XXII Encontro Nacional de Pesquisa em Ciência da Informação – XXII ENANCIB

ISSN 2177-3688

GT8 – Informação e Tecnologia

QUALIDADE DE DADOS EM ACERVOS MUSEAIS: UMA AVALIAÇÃO SEMIAUTOMÁTICA PARA OS ACERVOS SOB GESTÃO DO IBRAM

DATA QUALITY IN MUSEUM COLLECTION: A SEMI-AUTOMATIC EVALUATION FOR THE COLLECTIONS UNDER IBRAM'S MANAGEMENT

Abeil Coelho Júnior. UFES.

Daniela Lucas da Silva Lemos. UFES.

Modalidade: Resumo Expandido

Resumo: A demanda por dados com qualidade a partir de práticas maduras de catalogação tem sido uma realidade em instituições, como é o caso do Instituto Brasileiro de Museus (Ibram). O objetivo do artigo é apresentar uma avaliação diagnóstica de qualidade de dados realizada nas bases dos museus vinculados ao Ibram frente às boas práticas de catalogação indicadas no guia *Cataloging Cultural Objects* (CCO). Metodologicamente, a exploração dos dados foi realizada por solução semiautomática em 3 coleções de caráter museológico do Ibram. O diagnóstico permite à Instituição fazer recomendações visando qualificar seus atuais padrões de documentação pensando no usuário final.

Palavras-Chave: Organização da Informação. Qualidade de Dados. Padrões de Documentação.

Abstract: The demand for quality data from mature cataloging practices has been a reality in institutions, such as the Brazilian Institute of Museums (Ibram). The objective of the article is to present a diagnostic evaluation of data quality carried out in the databases of museums linked to Ibram given the good cataloging practices indicated in the Cataloging Cultural Objects (CCO) guide. Methodologically, data exploration was carried out using a semi-automatic solution in 3 Ibram museum collections. The diagnosis allows the Institution to make recommendations aimed at qualifying its current documentation standards with the end-user in mind.

Keywords: Information Organization. Data Quality. Documentation Standards.

1 INTRODUÇÃO

Durante as últimas décadas, constatou-se considerável adesão das instituições de patrimônio cultural a processos de digitalização e disponibilização de seus dados de acervos na internet, projetando maior acessibilidade e democratização de conhecimento científico e cultural à sociedade (MARTINS *et al.*, 2022).

Nesse contexto, instituições vêm investindo na adoção de padrões e práticas de documentação para a produção de metadados visando o alcance de interoperabilidade a nível sintático e semântico entre esquemas de metadados e suas aplicações (ZENG, 2019). Tal

processo geralmente é composto por orientações acerca do uso de padrões para tratamento nos dados (GILLILAND, 2016) focando normalização, qualidade e intercâmbio de descrições em ambiente digital, incluindo os seguintes elementos preponderantes: i) estrutura de dados: conjunto de elementos de metadados que formam um registro de informação; ii) valores dos dados: linguagens documentárias, vocabulários controlados, arquivos de autoridade e ontologias de domínio usados para preencher os dados nos elementos de metadados; iii) conteúdo dos dados: regras e códigos de catalogação que orientam em formatações, sintaxes e relacionamentos para os valores de dados usados para preencher os elementos de metadados; e iv) comunicação de dados: padrões de metadados expressados em uma linguagem legível para a máquina.

No setor cultural, foco desta pesquisa, Martins *et al.* (2022) constataram que a publicação de dados com qualidade na internet por instituições de memória representativas no Brasil, incluindo o próprio Instituto Brasileiro de Museus (Ibram), ainda é muito precária em termos de organização e representação de informações, dificultando o provimento de possíveis meios de agregar seus dados de acervos na web.

Na busca de se inserir nesse contexto de agregação de dados em rede, e vislumbrando benefícios com sofisticados mecanismos de busca, recuperação e gestão de acervos em interface única de suas bases de dados integradas, o Ibram vem adotando atualmente estratégias para o desenvolvimento de uma rede interoperável de museus digitais na web para o cenário brasileiro. Assim, desde o ano 2016, o Ibram vem implantando em seus 30 museus federais a plataforma digital Tainacan, objetivando preservação, difusão e integração dos acervos de suas instituições (SIQUEIRA; MARTINS, 2021).

