Avaliando a Integridade de um Repositório

David Marques[2,5], Emely Rossim[4], Jean-Rémi Bourguet[6,7], and Elias de Oliveira[1,4,5]

[1]Departamento de Arquivologia
[2]Departamento de Informática
[4]Programa de Pós-Graduação em Informática
[5]Universidade Federal do Espírito Santo
[6]Departamento de Ciência da Computação
[7]Universidade Vila Velha, Vila Velha, Brasil

Resumo

A identificação de objetos espúrios em um repositório é uma tarefa árdua. Dependendo do volume de itens no acervo da instituição, essa atividade pode demandar muito tempo dos especialistas na execução de um *procedimento do tipo censo*. Diante dessa dificuldade, pode-se adotar uma *metodologia de amostragem*, desde que o erro associado esteja dentro de um intervalo aceitável para os padrões de qualidade almejados pela instituição ou exigidos por normas específicas. Neste trabalho, propomos uma estratégia baseada em Estatística Bayesiana para reduzir o esforço do especialista no treinamento do algoritmo de inteligência artificial na identificação de objetos espúrios. Nossos experimentos mostram que é possível reduzir para cerca de 3.96% do tamanho do acervo, o esforço humano necessário para fornecer exemplos ao treinamento do algoritmo na identificação do restante do acervo.

1 Introdução

Com o propósito de ampliar a disseminação do conhecimento gerado nas instituições acadêmicas para o público em geral, um sistema informático anteriormente conhecido como *Biblioteca Digital Brasileira de Teses e Dissertações* (BDTD)

(KURAMOTO, 2006), atualmente denominado repositório institucional, tornou-se uma exigência nacional. Assim, o número de sistemas desse tipo no Brasil ultrapassa 63 unidades, equivalente ao número de universidades federais no país.

Um repositório permite o acesso às dissertações, teses, artigos e outros documentos produzidos pela instituição responsável. Carvalho, Silva e aes (2012) descreve a experiência do repositório ARCA, um grande acervo digital na área de saúde mantido pela Fiocruz[1]. Os autores dizem que

> *Recuperar, organizar e preservar a produção técnico-científi ca do Icict foi o mote do projeto para prover o acesso e disseminar o conhecimento acumulado. Assim, o Icict está contribuindo para a produção de novo conhecimento, aumentando a visibilidade da sua produção intelectual e fortalecendo o processo de comunicação científica.*

O Icict – *Instituto de Comunicação e Informação Científica e Tecnológica em Saúde* tem como missão participar da formulação, implementação e avaliação de políticas públicas, desenvolver estratégias e executar ações de informação e comunicação no campo da ciência, tecnologia e inovação em saúde, objetivando atender às demandas sociais do Sistema Único de Saúde - SUS e de outros órgãos governamentais Carvalho, Silva e aes (2012).

Em cada unidade universitária, os objetivos não são diferentes. O acesso a informação produzida pela comunidade acadêmica e, *a posteriori*, disseminada para a sociedade em geral é o mote da existência de quaisquer outros repositórios. Portanto, o repositório institucional da Universidade Federal do Espírito Santo segue o mesmo propósito. O RiUfes[2] implantado desde outubro de 2010 vem ampliando seu acervo desde então, como podemos observar na Figura 1.

É importante contextualizar a queda no número de documentos acadêmicos submetidos ao RiUfes nos últimos anos, evidenciada na 1, dentro de um contexto mais abrangente. De acordo com uma pesquisa realizada pela Elsevier[3] e Agência Bori[4] (ELSEVIER-BORI, 2023), que analisou a produção científica de 51 países entre 1996 e 2022, observou-se que a partir de 2020, a produção de artigos com autores Brasileiros diminuiu seu ritmo de crescimento e enfrentou, em 2022, sua primeira queda. Esse declínio é ainda mais evidente quando comparado ao crescimento observado em décadas anteriores. Os autores da pesquisa destacam que é muito provável que o decréscimo no ritmo de crescimento da produção científica se deva, em boa parte, aos efeitos da pandemia.

