PROJETO TAINACAN

Relatório referente ao produto Q do 6º Termo Aditivo do TED UFG e Ibram

Implementação em produção de ferramenta de interoperabilidade semântica dos acervos para busca e recuperação integrada

Setembro 2021

ibram
instituto brasileiro de museus

UFG
UNIVERSIDADE FEDERAL DE GOIÁS

TAINACAN

Sumário

1. Introdução

Este relatório é referente ao Produto Q, do 6º Termo Aditivo do TED UFG e Ibram, e descreve o processo de implementação do protótipo da ferramenta de interoperabilidade semântica dos acervos do Ibram para busca e recuperação integrada, intitulada de Brasiliana Museus.

Parte do processo para agregação de dados foi descrito no Relatório referente ao Produto S[1], do 6º Termo Aditivo do TED UFG e Ibram: Painel de monitoramento de dados e estatísticas centralizado da rede de acervos, que foi desenvolvido utilizando as ferramentas do *Elastic Stack*, que reúnem os projetos *open source*: *ElasticSearch*, *Beats*, *Logstash* e *Kibana*, base para agregação dos dados.

Em 2021, 18 museus, dos 30 geridos pelo Ibram, foram migrados para o Tainacan e estão disponíveis para acesso via web. Com este panorama e com o número crescente de museus migrados, o Laboratório vem realizando diversos estudos com o objetivo de criar um novo serviço de informação, o Brasiliana Museus, um agregador de objetos digitais culturais, dos acervos de diferentes museus, que visa oferecer uma interface única de busca e recuperação de informações, utilizando o Tainacan como repositório final.

Assim, são apresentadas as etapas percorridas para o desenvolvimento do protótipo do serviço de agregação, com ênfase nas soluções tecnológicas utilizadas, no caso, a pilha *Elastic Stack* e o repositório digital Tainacan, assim como o *workflow* de agregação, que traz cinco etapas: (1) migração do acervo digital para o Tainacan; (2) coleta automatizada dos acervos; (3) agregação dos acervos e transformação dos dados; (4) submissão dos itens no Tainacan e (5) publicação dos acervos agregados para busca e recuperação dos itens museológicos brasileiros.

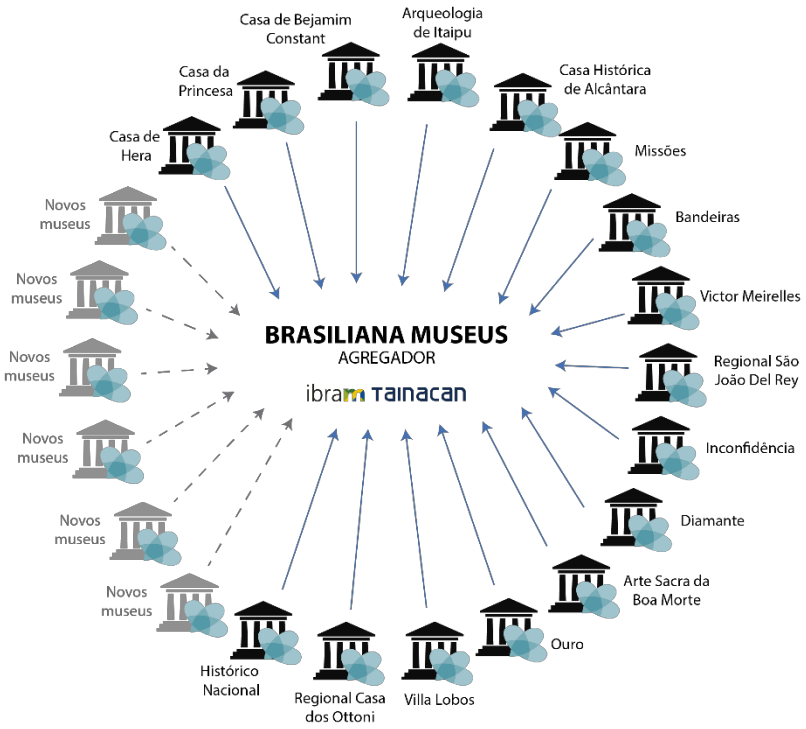
2. Metodologia

O serviço de agregação a princípio, é composto por acervos digitais de 16 museus, conforme pode ser visto na Figura 1, sendo 17 coleções, resultando em 15,312 objetos digitais, sendo eles: Museu Casa da Hera, com duas coleções, Acervo Museológico (1,124) e Indumentária (31); Museu Casa da Princesa (996); Museu Casa de Benjamin Constant (983);