No caso do procedimento de agregação dos museus, tornou-se necessário um alinhamento dos metadados que descrevem as bases de dados das coleções dos museus envolvidos na integração, de modo que os dados correspondentes pudessem fazer referência a um modelo comum (SIQUEIRA; MARTINS, 2021). Neste caso, foi adotado o padrão de dados orientado internamente pela Instituição, a saber: o modelo do Inventário Nacional de Bens Culturais Musealizados – INBCM (BRASIL, 2021).

A situação problemática que se configura a partir do uso da normativa INBCM pelos museus vinculados ao Ibram é que, a princípio, o INBCM surgiu para servir de instrumento de

inventário para gestão interna de acervos, não sendo, a priori, um modelo de catalogação que almeja requisitos descritivos únicos e singulares, vocabulários padronizados, indexação, localização, acesso e navegação em sistemas de recuperação da informação (SRIs) contemporâneos (LANCASTER, 2004; INTERNATIONAL FEDERATION OF LIBRARY ASSOCIATIONS AND INSTITUTIONS, 2016; MARTINS *et al.*, 2022).

Logo, a ausência de um modelo de catalogação apropriado pode comprometer a identificação de informações cruciais e necessárias para descrever um item museal, de modo a localizá-lo no acervo para fins de busca e recuperação (INTERNATIONAL FEDERATION OF LIBRARY ASSOCIATIONS AND INSTITUTIONS, 2016). A utilização de regras de catalogação em instituições como museus é essencial, uma vez que orientam os formatos e os valores adequados de preenchimento (GILLILAND, 2016) acerca dos elementos de metadados constitutivos de suas bases de dados, que podem, inclusive, serem utilizados como possíveis índices numa interface de busca e navegação.

Dentre os padrões de documentação recomendados no campo do patrimônio cultural (SILVA; LARA, 2021), incluindo modelos de dados, modelos ontológicos, padrões de metadados, guias de gestão de acervos e guias de catalogação, a presente pesquisa destaca o guia *Cataloging Cultural Objects* (CCO) pelos argumentos que se seguem.

O CCO, publicado pela *American Library Association* (ALA) em 2006, é um padrão de criação de conteúdo descritivo para recursos culturais, derivado do padrão semântico *Categories for the Description of Works of Art* (CDWA), que traz recomendações e regras de catalogação descritas com clareza e bem organizadas em grupos de informação sintetizados em 9 capítulos. Um dos destaques do CCO é que o padrão apresenta conceitos genéricos que podem ser utilizados com outros conjuntos de metadados, como, por exemplo, o MARC, o Dublin Core, e, inclusive, com os elementos descritivos do INBCM.

Assim sendo, o objetivo do presente artigo é apresentar o resultado de uma avaliação semiautomática realizada nas bases de dados dos museus vinculados ao Ibram, visando um diagnóstico da qualidade de dados desses museus frente às boas práticas de catalogação indicadas no guia de referência no campo da cultura digital, o CCO.

2 METODOLOGIA DE PESQUISA

A presente pesquisa foi classificada como sendo de natureza teórica e aplicada, qualiquantitativa, e de cunho bibliográfico, exploratório e descritivo, envolvendo a avaliação da conformidade dos elementos de descrição do INBCM que descrevem os acervos dos museus sob gestão do Ibram em consonância com os elementos recomendados pelo guia CCO. Trata-se, portanto, de um estudo de caso que visa avaliar a qualidade de dados de acervos digitais específicos de museus que usam a tecnologia de repositório digital Tainacan.

Algumas decisões metodológicas são importantes de serem esclarecidas inicialmente para fins de entendimento dos dados trabalhados na pesquisa. No que diz respeito ao INBCM, foram considerados apenas os 15 elementos de descrição para identificação do bem cultural de caráter museológico. Tal decisão foi feita após análise prévia dos dados captados dos acervos à luz das orientações do INBCM. Apenas o Museu Solar Monjardim apresentava acervo do tipo Arquivístico. Todos os outros acervos (22 coleções no total) dos 20 museus utilizavam metadados especificados pelo INBCM com carácter museológico.

Outro detalhe importante a ser destacado é que foram considerados 8 dos 9 capítulos do CCO. O Capítulo 9, denominado *View Information*, é endereçado à catalogação do substituto digital de uma obra, a exemplo de uma imagem. Os dados de catalogação dos acervos museais vinculados ao IBRAM são referentes às obras presentes nos museus, logo, não descreve as imagens representativas dessas obras.