[1] http://portal.fiocruz.br/

[2] http://repositorio.ufes.br/

[3] http://www.elsevier.com/

[4] https://abori.com.br/

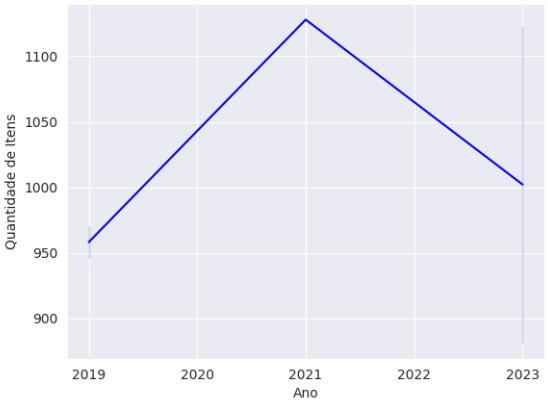

Figura 1: Crescimento do acervo do RiUfes

Ao longo do tempo, os repositórios têm modificado suas práticas de arquivamento. Atualmente, adota-se o auto arquivamento, onde cada autor de uma produção científica é responsável pela inserção dos elementos descritivos e pelo arquivamento do documento (ASSIS, 2013). Quando o arquivamento era realizado pelos profissionais mantenedores do repositório, isso frequentemente se tornava um um gargalo ao tempo de liberação dos objetos submetidos para disponibilização Por outro lado, o autoarquivamento introduziu desafios adicionais relacionados à qualidade dos elementos descritivos e conteúdo.

Para garantir o acesso confiável aos documentos, cada objeto descrito no repositório deve conter efetivamente o conteúdo correspondente. Qualquer inconsistência pode indicar que o objeto é espúrio. Experiências passadas incluem casos como o de uma Ata de defesa de dissertação presente no lugar da dissertação propriamente dita.

Uma abordagem para resolver esse problema seria a investigação item a item de todos os elementos garantidores da qualidade de um acervo, ou repositório. Esse tipo de verificação, similar a um censo, é altamente custoso e intensivo em mão de obra especializada. Para contextualizar, uma tentativa desse tipo em nosso repositório envolveu 26 profissionais e levou cerca de 10 meses para verificar aproximadamente 5.049 itens à época. Arbitrariamente estipulamos que esse processo consumiu cerca de 2 minutos por item.

Uma alternativa ao censo seria o uso de amostragens. Por ser um processo aproximativo, é crucial especificar um intervalo de erro aceitável para as amostras. Um erro estimado pequeno, como ± 10 itens espúrios, poderia ser considerado aceitável, considerando a quantidade de itens no repositório naquela época, e

ainda mais atualmente, com o aumento do volume. No entanto, um erro de ± 100 itens poderia ser inaceitável.

Dessa forma, a proposta discutida neste trabalho baseia-se na utilização de amostragens sucessivas, por meio de um ambiente eletrônico que servirá de interface para o processo. O foco principal será a integridade dos itens, ou seja, se um item estiver descrito como dissertação em seus metadados, é necessário que haja uma dissertação como objeto armazenado no repositório, e não outro tipo de documento. Para estimar o grau de integridade, serão realizadas amostragens sucessivas e aleatórias. Em cada etapa, itens serão selecionados para que especialistas avaliem a integridade do objeto, por meio do envio de seu endereço digital, a URL (*Uniform Resource Locator*). Por exemplo a URL `http://repositorio.ufes.br/handle/10/12183` levaria o avaliador a uma dissertação de mestrado do Programa de Programa de Pós-Graduação em Ciência da Informação com o título *A biblioteca escolar no processo de ensino-aprendizagem: estudo de caso da rede de ensino do município de Vila Velha, Espírito Santo*.

Os experimentos amostrais sugerem que podemos diminuir grandemente o custo da avaliação da qualidade do repositório em mais de 96.03% do tamanho do acervo, o que representa uma grande economia de tempo e trabalho para o setor. Além disso, o processo estatístico aplicado nos dá garantias de que, caso ainda haja algum objeto espúrio, a probabilidade desse evento ocorrer seria igual ou inferior a 3%, por exemplo. Para um acervo com a quantidade de itens existente hoje, é uma margem aceitável.