[1] https://drive.google.com/file/d/1BAsWAoExstrApz_vwo-1Nj3-iThYGiBf/view

Museu Casa Histórica de Alcântara (631); Museu da Inconfidência (4,624); Museu das Bandeiras (401); Museu das Missões (90); Museu de Arqueologia de Itaipu (1,040); Museu de Arte Sacra da Boa Morte (783); Museu do Diamante (895); Museu do Ouro (101); Museu Histórico Nacional (773); Museu Regional Casa dos Ottoni (463); Museu Regional São João Del Rey (328); Museu Victor Meirelles (237) e Museu Villa Lobos (1,812). Os demais museus serão, à medida que migrados, incluídos no agregador.

Figura 1. Acervos museológicos agregados na Brasiliana Museus



Fonte: elaborado pelos autores (2021)

Para efetivar a agregação dos dados foram utilizadas ferramentas gratuitas e de código aberto: Tainacan[2] e *Filebeat* e *Logstash*, do *Elastic Stack*[3], além de um plugin para *Logstash* desenvolvido pelos autores, utilizando a linguagem de programação *Ruby*[4].

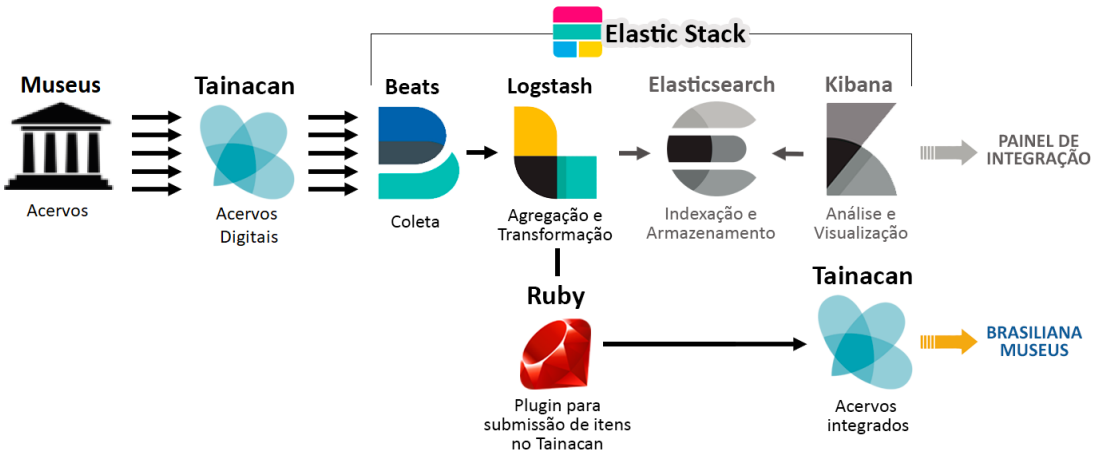
As instalações dos museus provedores de dados utilizam o Tainacan versão 0.15 ou superior. O Tainacan de agregação utiliza a versão mais recente. O *Filebeat* e *Logstash* utilizam a versão 7.9. Para que todo processo funcione corretamente é preciso que o Tainacan esteja com a versão mínima 0.15 e o *Filebeat* e *Logstash*, 7.9.

3. Processo de Agregação de Acervos Museológicos

O Brasiliana Museus é um agregador de objetos digitais culturais dos museus administrados pelo Ibram, que visa oferecer um painel de integração e uma interface única de busca. Destacamos alguns dos principais benefícios de se disponibilizar um agregador de dados culturais: oferecer recursos para melhorar a gestão da documentação museológica, pois apresenta, de forma unificada e interativa, uma visão macro dos museus; demonstrar inconsistências sintáticas ou semânticas, possibilitando aos museus melhorarem, na base, suas informações; comparar diferentes instituições, tornando possível encontrar padrões e conexões de interesse, encontrando indicadores que possam garantir um melhor planejamento estratégico, maior produtividade e agilidade no acesso aos dados, gerando efeitos nas dimensões administrativas, culturais e educacionais; enriquecer a fonte de dados para pesquisa; permitir aos usuários uma busca simplificada dos objetos museológicos sem a necessidade de conhecimentos mais avançados sobre os museus e; monitorar o acervo em comparação aos demais.