Como este trabalho traz resultados parciais de uma pesquisa em andamento, optou-se por uma exploração e análise mais apurada das bases de dados de 3 coleções de caráter museológico de 3 instituições, a saber: Museu Casa da Hera, coleção de Indumentárias; Museu Solar Monjardim, coleção museológica; e Museu das Missões, coleção de Arte Sacra. O critério usado para a seleção desses museus se deu pela quantidade de objetos nas coleções, visto que seria possível validar com mais acurácia o algoritmo em si e obter resultados satisfatórios para o progresso do experimento. As etapas do método de avaliação dos dados dos acervos elencados são descritas nas subseções a seguir.

2.1 Alinhamento entre elementos de descrição: INBCM e CCO

O primeiro passo foi realizar o alinhamento (mapeamento) entre os elementos descritivos da normativa do INBCM e do guia de catalogação CCO, ambos previamente

estudados e analisados em suas respectivas fontes. Essa fase tornou-se o ponto de partida para que as regras de catalogação do CCO pudessem ser sugeridas de aplicação aos elementos descritivos do INBCM e, por conseguinte, aos metadados constitutivos das bases de dados dos museus explorados na presente pesquisa.

O alinhamento (Quadro 1) se deu a partir de um procedimento manual e intelectual baseado na aquisição de conhecimento sobre os dois instrumentos de pesquisa, com destaque para o aspecto de ordem semântica (papel das entidades de informação) nas decisões de cotejamento dos elementos descritivos destinados a um recurso de informação.

Quadro 1 – Alinhamento entre elementos descritivos – INBCM e CCO.

Capítulo CCO	Elemento CCO	Obrigatório CCO	Vocabulário Controlado CCO	Elemento INBCM	Obrigatório INBCM
I-Part 2	Work Type	Sim	Sim	Denominação	Sim
I-Part 2	Title	Sim	Não	Título	Não
II-Part 2	Creator	Sim	Sim	Autor	Sim
III-Part 2	Measurements	Sim	Sim	Dimensões	Sim
III-Part 2	Materials and Techniques	Sim	Sim	Material/Técnica	Sim
III-Part 2	Physical Description	Não	Sim	Estado de Conservação	Sim
III-Part 2	Inscription	Sim	Sim	Número de Registro	Sim
IV-Part 2	Date	Sim	Não	Data de Produção	Não
V-Part 2	Creation Location	Não	Sim	Local de Produção	Não
V-Part 2	Location	Sim	Sim	Situação	Sim
VII-Part 2	Class	Sim	Sim	Classificação	Não
VIII-Part 2	Description	Não	Não	Resumo Descritivo	Sim
VIII-Part 2	Other Descriptive Notes	Não	Não	Condições de Reprodução	Sim
IX-Part 2	View Description	Sim	Não	Mídias Relacionadas	Não
NA	NA	NA	NA	Outros Números	Não
Legenda: NA (não aplicado)					

Fonte: elaborado pelos autores.

Considera-se importante salientar que o capítulo VI, parte 2 do guia, dedicado ao elemento central "assunto" (*Subject*), não é considerado nos elementos de descrição para identificação do bem cultural de caráter museológico no INBCM, sugerindo que esse tipo de representação temática não é relevante para o contexto dos museus vinculados ao Ibram. Logo, o experimento da presente pesquisa considerou 7 dimensões analíticas.

2.2 Regras de catalogação mapeadas e selecionadas no CCO

Cada um dos elementos CCO (Quadro 1) possui regras específicas de uso e preenchimento, com vista a nortear o catalogador no uso de regras claras para o preenchimento do conteúdo dos dados. De acordo com o mapeamento realizado em todas as regras explicitadas nos capítulos ora elencados do guia (I, II, III, IV, V, VII e VIII) foram identificadas 244 regras, incluindo 122 regras distintas.

Contudo, dentre o conjunto de regras mapeadas, foram elencadas apenas as regras pertencentes aos elementos descritivos alinhados com o INBCM, que não apresentassem fator subjetivo e que, portanto, pudessem inviabilizar tecnicamente a avaliação por algoritmo computacional. Uma regra inviável, por exemplo, seria "Caso o local não tenha nome, registre o nome do local mais próximo" apontado pelo CCO para o elemento Local. Assim sendo, foram elencadas 51 regras, incluindo 17 regras distintas, organizadas em 7 dimensões analíticas, consideradas viáveis e, portanto, usadas no experimento.