Esse artigo está estruturado da seguinte forma. Na Seção 2, discutimos alguns dos trabalhos recentes que servirão de base de comparação com a nossa proposta. Na Seção 3 descrevemos nossa proposta à luz da economia de recursos e sucessos em outras áreas. Na Seção 4 discutimos os experimentos e os resultados. Por fim, na Seção 5, nossas conclusões e algumas sugestões para trabalhos futuros.

2 Revisão de Literatura

Kuramoto (2006) discute a importância da criação de um sistema capaz de difundir a produção científica brasileira tendo em vista os *altos custos custos na manutenção das assinaturas das revistas científicas*. É patente que o acesso à *informação científica é o insumo básico para o desenvolvimento científico e tecnológico de um país*. É nesse contexto que surgem as BDTDs. Estímulos foram dados a que as instituições de ensino superior mantenedoras de programas de pós-graduação fossem pioneiras nesse processo em nosso país. Os dados seriam armazenados em suas respectivas instituições. Contudo, em decorrência da arquitetura de interoperabilidade com os demais entes da rede e o sistema do centralizador no Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT), todos

poderiam ter acesso a quaisquer dos dados coletados pelas partes. Essa interoperabilidade é possível pelo fato de que a BDTD utiliza-se do *Padrão Brasileiro de Metadados de Teses e Dissertações* (MTD-BR). À época, o depósito dos objetos na BDTD, em nossa instituição, era feito por um profissional do setor, o que se verificou ser um gargalo. O auto-arquivamento, processo em que o autor envia o texto para publicação, sem a intervenção de outras pessoas (BOMFÁ et al., 2008), foi posteriormente adotado.

No estudo de Bourguet, Silva e Oliveira (2021), é desenvolvida uma estratégia para identificar e corrigir objetos espúrios em repositórios acadêmicos, focando principalmente em dissertações de mestrado e teses de doutorado. Utilizando metodologias como Estimação de Máxima Verossimilhança, modelagem de distribuição gama, aplicação do Teorema de Bayes (BUSSAB; MORETTIN, 2013, Cap. 11) e a estratégia *Bag of Words*, os autores construíram um modelo probabilístico capaz de categorizar documentos de forma eficaz.

Por outro lado, Assis (2013) explora um panorama mais amplo das políticas de acesso aberto e às suas implicações para a comunidade acadêmica, principalmente em relação às práticas de auto arquivo dentro dos repositórios institucionais. O autor tem interesse em compreender a interação entre mandatos políticos e práticas acadêmicas, mais especificamente em como essas políticas afetam o comportamento dos pesquisadores quando se trata de colocar suas próprias pesquisas *online*. Atualmente, o autoarquivamento é uma prática comum. No entanto, esta pode trazer problemas, como inconsistências nos elementos descritivos e no conteúdo dos documentos, o que compromete a integridade dos repositórios.

Apesar das abordagens distintas, os estudos mencionados acima compartilham o objetivo comum de promover a integridade de repositórios e facilitar o acesso e a compreensão da pesquisa acadêmica. O trabalho atual também está alinhado a este objetivo. Para alcançá-lo, propomos uma metodologia baseada em Estatística Bayesiana para a classificação técnica de documentos. Além disso, abordamos questões mais abrangentes relacionadas à gestão de repositórios institucionais, enfatizando a importância desses repositórios, os desafios enfrentados e a garantia da integridade dos documentos.

Na próxima seção, discutimos nossa proposta para a aquisição de conhecimento, via documentos, que utilizaremos para a construção de conhecimento para o domínio da instituição de arquivo, nossa parceira nesse projeto.

3 Uma Estratégia Incremental de Amostragem

Esse projeto propõe uma estratégia incremental para medir-se, de forma quantitativa e precisa, um repositório para que o mesmo possa estar de acordo com o esperado para melhor servir seus usuários. Para tanto, discutimos aqui apenas um

dos muitos outros aspectos relacionados a um contexto maior de certificação desse tipo de sistema computacional. O auto-arquivamento trouxe vantagens em termos de agilidade na disponibilização do objeto no repositório (ASSIS, 2013), outros problemas surgiram e precisam ser tratados.