O processo de agregação dos acervos museológicos geridos pelo Ibram é realizado em cinco etapas distintas: etapa 1, migração do acervo digital para o Tainacan; etapa 2, coleta automatizada dos itens dos acervos; etapa 3, agregação dos acervos e transformação dos dados; etapa 4, submissão dos itens no Tainacan e etapa 5, publicação dos acervos agregados para busca e recuperação dos itens museológicos brasileiros. A Figura. 2 apresenta sua arquitetura, considerando, em cinza, as etapas para criação do Painel Integrado, descrito no relatório do Produto S, do 6º Termo Aditivo do TED UFG e Ibram.

Figura 2. Arquitetura do Brasiliana Museus



Fonte: Elaborado pelos autores (2021)

Cada etapa está descrita a seguir.

3.1 Migração do acervo digital para o Tainacan

Uma vez que o museu possui seu acervo, parte ou integralmente digitalizado e sua documentação consolidada, ocorre a primeira etapa do processo de agregação, que consiste na migração para o repositório Tainacan, contemplando passos que garantam a viabilidade e a qualidade do acervo. A Figura 3 demonstra os passos para a efetivação da migração, que podem sofrer algumas variações a depender da realidade de cada instituição.

Figura 3. Processo de migração do acervo do museu administrado pelo Ibram para uma base de dados no Tainacan

Fonte: Elaborado pelos autores (2021)

No passo 1, Análise, as características técnicas dos acervos, como padrões de metadados, políticas de direitos autorais e de digitalização, regras de catalogação, linguagens documentárias e softwares utilizados, são elencadas. No passo 2, Coleta, os dados são extraídos da solução atual, como softwares legados, fichas catalográficas, físicas ou digitalizadas, planilhas eletrônicas, documentos de texto, arquivos em PDF ou bases de dados relacionais e convertidos para o formato *Comma-Separated Values* - CSV, possibilitando o tratamento dos dados. Para cada extração são utilizadas diferentes estratégias de Ciência de Dados, como, por exemplo, *script* em *Python*. No passo 3, Tratamento, são realizados procedimentos de normalização, limpeza e preparação dos dados, utilizando a ferramenta *OpenRefine*[6] e fórmulas em linguagem *General Refine Expression Language* - GREL[7]. A normalização consiste em agrupar termos comuns (ex.: João José, J.J, João J, J. José), padronizar caixa-alta e caixa-baixa, separadores etc.; a limpeza, em realizar correções gramaticais e retirar termos estranhos (ex.: "???", "/" " - ", além de caracteres sem sentido, como um valor de texto em um campo para valores monetários) e, por fim, os dados são adequados ao formato exigido pelo importador do Tainacan. No passo 4, Migração, os dados são modelados e migrados para uma nova base de dados no sistema Tainacan, via funcionalidade de importação do software[8]. No passo 5, Validação, os responsáveis pelo museu navegam e realizam a busca e recuperação de dados para identificar eventuais problemas, caso esteja tudo correto, o acervo é publicado na rede.

3.2 Coleta automatizada dos itens dos acervos

No contexto do Tainacan, "item" se refere ao conteúdo propriamente dito, representado por pinturas, filmes, livros e etc. É o conjunto de um documento (mídia, texto ou URL), metadados e, também, eventuais documentos em anexo, que são organizados em coleções (TAINACAN.ORG, 2021b). No caso da agregação, apenas os metadados são coletados e a eles, incluídos os metadados que armazenam o nome do museu e o id e o URL do item no museu de origem, para acesso direto do usuário. Dessa forma, durante a agregação, os documentos ou anexos são não considerados.

Por padrão, o Tainacan oferece *URLs* alternativos para visualização dos dados nos formatos *Application Programming Interface JavaScript Object Notation - API JSON*; *HyperText Markup Language - HTML* e arquivos CSV (TAINACAN.ORG, 2021a), possibilitando que novas aplicações possam se beneficiar dos dados. Especificamente no caso Ibram, utiliza-se a API no formato JSON.

No entanto, apenas reunir os dados dos museus não é suficiente. É preciso assegurar que estes dialoguem entre si, ou seja, garantir que todos possuam o mesmo padrão de metadados, resultando em uma mesma sintática e semântica na agregação. Tal procedimento é denominado na literatura como *Crosswalk:* os elementos de um esquema de metadados são associados a outro, viabilizando a interoperabilidade e permitindo que coleções heterogêneas possam ser pesquisadas simultaneamente (CHAN & ZENG, 2006).