2.3 Automação dos métodos aplicados na exploração das bases de dados

Para a obtenção dos dados para a presente pesquisa, foi desenvolvido um script (GITHUB, 2022) por meio da utilização da linguagem de programação *Python*, e com o uso das bibliotecas *Pandas*, *BeautifulSoup* e *Requests*, para realizar a exportação em massa de todos os dados dos acervos dos museus no formato "CSV: inbcm-ibrammapper". Ressalta-se que esse formato é um dos disponíveis para exportação no *software* Tainacan.

No que se refere às regras implementadas no algoritmo em *Python* (GITHUB, 2022), as avaliações dos dados das bases foram determinadas a partir dos fundamentos do campo da catalogação descritiva (GILLILAND, 2016) quanto às orientações acerca do uso de padrões para tratamento nos dados, focando padronização, normalização e qualidade nas descrições em ambiente digital, quais sejam: padrão de conteúdo de dados e padrão de valor de dados nos acervos dos museus envolvidos na análise.

No caso do padrão de conteúdo de dados, a avaliação do dado foi implementada no algoritmo a partir de uma técnica conhecida como Expressões Regulares (CROCHEMORE; RYTTER, 1994, p. 157). Expressões Regulares (*regex*) são escritas em uma linguagem formal e podem ser interpretadas por um processador de expressão regular. Um processador de expressão regular é um programa que serve como um analisador sintático ou examinador de

texto, identificando as partes que casam com a especificação dada, neste caso a *regex*. Várias linguagens de programação possuem formas diferentes de lidar com *regex*. No *Python*, há uma biblioteca chamada *re*[1]que trabalha bem com *regex*, e esta foi utilizada no algoritmo da presente pesquisa.

Já para o padrão de valor de dados, a avaliação da utilização de vocabulário controlado foi feita a partir dos dados disponibilizados pela API do Tainacan, disponível no painel de exportação com nome "API do Tainacan em formato JSON". Essa API disponibiliza dados para além dos elementos de metadados do INBCM, e que indicam se a configuração do elemento de metadado é do tipo taxonomia para uma determinada coleção.

Para cada regra (documentada em GITHUB, 2022) associada ao elemento de metadado pertencente a uma dimensão, o registro de dado correspondente (string avaliada) recebeu o valor 0 ou 1. O valor 1 foi atribuído quando o registro de dado atendeu ao critério (regra) recomendado pelo CCO; e o valor 0 quando não atendeu. Por fim, o índice de adequação é dado pela seguinte fórmula:

índice b = (∑ Valor1 / (∑ Valor1+ ∑ Valor0)) *100

Onde:

- "b" é a base com a amostra de dados de uma coleção em particular.
- índice é o percentual de adequação obtido a nível de dimensão, elemento de metadado e regra de catalogação para um determinado museu e coleção.
- Valor1 é a indicação de ocorrência do registro de dado que **se adequou** a regra.
- Valor0 é a indicação de ocorrência do registro de dado que **não atendeu** a regra.

