ProgDOC

PRODUTO 02

Projeto 914BRZ4019 – Unesco - Contrato nº. SA-47/2018

Projeto: Consultor em documentação de acervos digitais indígenas

Pesquisador: André Luiz Dadona Benedito

Data: 23/05/2018

Produto 2

a) Estudo de interconexão entre a plataforma livre de acervo digital Tainacan, o site e as bases de dados do Museu do Índio visando o desenvolvimento do site institucional integrado ao acervo digital em plataforma livre;

b) primeiro relatório técnico da instalação e estruturação na plataforma livre de acervos digitais, dos bancos de dados de outros sistemas do Museu do Índio: PHL, ATOM.

Sumário

1. Introdução

A plataforma livre de acervo digital Tainacan, a partir de sua proposta de repositório digital, integra bases de dados de diversas fontes de maneira manual, através do preenchimento de coleções e itens, ou pela interconexão com bases de dados já existentes, pelas ferramentas de importação, permitindo assim a migração de dados armazenados em sistemas de informação.

O processo de integração do Tainacan com bases de dados de terceiros se torna um desafio quando os formatos não são os mesmos, já que a plataforma oferece suporte para arquivos no formato de planilha via interface do usuário, e via banco de dados relacional no formato SQL (Linguagem Estruturada de Consulta) pelo sistema Wordpress.

Sendo assim, qualquer processo de integração de dados à plataforma que estejam em formatos diferentes dos suportados envolve a conversão destes um formato utilizável pelo Tainacan para importação de dados, além disso é um processo que necessidade de um acompanhamento especial, com atenção à integridade dos dados, eles devem ser migrados com a mesma representatividade do sistema anterior, nenhum dado pode ser negligenciado.

A partir da parceria do Museu do Índio com o Laboratório de Políticas Públicas Participativas/UFG (L3P), esse desafio ganha materialidade ao se propor a interconexão dos dados do museu com a plataforma Tainacan, já que esses dados estão em um sistema denominado PHL Elysio, que possibilita a exportação de dados somente no formato XML.

O PHL Elysio[1] é um sistema de gerenciamento de informações desenvolvido pelo professor Elysio Mira Soares de Oliveira para “administrar coleções, serviços de biblioteca e centros de informações”, ele utiliza uma interface web para lidar com a consulta e preenchimento de dados de repositórios, e o modo como foi construído foi voltado para que as funcionalidades ocorressem nesse meio, assim toda a estrutura do banco de dados só pode ser exportada no formato XML.

O XML (Figura 1) é a denominação da Linguagem de Marcação Extensível, ou seja, é uma estrutura utilizada para organizar dados em pequenos compartimentos chamados *tags*, elas são expressas no formato: *<item_museológico>Nome do Item*

1 PHL Elysio – Site: http://www.elysio.com.br/

Museológico</item_museológico>, o dado representado, no exemplo, o nome do item, fica entre duas marcações que representam o atributo, nesse caso, o item museológico. Na figura 1 é apresentada uma visualização parcial do documento obtido da base de Tombo no museu:

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<db>
  <rec>
    <v800>1000203172651 91A01</v800>
    <v801>916art</v801>
    <v807>2</v807>
    <v819>3</v819>
    <v820>00000000</v820>
    <v997>2</v997>
    <v998>1</v998>
    <v999>^d20070429^h0816^bsuper</v999>
    <v999>^d20180315^h1212^bIone</v999>
  </rec>
  <rec>
    <v800>1000208153814 07A01</v800>
    <v801>718art</v801>
    <v803>1</v803>
    <v807>2</v807>
    <v819>3</v819>
    <v820>19970700</v820>
    <v823>R$ 4.000,00 (quatro mil reais) ou US$ 2.000 (dois mil dólares)</v823>
    <v997>20</v997>
    <v998>15</v998>
    <v999>^d20081113^h1513^bFabiane</v999>
    <v999>^d20120510^h1406^bfabiana</v999>
  </rec>
</db>
```

Figura 1 – Exemplo de documento XML (Base de Tombo – Museu do Índio)

Nesse exemplo, a *tag* “<db></db>” significa que o que tiver dentro dela é o banco de dados, a *tag* “<rec></rec>” representa os registros desse banco (no exemplo são dois registros), e para cada registro existem, no caso dessa base de tombo, até 28 outras *tags*, que representam os atributos/campos da base, com seus respectivos valores entre as *tags*.

Então, o desafio de migração dos conjuntos de dados do museu foi transformar esse tipo de arquivo XML para o formato de tabela em planilha, com colunas e células, onde as colunas são os atributos, representados pelas *tags* dentro de cada registro, e as células são os valores dentro dessas *tags*.

Esse processo se deu em três macro etapas: 1 – Análise do sistema utilizado pelo museu e coleta das bases de dados; 2 – Conversão do formato da base de dados; 3 – Validação das bases convertidas. Ao concluir, foram geradas duas planilhas com o conjunto de dados exportado do sistema de informações do museu, já estruturadas para importação na plataforma livre de acervos digitais Tainacan.

Além da conversão e migração das bases de dados do sistema PHL Elysio para a plataforma Tainacan, gerando o produto das bases de formato tabular em documentos de planilha para importação, é objetivo deste relatório descrever outro produto, que é uma descrição técnica da instalação e estruturação na plataforma livre de acervos digitais, dos bancos de dados e outros sistemas do Museu do Índio.

Esse produto foi executado a partir de entrevistas com os técnicos do museu do índio, cujo objetivo foi fazer uma análise de necessidades que apontasse saídas para a otimização dos recursos de gestão da informação do museu. Resultando em um parecer técnico que recomenda a migração das bases de dados do PHL Elysio para sistemas de informação que permitem integração entre os dados.

2. Análise do Sistema e Coleta de Dados

