p. 19985, 5 ago. 1995. Disponível em: http://www.planalto.gov.br/ccivil_03/leis/l9140.htm. Acesso em: 23 jan. 2020.

BRASIL. **Mortos e Desaparecidos Políticos**. Brasília: Secretaria de Direitos Humanos da Presidência da República, 2008. Disponível em: https://web.archive.org/web/20080628095531/http://www.desaparecidospoliticos.org.br/pessoas.php. Acesso em: 21 jan. 2020.

CANABARRO, Ivo. Caminhos da comissão nacional da verdade (CNV): memórias em construção. **Sequência (Florianópolis)**, Florianópolis, n. 69, p. 215-234, 2014. Disponível em: https://doi.org/10.5007/2177-7055.2014v35n69p215. Acesso em: 24 jan. 2020.

CENTRO DE DOCUMENTAÇÃO EREMIAS DELIZOICOV; DOSSIÊ MORTOS E DESAPARECIDOS POLÍTICOS NO BRASIL. **Mortos e Desaparecidos Políticos**. 2020. Disponível em: http://web.archive.org/web/20191231131345/http://www.desaparecidospoliticos.org.br. Acesso em: 17 jan. 2020.

COMISSÃO DE FAMILIARES DE MORTOS E DESAPARECIDOS POLÍTICOS; INSTITUTO DE ESTUDOS DA VIOLÊNCIA DO ESTADO; GRUPO TORTURA NUNCA MAIS - RJ; GRUPO TORTURA NUNCA MAIS - PE. **Dossiê dos mortos e desaparecidos políticos a partir de 1964**. Recife: Companhia Editora de Pernambuco, 1995, 444 p. Disponível em: http://www.dhnet.org.br/dados/dossiers/dh/br/dossie64/br/dossmdp.pdf.

Acesso em: 23 jan. 2020.

COSTA, Alessandra de Sá Mello da; SILVA, Marcelo Almeida de Carvalho. Empresas, violação dos direitos humanos e ditadura civil-militar brasileira: a perspectiva da Comissão Nacional da Verdade. **Organizações & Sociedade**, Salvador, v. 25, n. 84, p. 15-29, 2018. Disponível em: https://doi.org/10.1590/1984-9240841. Acesso em: 24 jan. 2020.

DRULIOLLE, Vincent. Recovering Historical Memory: A Struggle against Silence and Forgetting? The Politics of Victimhood in Spain. **International Journal of Transitional Justice**, Oxônia, Reino Unido, v. 9, n. 2, p. 316-335, 2015. Disponível em: https://doi.org/10.1093/ijtj/ijv008. Acesso em: 23 jan. 2020.

ESTELLÉS-AROLAS, Enrique; GONZÁLEZ-LADRÓN-DE-GUEVARA, Fernando. Towards an integrated crowdsourcing definition. **Journal of Information Science**, v. 38, n. 2, p. 189-200, 2012. Disponível em: https://doi.org/10.1177/0165551512437638. Acesso em: 7 fev. 2020.

FERNANDES, Pádua. Justiça de transição e o fundamento nos direitos humanos: perplexidades do relatório da Comissão Nacional da Verdade brasileira. *In*: KASHIURA JÚNIOR, C. N.; AKAMINE JÚNIOR, O.; MELO, T. M. (eds.), **Para a crítica do direito: reflexões sobre teorias e práticas jurídicas**. São Paulo: Dobra Universitário, 2015. p. 717-745. Disponível em: https://www.academia.edu/24092436/Justiça_de_transição_e_o_fundamento_nos_direitos_humanos_perplexidades_do_relatório_da_Comissão_Nacional_da_Verdade_brasileira. Acesso em: 24 jan. 2020.

FROTA, Maria Guiomar da Cunha. Memória e produção social da informação em direitos humanos: uma perspectiva latino-americana. **Perspectivas em Ciência da Informação**, v. 24, n. esp., p. 162-175, 2019. Disponível em: https://doi.org/10.1590/1981-5344/3900. Acesso em: 23 jan. 2020.

GATTI, Gabriel. **Un mundo de víctimas**. Barcelona, Espanha: Anthropos Editorial, 2017, 431 p.