O Tainacan, a partir da versão 0.15, permite, por meio de um plugin desenvolvido para o Ibram, mapear os metadados para o padrão normatizado pelo Ibram e intitulado como Inventário Nacional dos Bens Culturais Musealizados – INBCM (BRASIL, 2014). Dessa forma, os elementos de metadados constituintes dos esquemas das bases de dados dos museus foram mapeados para o INBCM, no próprio Tainacan. Após o mapeamento, o Tainacan passa a disponibilizar a *API JSON*, "JSON simples: inbcm-ibram mapper", que será coletada pelo *Filebeat*.

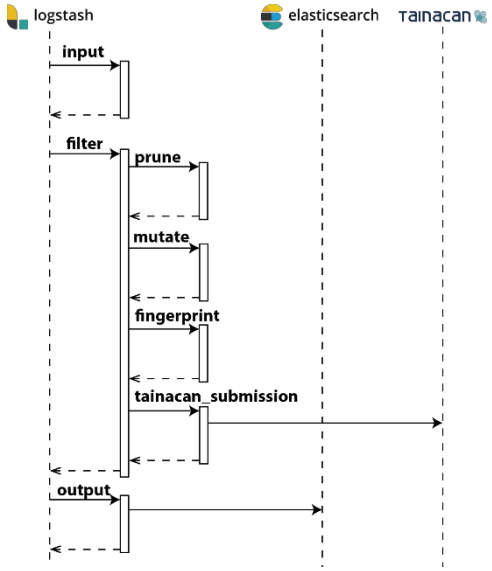
Foram utilizados dois arquivos de configuração para o *Filebeat*, o primeiro, se trata da configuração necessária para cada coleção, de cada museu, que terá seus dados coletados. De forma geral, o *Filebeat* realiza o acesso a API no formato JSON especificado no arquivo de configuração, recebendo os metadados mapeados e o id e o URL do item e, por fim, adiciona

um novo metadado com o nome do museu de origem. O segundo, se trata de um arquivo de configuração geral, que orquestra o acesso aos arquivos individuais (.yml) e envia seus resultados ao *Logstash*.

3.3 Agregação dos acervos e transformação dos dados

No *plugin input*, os dados são recebidos do *Filebeat*. Na sequência, são executados três filtros padrão do *Logstash: prune, mutate e fingerprint*, e um, o tainacan_submission, desenvolvido pelo Laboratório utilizando a linguagem de programação Ruby, com a finalidade de submeter os itens a uma instalação Tainacan. No *plugin output* os dados são enviados para o armazenamento no *Elasticsearch*. A Figura 4 apresenta o *pipeline* com as principais etapas.

Figura 4. Pipeline do Logstash



Fonte: Elaborado pelos autores (2021)

O filtro *prune* executa a limpeza dos dados recebidos, deixando apenas os dados relevantes, configurados no *Logstash*, removendo quaisquer outros dados enviados pelo *Filebeat*. O filtro *mutate* realiza alterações nos campos, assim, foram realizadas as alterações: *capitalize*[9], que transforma os dados em caixa alta; *lowercase*[10], que transforma os dados em caixa baixa; *remove_field*[11], que remove campos desnecessário; *rename*[12], que renomeia os campos para nomes mais adequados; *strip*[13], que retira espaços que possam existir antes ou depois do dado e *split*[14], que separa os dados em campos menores.

Na sequência, é utilizado o *Fingerprint filter*, que cria um novo campo, uma espécie de novo id, que garante a unicidade do documento, permitido que novas inserções de dados, não gerem dados duplicados, principalmente. Para criação do código único, foi criado um *hash* utilizando três campos: a instalação, ou seja, o nome do museu, o id e o número de registro do item, coletados dos museus de origem.

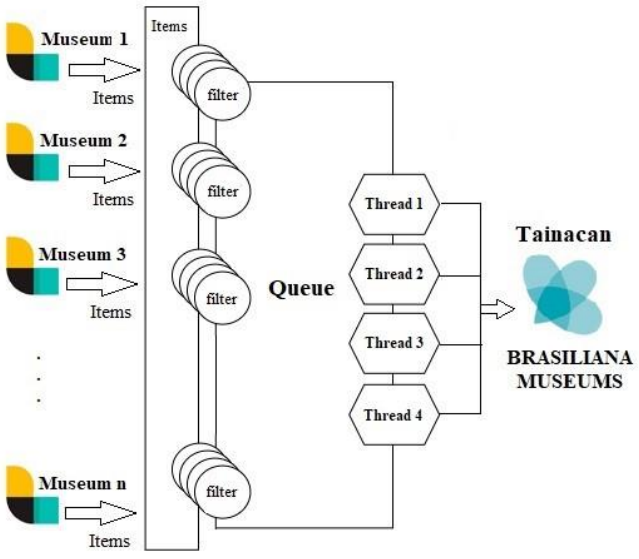
O último filtro, o *tainacan_submission*, foi desenvolvido pelo Laboratório para realizar a submissão dos itens coletados no Tainacan.