Nossa proposta consiste na utilização de modelos estatísticos para medir o parâmetro de integridade entre o que é descrito no metadado do objeto e o real conteúdo do objeto armazenado no repositório. A Estatística nos oferece forma de estimação de valores e parâmetros. Contudo, por ser uma estimação, teremos sempre um quantitativo de erro associado ao processo. O que buscamos é uma estimativa de mínimo erro estimado e uma máxima redução do esforço humano para realizar o trabalho. Um compromisso comum a ser ponderado em qualquer tipo de empreendimento.

Nosso procedimento inicia-se com a mineração dos itens existentes no repositório alvo. Para tanto, precisamos colher todos os *hiperlinks* que identifique cada um dos itens do repositório a ser avaliado. Utilizamos o *crawler-RiUfes*, um programa escrito em *Shell Script* do Linux. O Linux é um sistema operacional de código aberto, livre e gratuito. A linguagem *Shell Script* está disponível em todos os sistemas operacionais Linux, portanto, acessível e de fácil uso para quem quiser replicar nossos experimentos.

Tendo todos os *hiperlinks* coletados, ou hiperreferências, o gestor pode especificar a quantidade desse *links* que serão enviados para cada especilista avaliar. A escolha do conjunto dos *links* a serem enviados para um dos especialistas participantes do processo de avaliação se dá de forma aleatória. O objetivo é avaliarmos itens de forma bem abrangente no acervo e não apenas alguns com alguma característica específica, o que incutiria um viés no procedimento. A escolha aleatória dos *links*, que levam aos itens, garante que as propriedades estatísticas sejam preservadas e, portanto, teremos como inferir um resultado para predizer parâmetros estatísticos para a população, representada por essa amostra avaliada. Ou seja, avaliaremos apenas uma porção de todo o acervo e, pela estatística, termos garantias de que todo o acervo esteja tão íntegro como avaliado pela amostra. É claro, considerando as margens de erro.

Para a execução da amostragem, a proposta inclui a utilização de um ambiente de eletrônico como interface para lidar com os susuários especialistas que apoiarão o processo de amostragem. Adotamos o sistema do Moodle[5], também *software* aberto e livre e de grande utilização nas instituições de ensino no Brasil. Junto a esse *software*, desenvolvemos um conjunto de outros utilitários para comunicação com o Moodle. Esses utilitários se encarregarão do envio e coleta das informações que os especialistas nos fornecerão. Por intermédio desses utilitários, que são também *software*, um conjunto dessas hiperrerências são enviadas