HALL, Gary. Towards a post-digital humanities: cultural analytics and the computational turn to data-driven scholarship. **American Literature**, v. 85, n. 4, p. 781-809, 2013. Disponível em: https://pureportal.coventry.ac.uk/en/publications/towards-a-post-digital-humanities-cultural-analytics-and-the-comp-2. Acesso em: 23 jan. 2020.

INSTITUTO VLADIMIR HERZOG. **Memórias da ditadura**. 2020. Disponível em: http://memoriasdaditadura.org.br. Acesso em: 6 ago. 2020.

KAFFEE, Lucie-Aimée; PISCOPO, Alessandro; VOUGIOUKLIS, Pavlos; SIMPERL, Elena; CARR, Leslie; PINTSCHER, Lydia. A Glimpse into Babel: An Analysis of Multilinguality in Wikidata. *In*: OPENSYM'17: 13TH INTERNATIONAL SYMPOSIUM ON OPEN COLLABORATION, 2017, Galway, Irlanda. **Anais** [...]. Galway, Irlanda: Association for Computing Machinery. Disponível em: https://doi.org/10.1145/3125433.3125465. Acesso

em: 24 jan. 2020.

KIELING, Camila Garcia. Portal Memórias da Ditadura: Uma subversão cartográfica sobre a memória da ditadura militar no Brasil. *In*: XXXIX CONGRESSO BRASILEIRO DE CIÊNCIAS DA COMUNICAÇÃO, 2016, São Paulo. **Anais** [...]. São Paulo: Sociedade Brasileira de Estudos Interdisciplinares da Comunicação. Disponível em: http://portalintercom.org.br/anais/nacional2016/resumos/R11-0418-1.pdf. Acesso em: 23 jan. 2020.

KULKARNI, Akshay; SHIVANANDA, Adarsha. Extracting the Data. *In*: KULKARNI, A.; SHIVANANDA A. **Natural Language Processing Recipes: Unlocking Text Data with Machine Learning and Deep Learning using Python**. Berkeley, Califórnia, Estados Unidos da América: Apress, 2019, p. 1-35. Disponível em: https://doi.org/10.1007/978-1-4842-4267-4_1. Acesso em: 7 fev. 2020.

LAITANO, Bruno Grigoletti. Reflexões acerca da digitalização de arquivos da ditadura civil-militar brasileira. *In:* 30º SIMPÓSIO NACIONAL DE HISTÓRIA, 2019, Recife. **Anais** [...]. Recife: Associação Nacional de História. Disponível em: https://www.snh2019.anpuh.org/resources/anais/8/1554691778_ARQUIVO_BrunoGrigolettiLaitano.pdf. Acesso em: 23 jan. 2020.

LISTA DE MORTOS E DESAPARECIDOS POLÍTICOS NA DITADURA MILITAR BRASILEIRA. *In*: WIKIPÉDIA: A enciclopédia livre. [São Francisco, Califórnia, Estados Unidos da América: Fundação Wikimedia, 2014]. Disponível em: https://pt.wikipedia.org/w/index.php?title=Lista_de_mortos_e_desaparecidos_políticos_na_ditadura_militar_brasileira&oldid=38917886. Acesso em: 27 jan. 2020.

LISTA DE MORTOS E DESAPARECIDOS POLÍTICOS NA DITADURA MILITAR BRASILEIRA. *In:* WIKIPÉDIA: A enciclopédia livre. [São Francisco, Califórnia, Estados Unidos da América: Fundação Wikimedia, 2020]. Disponível em: https://pt.wikipedia.org/w/index.php?title=Lista_de_mortos_e_desaparecidos_políticos_na_ditadura_militar_brasileira&oldid=57195374. Acesso em: 17 jan. 2020.

LUZ, Larissa Pavarini da; CONEGLIAN, Caio Saraiva; SANTAREM SEGUNDO, José Eduardo. Tecnologias da Web Semântica para a recuperação da informação no Wikidata. **Revista Digital de Biblioteconomia e Ciência da Informação**, Campinas, v. 17, n. e019003, 2019. Disponível em: https://periodicos.sbu.unicamp.br/ojs/index.php/rdbci/article/view/8651791. Acesso em: 24 jan. 2020.

MARTINS, Allysson; MIGOWSKI, Ana. Cartografando a Ditadura Militar no Brasil: memórias coletivas e mapas digitais colaborativos. *In*: 24º ENCONTRO NACIONAL COMPÓS, 2015, Brasília. **Anais** [...]. Brasília. Disponível em: http://www.compos.org.br/biblioteca/compós2015-autores_2884.pdf. Acesso em: 23 jan. 2020.