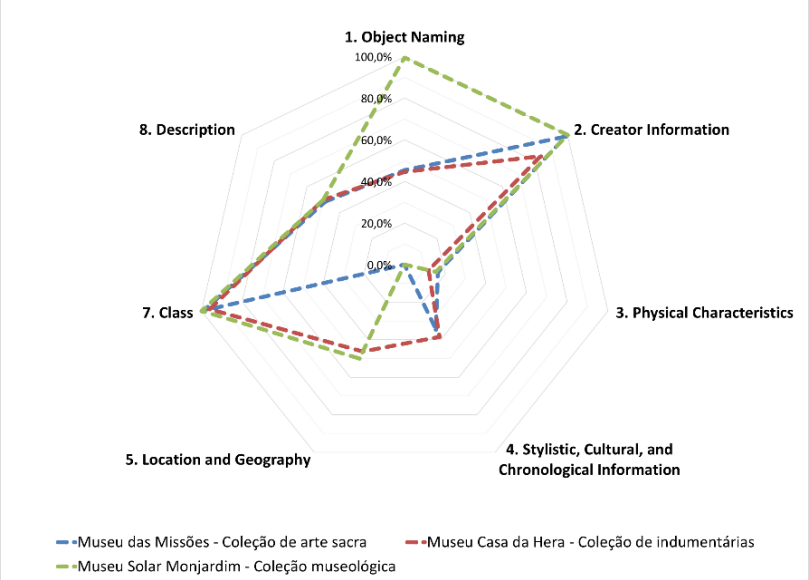
3 RESULTADOS E DISCUSSÕES

Os índices de adequação das coleções selecionadas podem ser observados no Gráfico 1 em que é possível observar a taxa de adequação das 3 coleções museais, cujas bases de dados foram descritas pelas orientações do INBCM, frente às 7 dimensões analíticas.

Como resultado geral para os 3 museus, destaca-se o alto índice de adequação média das dimensões *Creator information* (94,2%) e *Class* (98,3%). Por outro lado, é visível o baixo índice de adequação média para as dimensões *Physical Characteristics* (14,5%), *Stylistic, Cultural, and Chronological Information* (24,3%), *Location and Geography I*32,1%) *e Description* (49,6%).

[1] https://docs.python.org/3/library/re.html

Gráfico 1 – Diagnóstico da adequação dos metadados INBCM às dimensões CCO.



Fonte: elaborado pelos autores.

Como destaque para a coleção de Indumentárias do Museu Casa da Hera, é possível observar que a dimensão *Creator Information* tem uma adequação positiva de 83,6%; seguida da *Class* com 96,0%, e *Location and Geography* com 46,3%. Observou-se que o principal causador da inadequação desta coleção foi a quantidade de metadados com valor nulo (vazio). Em destaque, o elemento CCO *Location* alinhado ao elemento *Situação* do INBCM com 100% de inadequação por estar vazio e possuir a orientação de obrigatoriedade de preenchimento tanto no COO quanto no INBCM. Tal prática de catalogação por parte de instituições museais pode acarretar a ausência do registro de um item numa dada situação de busca e recuperação de informação, e, consequentemente, prejudicar a criação de possíveis índices numa solução de busca agregada (SIQUEIRA; MARTINS, 2021).

Na coleção de Arte Sacra do Museu das Missões é possível observar as dimensões *Object Naming* com um índice de adequação de 45,5%; *Creator Information* com 98,9%; e *Class* com 98,9%. Apesar do *Object Naming* estar entre as dimensões com maior taxa de adequação, a mesma foi fortemente impactada pela ausência de valores no elemento de metadado *Work Type*. A dimensão *Creator Information* recebeu alta taxa de adequação apesar de a maioria dos registros terem o valor de dado "Não identificado", o que cumpriu bem as regras de não ficar vazio, usar termos controlados e não apresentar abreviações. A dimensão *Class* também teve destaque pelo fato de os metadados associados fazerem uso de vocabulário controlado, não apresentarem plural e por possuírem valores de dados em boa

parte dos registros. Como destaque negativo, as dimensões *Physical Characteristics* com um índice de adequação de 16,4%; *Stylistic, Cultural, and Chronological Information* com 34,4%; e *Location and Geography* com 0,0%.

Na coleção museológica do Museu Solar Monjardim é possível observar as dimensões *Object Naming* com um índice de adequação de 99,5%; Creator *Information* com 100%; e *Class também* com 100%, resultando numa adequação com o COO bem positiva, e, portanto, qualificada, para os elementos de metadados envolvidos na base de dados deste museu. Como destaque negativo, tem-se a dimensão *Physical Characteristics* com índice de 15,4% e Stylistic*, Cultural, and Chronological Information* com 0,0%.

Finalmente, o uso de vocabulário controlado é recomendado por 9 dos 14 elementos de descrição alinhados entre o CCO e INBCM. Desses 9, apenas 5 deles apresentaram algum índice de adequação nas coleções avaliadas, a saber: *Work Type*, *Creator*, *Materials and Techniques*, *Creation Location* e *Class*. Essa última dimensão, alinhada com o elemento INBCM Classificação, foi a que apresentou melhor índice de adequação dentre as coleções analisadas. Pondera-se, assim, que o uso de taxonomia pelos museus do Ibram, por meio do metadado classificação, é um aspecto importante na qualidade de dados das coleções, pois normaliza e padroniza a terminologia que será usada nos processos de busca e recuperação da informação (LANCASTER, 2004), além de ajudar no alcance da interoperabilidade semântica dos dados entre diferentes esquemas de metadados e aplicações (ZENG, 2019).