[5]http://www.moodle.org

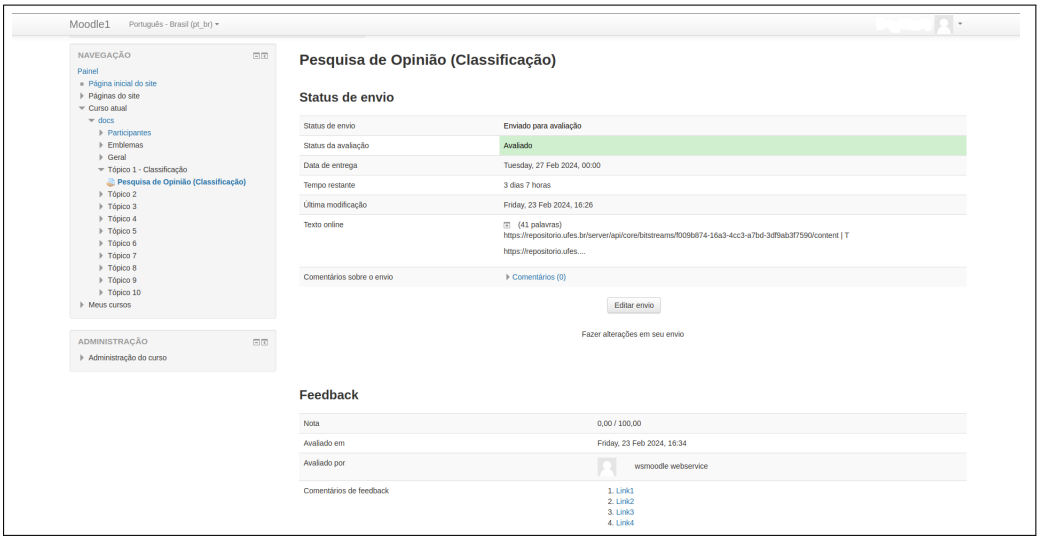

Figura 2: Recebendo hiperreferências do repositório em um Moodle

para cada especialista que aceite participar do processo de auditoria do repositório, via o uso do Moodle, como mostra a Figura 2. No final da figura, o usuário recebeu 4 *links* para serem avaliados. Ao final esse especialista responderá se o que pode avaliar na ponta é de fato o que os metadados diz ser, ou não. Essas respostas são agregadas em um servidor para posterior processamento estatístico dos resultados.

4 Experimentos e Resultados

Processos de amostragem consistem em examinarmos uma *porção de objetos – (amostra)*, escolhidos de forma aleatória, dentre o universo total de itens no conjunto, e medirmos as respectivas características de interesse. No nosso caso, estamos interessados em verificar se o metadado que designa se um objeto seja Dissertação de mestrado, ou Tese de doutorado, de fato sejam, respectivamente, o que se registrou no metadado. Qualquer contradição consideraremos tais itens como objetos espúrios, corroendo portanto, a qualidade da integridade do repositório como um todo.

Em nossa metodologia, um especialista receberá uma hiperreferência de um conjunto de objetos do acervo de itens de um repositório. Essas hiperreferências, ou *hiperlinks*, levarão esse especialista diretamente ao item no ciberespaço. O exemplo, que já apresentamos anteriormente, `http://repositorio.ufes.br/handle/10/12183`, nos leva diretamente a uma dissertação defendida no programa de Pós-graduação

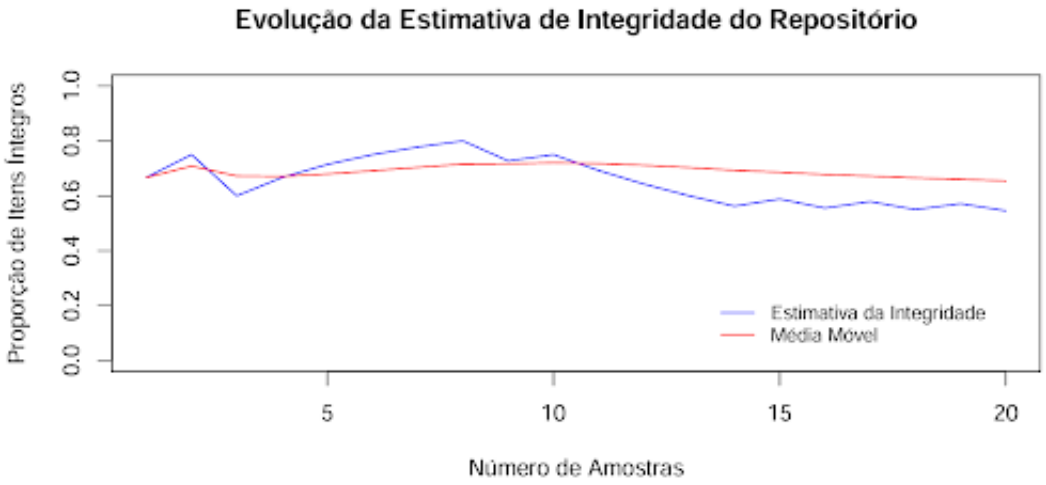

Figura 3: Evolução da quatidade de espúrios por amostragem

em Ciência da Informação da Universidade Federal do Espírito Santo. O metadado, portanto, confere com o conteúdo apresentado, portanto, não se trata de um objeto espúrio, pelo contrário, trata-se de um objeto íntegro no repositório.

Nossa proposta realiza-se de forma incremental, ou seja, algumas amostragens são feitas e avaliadas. Caso algum objeto espúrio seja identificado, esse será enviado para os funcionários do setor para recuperação do objeto espúrio. Uma nova amostragem é feita e, novamente, uma nova avaliação é realizada. Esse processo é repetido até que o valor de erro figure dentro dos padrões esperados.