MARTINS, Dalton Lopes; CARMO, Danielle do. Dinâmica de participação social na

Revista ACB: Biblioteconomia em Santa Catarina, Florianópolis (Brasil) - ISSN 1414-0594

construção coletiva de informação no campo museal: estudo de caso dos museus na Wikipédia no âmbito do Instituto Brasileiro de Museus. **Liinc em Revista**, Rio de Janeiro, v. 15, n. 1, 2019. Disponível em: http://revista.ibict.br/liinc/article/view/4607. Acesso em: 23 jan. 2020.

MITCHELL, Ryan. **Web Scraping with Python: Collecting More Data from the Modern Web**. 2. ed. [S.l.]: O'Reilly Media, 2018, 310 p.

MORAES, Renato; PESCHANSKI, João Alexandre; DIELLO, Mariana; CARRERA, Marília. A wiki-pedagogia no Jornalismo: o caso do Projeto Wikipédia da Faculdade Cásper Líbero. **Revista Brasileira de Ensino de Jornalismo**, Brasília, v. 6, n. 18, p. 75-100, 2016. Disponível em: http://rebej.abejor.org.br/ index.php/rebej/article/view/184/115. Acesso em: 23 jan. 2020.

NOIRET, Serge. História Pública Digital. **Liinc em Revista**, Rio de Janeiro, v. 11, n. 1, 2015. Disponível em: http://revista.ibict.br/liinc/article/view/3634. Acesso em: 23 jan. 2020.

PEDRETTI, Lucas. Silêncios que gritam: Apontamentos sobre os limites da Comissão Nacional da Verdade a partir do seu acervo. **Revista do Arquivo**, São Paulo, v. II, n. 5, p. 62-76, 2017. Disponível em: http://www.arquivoestado.sp.gov.br/revista_do_arquivo/05/artigo_04.php. Acesso em: 24 jan. 2020.

PELLISSIER TANON, Thomas; VRANDEČIĆ, Denny; SCHAFFERT, Sebastian; STEINER, Thomas; PINTSCHER, Lydia. From Freebase to Wikidata: The Great Migration. *In*: 25TH INTERNATIONAL CONFERENCE ON WORLD WIDE WEB, 2016, Montréal, Québec, Canadá. **Anais** [...]. Montréal, Québec, Canadá: Association for Computing Machinery. Disponível em: https://doi.org/10.1145/2872427.2874809. Acesso em: 6 fev. 2020.

PEREIRA, Mateus Henrique de Faria. Nova direita? Guerras de memória em tempos de Comissão da Verdade (2012-2014). **Varia Historia**, Belo Horizonte, v. 31, n. 57, p. 863-902, 2015. Disponível em: https://doi.org/10.1590/0104-87752015000300008. Acesso em: 23 jan. 2020.

PESCHANSKI, João Alexandre. Tarefa 2 do Projeto Wikipédia: melhorar itens de mortos e desaparecidos no Wikidata. In: PESCHANSKI, J. A. **Ciência política para jornalistas**. São Paulo, 18 out. 2019. Disponível em: http://cienciapoliticaparajornalistas.blogspot.com/2019/10/tarefa-2-do-projeto-wikipedia-melhorar.html. Acesso em: 4 fev. 2020.

PREDEFINIÇÃO:LISTA DO WIKIDATA. *In:* WIKIPÉDIA: A enciclopédia livre. [São Francisco, Califórnia, Estados Unidos da América: Fundação Wikimedia, 2018]. Disponível em: https://pt.wikipedia.org/w/index.php?title=Predefinição:Lista_do_Wikidata&oldid=53662896. Acesso em: 17 jan. 2020.

Revista ACB: Biblioteconomia em Santa Catarina, Florianópolis (Brasil) - ISSN 1414-0594

ROTTA, Vera. Comissão Especial de Mortos e Desaparecidos Políticos. **Acervo, Comissão Especial de Mortos e Desaparecidos Políticos**, Rio de Janeiro, v. 21, n. 2, p. 193-200, 2008. Disponível em: http://revista.arquivonacional.gov.br/index.php/revistaacervo/article/view/302. Acesso em: 24 jan. 2020.