3.4 Submissão dos itens no Tainacan

O filtro desenvolvido para *Logstash*, *tainacan_submission*, tem a finalidade de recuperar os dados processados pelo *Logstash* e enviá-los para uma nova instalação do Tainacan, por meio de sua API de submissão[15].

Para melhor desempenho das submissões, foi implementada uma fila, Figura 5, de processos que as organizam e encaminham para a API do Tainacan, visto que, o *Logstash* envia um alto número de itens, de forma simultânea. Todos os itens processados são incluídos no final da fila, enquanto um conjunto de *threads* (quatro, por padrão, na Figura 5), são responsáveis por processar as saídas dos itens dessa fila, efetivando a submissão do item no Tainacan.

Figura 5. Esquema da fila de submissão



Fonte: Elaborado pelos autores (2021)

Após todos os itens da fila serem processados, estes estarão disponíveis para acesso na instalação do Tainacan. É importante ressaltar que após esta etapa, os itens também serão enviados para o *Elasticsearch*.

4. Considerações Finais

A evolução tecnologia vem propiciando a reinvenção do espaço museal, abrindo caminhos potenciais de difusão e socialização dos seus acervos. Dessa forma, instituições internacionais e nacionais, responsáveis pela guarda de acervos do patrimônio cultural, têm investido na disponibilização de seus acervos digitais, resultando em inúmeros acervos digitais de museus distribuídos na *web*, gerando assim diferentes possibilidades de busca e recuperação da informação.

Esse cenário, embora muito rico, trouxe desafios à descoberta dos recursos do patrimônio cultural. Entretanto, para otimizar a localização dos objetos digitais presente em diferentes sítios de museus espalhados pela internet, diversas instituições ao redor do mundo se tornaram provedores de dados para serviços especializados na agregação de acervos culturais do patrimônio cultural, como é o caso da Europeana e da DPLA. Nesse sentido, a agregação torna a busca e a recuperação de informações sobre acervos uma operação mais eficiente para os usuários destes acervos.

Por exemplo, para se encontrar um objeto digital específico, é preciso conhecer o acervo ao qual este pertence e algumas de suas características, o que para um usuário leigo seria praticamente impossível, limitando, sobremaneira, a real difusão do acervo. Por isso, prover aos museus e aos usuários serviços para busca e recuperação eficientes se torna um tema essencial.

Nesse sentido, o estudo e a prototipagem do serviço de agregação de acervos digitais de museus, o Brasiliana Museus, se revela como uma solução eficiente, tanto do ponto de vista da socialização e difusão dos objetos digitais, quanto no acompanhamento e monitoramento das informações dos acervos por meio dos painéis analíticos, podendo mudar a forma como gestores, pesquisadores e usuários interagem com dados culturais, trazendo inúmeros

benefícios à sociedade. Além de apresentar uma solução tecnológica barata e viável de ser implantada.

Como trabalhos futuros, a pesquisa continua e melhorias serão realizadas do decorrer dos estudos, sendo o próximo passo, a coleta das miniaturas das imagens dos objetos digitais dos museus de origens para o agregador.

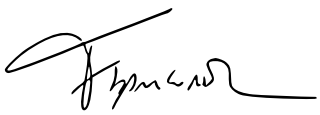
Referências

Brasil: Resolução Normativa n.2, de 29 de agosto de 2014. Diário Oficial da República Federativa do Brasil (2014), https://pesquisa.in.gov.br/imprensa/jsp/visualiza/index.jsp?jornal=1pagina=14data=01/09/2014, last accessed on 07/09/2021.

Chan, L.M., Zeng, M.L.: Metadata Interoperability and Standardization – A Study of Methodology Part I. D-Lib Magazine, 12(6), pp. 1082–9873 (2006), https://doi.org/10.1045/june2006-chan

Tainacan.org: Páginas do Tainacan (2021a), https://tainacan.github.io/tainacan-wiki//tainacan-pages?id=as-páginas-especiais-do-tainacan, last accessed on 07/09/2021.

Tainacan.org: Tainacan. Itens (2021b), https://tainacan.github.io/tainacan-wiki/#/pt-br/items, last accessed on 07/10/2021.



Prof.ª Dr.ª Flavia Maria Cruvinel
Coordenadora Projeto Tainacan