Essa estratégia incremental é baseada na Teoria de Bayes (BESSIERE et al., 2014). A cada amostragem os valores estatísticos esperados são reavaliados e atualizados até que esses valores fiquem estabilizados.

Para iniciar nossos experimentos, adotaremos, como hipótese *a priori*, de que nosso repositório tenha uma proporção de 50% de integridade. Veja que partimos dessa hipótese, o que seria uma situação idealmente inaceitável para o caso concreto.

Nesse sentido, convidamos um conjunto de 27 avaliadores. Nesses experimentos utilizamos um grupo de alunos de graduação para validarmos nossa metodologia. Nessa primeira etapa, esses alunos receberam, cada um, 3 hiperreferências. Como resultados dessa primeira etapa foram avaliados um total de 81 itens do repositório. Encontrou-se uma quantidade de sete itens espúrios nessa primeira investida. Em uma próxima amostragem, encontrou-se dois itens espúrios, ou seja, uma variação de cinco itens.

A hipótese inicial de que o repostiório teria 50% de integridade já foi atualizada para 55% e o erro estimado de ± 21.8.

A redução da expectativa sobre a existência de objetos espúrios é considerável. Entretanto, a quantidade de amostras, já avaliadas até esse ponto, é também pequena em relação ao total de itens no repositório para, com uma certa segurança, extrapolarmos esses valores estatísticos para o acervo do repositório como um todo.

Assim, em nossa metodologia, continuamos fazendo amostragens e observando a quantidade de itens espúrios encontrados. Essa evolução está descrita na Figura 3.

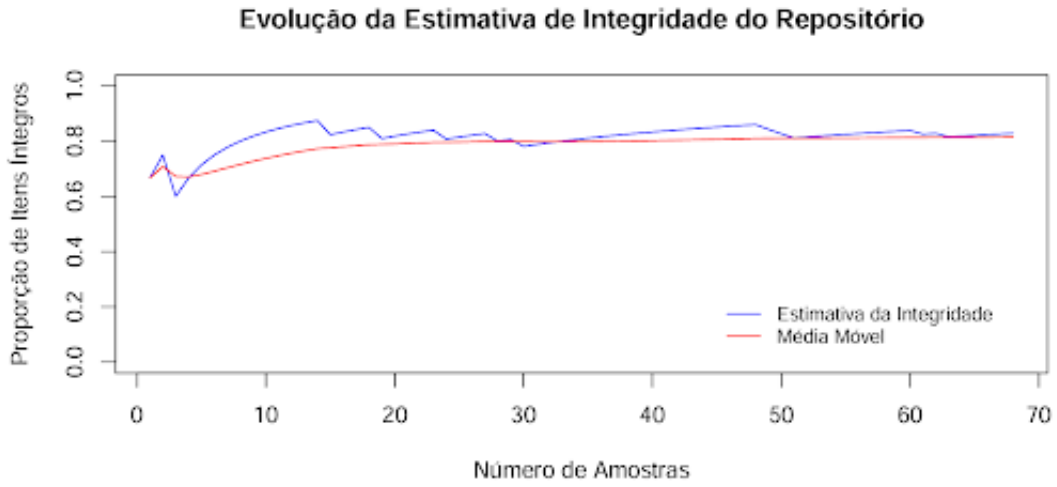

Figura 4: Evolução do processo de amostragem

A Figura 3 apresenta um resultado de estimativa de integridade de 94.74% do acervo com intervalo de erro entre ± 7.1. Até esse ponto, foram feitas 58 avaliações de itens, com a identificação de 9 itens espúrios.

Após termos feitos uma quantidade suficiente de amostragens para termos as garantias estatísticas indicadas pela literatura, em geral uma sequência de mais de 200 eventos de amostragens. Por outro lado, utilizando a abordagem frequentista, a clássica da Estatística (PEK; ZANDT, 2020) (BUSSAB; MORETTIN, 2013, Cap. 10), o tamanho da amostra necessária seria de aproximadamente 381 itens, com margem de erro de $\pm 5\%$. Contudo, ao final, fizemos apenas 200 avaliações de itens, sendo que, ao final do processo, identificamos um total de 10 espúrios itens no acervo. Uma proporção de 0.00198 itens espúrios no acervo, como indicado

na Figura 4.