SALGADO, Lívia de Barros. A Comissão Nacional da Verdade: espaço de política e poder. *In*: 29º SIMPÓSIO NACIONAL DE HISTÓRIA, 2017, Brasília. **Anais** [...]. Brasília: Associação Nacional de História. Disponível em: https://anpuh.org.br/uploads/anais-simposios/pdf/2019-01/1548953099_e8c972e36660ca4f6bdc2ed5d485803d.pdf. Acesso em: 24 jan. 2020.

SARTI, Cynthia. A construção de figuras da violência: a vítima, a testemunha. **Horizontes Antropológicos**, Porto Alegre, n. 42, p. 77-105, 2014. Disponível em: https://doi.org/10.1590/S0104-71832014000200004. Acesso em: 23 jan. 2020.

SCHINKE, Vanessa Dorneles; CASTRO, Ricardo Silveira. Relatório da Comissão Nacional da Verdade: O discurso sobre o judiciário. **Revista Direito e Práxis**, Rio de Janeiro, v. 7, n. 14, pp. 291-316, 2016. Disponível em: https://www.redalyc.org/pdf/3509/350945825011.pdf. Acesso em: 24 jan. 2020.

SIDI, Fatimah; PANAHY, Payam Hassany Shariat; AFFENDEY, Lilly Suriani; JABAR, Marzanah A.; IBRAHIM, Hamidah; MUSTAPHA, Aida. Data quality: A survey of data quality dimensions. *In*: 2012 INTERNATIONAL CONFERENCE ON INFORMATION RETRIEVAL KNOWLEDGE MANAGEMENT, 2012, Kuala Lumpur, Malásia. **Anais** [...]. Kuala Lumpur, Malásia: Universiti Putra Malaysia, 2012. Disponível em: https://ieeexplore.ieee.org/document/6204995. Acesso em: 24 jan. 2020.

VECCHIOLI, Virginia. Políticas de la Memoria y Formas de Clasificación Social. ¿Quiénes son las "Víctimas del Terrorismo de Estado" en la Argentina?. *In*: GROPPO, B.; FLIER, P. (eds.). **La imposibilidad del olvido: Recorridos de la memoria en Argentina, Chile y Uruguay**. La Plata, Argentina: Al Margen, 2001, p. 83-102. (Diagoníos)

VISUALIZAÇÕES DA PÁGINA. *In*: PAGEVIEWS ANALYSIS. [São Francisco, Califórnia, Estados Unidos da América: Fundação Wikimedia, 2020]. Disponível em: https://pageviews.toolforge.org/?project=pt.wikipedia.org&platform=all-access&agent=user&redirects=0&start=2015-07-01&end=2020-01-26&pages=Lista_de_mortos_e_desaparecidos_políticos_na_ditadura_militar_brasileira. Acesso em: 27 jan. 2020.

WIKIPÉDIA:GLAM/ARQUIVO NACIONAL. In: WIKIPÉDIA: A enciclopédia livre. [São Francisco, Califórnia, Estados Unidos da América: Fundação Wikimedia, 2019]. Disponível em: https://pt.wikipedia.org/w/index.php?title=Wikipédia:GLAM/Arquivo_Nacional&oldid=54920166. Acesso em: 27 jan. 2020.

Revista ACB: Biblioteconomia em Santa Catarina,
Florianópolis, v. 26, n. 4, p. 1-30, set./dez., 2021.

Completeness, consistency and correctness in digital databases on the politically killed and disappeared during the Brazilian civil-military dictatorship

Abstract: The article compares four databases on the politically killed and disappeared during the Brazilian civil-military dictatorship: the Special Commission on Political Killed and Disappeared; the Political Disappeared website; the Memories of the Dictatorship portal; and the report of the National Truth Commission. These DBs are evaluated according to the completeness, consistency and correctness of information, using methods to investigate heterogeneous databases. The results of the comparison indicate variations in the quality of the analyzed bases, when viewed in relation to each other and when evaluating the internal properties of each base. The comparison involved transferring the data on these bases to Wikidata; this transfer led to transforming the heterogeneous information into a more complete, consistent and accurate semantic database, after computational curation.

Keywords: Brazilian dictatorship. Politically killed and disappeared. Digital databases. Wikidata. Digital humanities.

Revista ACB: Biblioteconomia em Santa Catarina, Florianópolis (Brasil) - ISSN 1414-0594