4 CONSIDERAÇÕES FINAIS

A presente pesquisa apresentou o resultado de uma avaliação diagnóstica semiautomática realizada em coleções de 3 instituições museais vinculadas ao Ibram: Museu Casa da Hera, Museu Solar Monjardim e Museu das Missões. O propósito foi verificar o quão adequados encontram-se os dados dessas coleções em relação às boas práticas de catalogação indicadas no guia de referência no campo da cultura digital, o CCO.

Mesmo sendo uma pesquisa em andamento, realizada apenas em 3 coleções, o presente diagnóstico já apresentou alguns resultados interessantes que podem servir de insumo para que práticas de documentação maduradas, como as orientadas pelo CCO, sejam incorporadas na modelagem de metadados das bases de dados dos museus sob gestão do Ibram.

Por fim, como continuidade de pesquisa, almeja-se estender o diagnóstico para todas as 23 coleções dos 20 museus digitais disponíveis atualmente pelo Ibram, dando condição de se obter um conjunto de resultados mais completo para fins de proposição de um esquema de metadados formal à normativa do INBCM, intencionando a oferta de um serviço de busca e recuperação aprimorado a partir das necessidades de informação do usuário final.

REFERÊNCIAS

BRASIL (País). Ministério da Cultura. **Resolução Normativa n. 6, de 31 de agosto de 2021**. Disponível em: https://www.in.gov.br/web/dou/-/resolucao-normativa-ibram-n-6-de-31-de-agosto-de-2021-342359740. Acesso em: 12 ago. 2022.

CROCHEMORE, Maxime; RYTTER, Wojciech. **Text algorithms**. New York: Oxford University Press, 1994.

GILLILAND, Anne Jervois. Setting the stage. In: **Introduction to metadata**. 3. ed. Los Angeles: Getty Research Institute: Murtha Baca, 2016.

GITHUB. **AbeilCoelho. Qualidade_dados_IBRAM**. 2022. Disponível em: https://github.com/AbeilCoelho/Qualidade_dados_IBRAM. Acesso em: 22 jul. 2022.

INTERNATIONAL FEDERATION OF LIBRARY ASSOCIATIONS AND INSTITUTIONS (IFLA). **Declaração dos Princípios Internacionais de Catalogação**. Haia, 2016. Disponível em: https://www.ifla.org/wp-content/uploads/2019/05/assets/cataloguing/icp/icp_2016-pt.pdf. Acesso em: 22 maio 2022.

LANCASTER, Frederic Wilfrid. **Indexação e resumos**: teoria e prática. 2. ed. Brasília: Briquet de Lemos, 2004.

MARTINS, Dalton Lopes *et al*. Information organization and representation in digital cultural heritage in Brazil: Systematic mapping of information infrastructure in digital collections for data science applications. **Journal of the Association for Information Science and Technology**, [S. l.], p. asi.24650, 2022. DOI: 10.1002/asi.24650.

SILVA, Camila Aparecida Da; LARA, Marilda Lopes Ginez De. Esquema básico de metadados para representação descritiva de obras de arte em museus brasileiros. **Transinformação**, [online], v. 33, p. e200050, 2021. DOI: 10.1590/2318-0889202133e200050.

SIQUEIRA, Joyce; MARTINS, Dalton Lopes. Painel de visualização analítica dos acervos digitais integrados do instituto brasileiro de museus: o uso das tecnologias Tainacan e Elastic Stack. In: **XXI Encontro Nacional de Pesquisa e Pós-graduação em Ciência da Informação**, 2021, Rio de Janeiro. XXI Enancib, 2021. Disponível em: https://enancib.ancib.org/index.php/enancib/xxienancib/paper/view/95. Acesso em 21 mai. 2022.

ZENG, Marcia Lei. Interoperability. **Knowledge Organization**, v.46, n.2, p. 122-146, jan. 2019. Disponível em: https://www.ergon-verlag.de/isko_ko/downloads/ko_46_2019_2_d.pdf. Acesso em: 03 mai. 2021.