O procedimento proposto, e executado experimentalmente nesse trabalho, mostra que o repositório analisado tem integridade de 95% em seu acervo com uma garantia estimada de erro de ± 3.02 itens. Outro resultado importante é a grande redução do esforço necessário para a realização dessa avaliação, em comparação ao procedimento censitário empreendido anteriormente.

Com a proposta atual, estima-se que serão necessários 400 minutos de trabalho, o que corresponde a menos de um dia de trabalho de um especialista. Isso representa uma redução de 96.03% em relação à força de trabalho necessária anteriormente, quando foram avaliados 5.049 itens, consumindo cerca de 2 minutos por item.

5 Conclusões

Esse trabalho procurou mostrar algumas vantagens do método amostral incremental sobre o censo para medir-se a qualidade de integridade dos itens em um repositório. A prosta dessa estratégia nos permitiu reduzir a força de trabalho dos especialistas para 96.03%, contra ter que avaliar todos os itens do acervo. Dizendo de outra forma, precisamos avaliar apenas 200 itens para chegarmos a garantir que o repositório seja, pelo menos, 95% íntegro com erro de ± 3. As primeiras estimativas de acordo com o método clássico seria de 381 itens. Note que, embora o método amostral incremental traga a desvantagem de também ser um processo aproximativo do real valor medido, contra o método censitário, pode-se ainda assim calcular o nível de (*in-*)certeza da aproximação obtida tal como o método amostral clássico.

Com o apoio de procedimentos de Inteligência Artificial ao longo do processo consegui-se uma economia do uso do tempo de profissionais especializados na execusão dessa tarefa. Isso abre espaço para uma ação de auditoria do processo com mais frequência do queer a feito anteriormente. Além disso, diante da metodologia e tecnologia proposta, outros profissionais e até mesmo alunos podem participar do processo, o que reduziria ainda mais os custo de mantermos um repositório em sua maior integridade.

Como futuro trabalhos, planeja-se a extensão dessa mesma metodologia para a avaliação de outros itens relacionadas a certificação de um repositório. Com o apoio da metodologia e tecnologia desenvolvida nesse trabalho, conta-se que a diminuição do esforço será também proporcional.

Referências

ASSIS, T. B. Análise das Políticas de Autoarquivamento nos Repositórios Institucionais Brasileiros e Portugueses. *InCID: Revista de Ciência da Informação e Documentação*, v. 4, n. 2, p. 212–227, 2013.

BESSIERE, P. et al. *Bayesian Programming*. 1st. ed. Boca Baton, FL: Chapman & Hall/CRC, 2014.

BOMFÁ, C. R. Z. et al. Acesso Livre à Informação Científica Digital: Dificuldades e Tendências. *TransInformação*, SciELO Brasil, v. 20, p. 309–318, 2008.

BOURGUET, J.-R.; SILVA, W.; OLIVEIRA, E. Minimalist Fitted Bayesian Classifier-Based on Likelihood Estimations and Bag-Of-Words. In: . Online: Springer, 2021.

BUSSAB, W. O.; MORETTIN, P. A. *Estatística Básica*. 8. ed. São Paulo: Saraiva, 2013.

CARVALHO, M. C. R.; SILVA, C. H.; AES, M. C. S. G. *Repositório Institucional da Saúde: a Experiência da Fundação Oswaldo Cruz*. [S.l.], 2012.

ELSEVIER-BORI. *Análise da Produção Científica de 1996-2022: Queda Inédita no Número de Artigos Científicos do Brasil*. São Paulo, jun. 2023. Disponível em: https://abori.com.br/publicacoes/.

KURAMOTO, H. Informação Científica: Proposta de um Novo Modelo para o Brasil. *Ciência da Informação*, SciELO Brasil, v. 35, p. 91–102, 2006. Disponível em: <https://doi.org/10.1590/S0100-19652006000200010>.

PEK, J.; ZANDT, T. V. Frequentist and Bayesian Approaches to Data Analysis: Evaluation and Estimation. *Psychology Learning & Teaching*, v. 19, n. 1, p. 21–35, 2020. Disponível em: <https://doi.org/10.1177/1475725719874542>.