A primeira etapa da integração dos dados Museu do Índio à plataforma Tainacan foi marcada por uma visita presencial de pesquisadores do L3P, que através do contato foi conhecida a plataforma de gestão de informações utilizada (PHL Elysio), bem como estudado meios de exportação de dados dessa plataforma, culminando na conclusão da exportação em XML devido à impossibilidade de obter os dados em outro formato.

Dessa forma, foram coletadas duas bases de dados, a Base Tombo, com informações de registro dos objetos do museu, e a base da Ficha Catalográfica contendo a caracterização dos itens. As bases no formato XML foram armazenadas em um pen drive pela equipe do laboratório para posterior processamento e conversão para integração na plataforma Tainacan.

3. Conversão do Formato da Base de Dados

Como próximo passo, foi necessário converter as bases de dados do museu em formato XML para o formato tabular de planilha para posterior importação na plataforma Tainacan. Esse processo demandou duas etapas principais: o estudo e validação dos arquivos em XML para entender a estrutura dos dados, quais os registros e campos estavam presentes em cada base e se havia alguma inconsistência; e a utilização de um *script* de conversão de dados em na linguagem de programação Python para estruturar as bases no formato de planilha.

Para ler o arquivo XML e validá-lo, foi utilizado o programa XML Viewer Plus[2], que é um simples editor de texto que valida arquivos XML, mostrando se existem ou não inconsistências. Ao abrir a base de Tombo com este programa, e utilizar a função de validar, que basicamente verifica se todas as *tags* do arquivo estão corretamente dispostas e se existe algum caractere que cause interferência na estrutura do documento, foi retornado o valor de arquivo válido. (Figura 2)

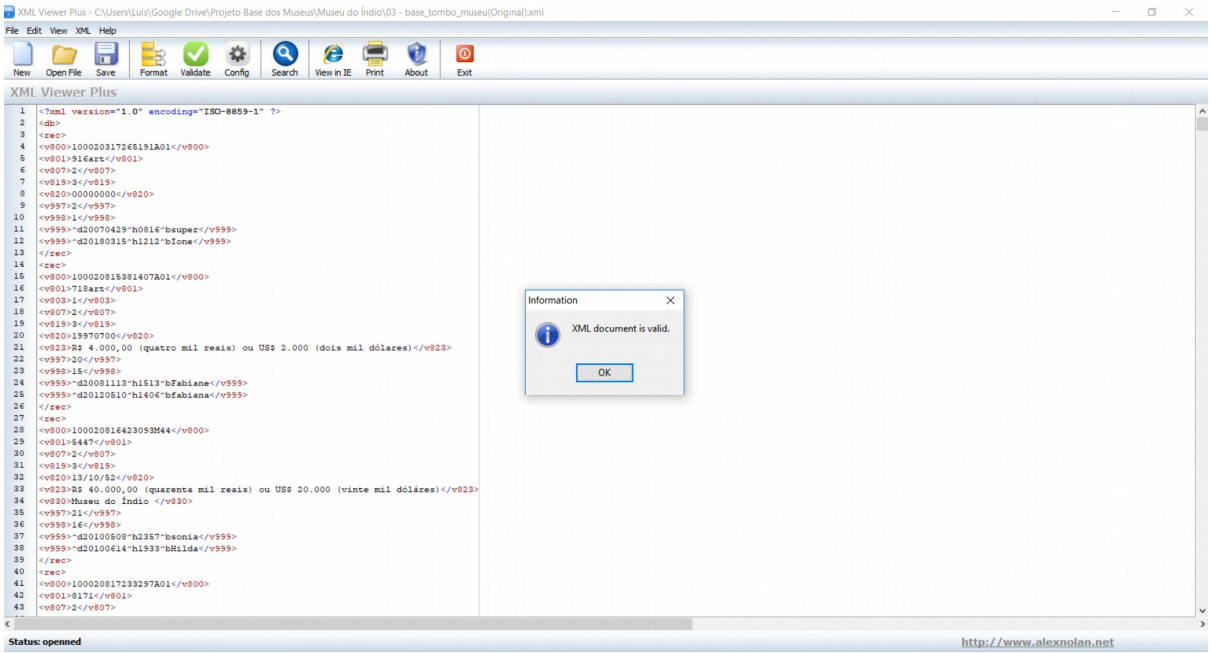

Figura 2 – Validação do arquivo XML da base de Tombo.

Já ao executar o mesmo processo com a base da Ficha Catalográfica, foram retornadas algumas inconsistências na formatação do arquivo XML (Figura 3). Os problemas encontrados tinham três características em comum, com as respectivas soluções: 1 – *Tags* de links malformadas "<Link>Descrição" foram organizados, onde

2 XML Viewer Plus – http://www.alexnolan.net/software/freexmleditor.htm

faltou a abertura ou fechamento da *tag* foi corrigido; 2 – Caractere "&" com interferência no documento foram substituídos por “e”; 3 – Caracteres "<", ">" no meio de textos sem indicação de *tags*, com interferência na leitura do XML foram removidos.

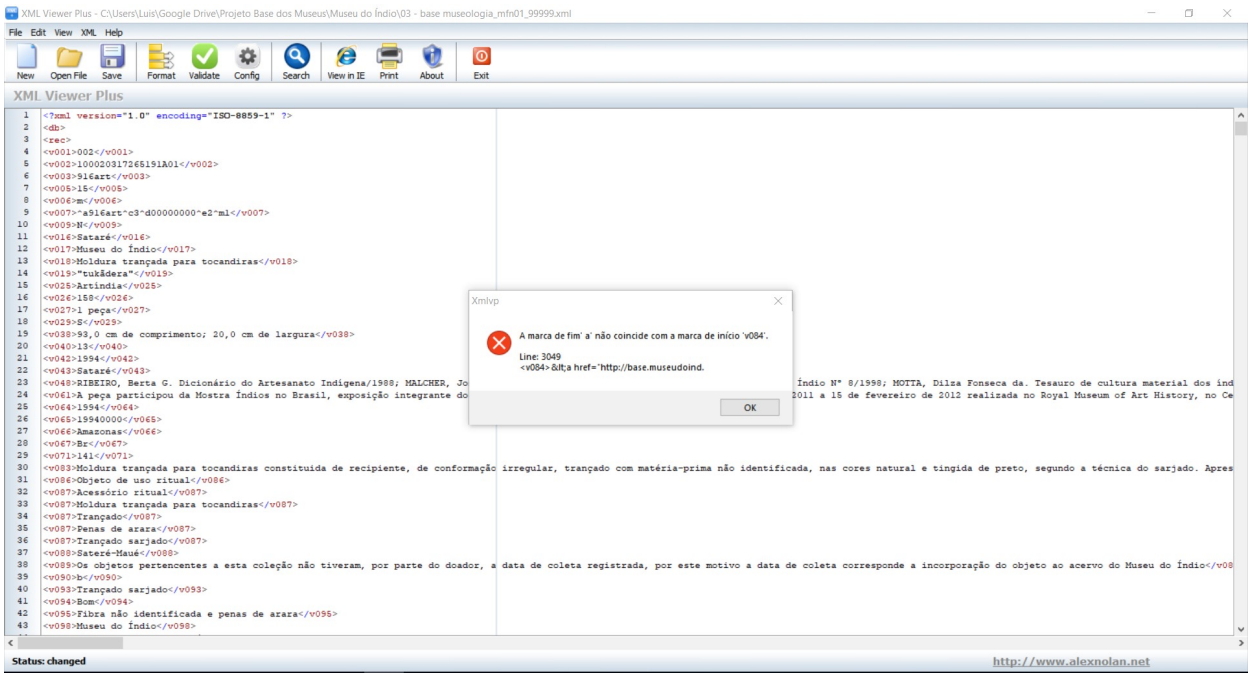

Figura 3 – Validação do arquivo XML da base da Ficha Catalográfica.

Segue alguns exemplos de inconsistências encontradas e corrigidas:

- Problema:<v084><ahref="http://base.museudoindio.gov.br/memoteca/semu/ceramica/a06/6793.html" target="_new">Xícara</v084> Linha 3049 - os caracteres <a> de abertura da *tag* do link não foram reconhecidos. Solução: Foi adicionado a abertura da tag.
- Problema:<v084><ahref="http://base.museudoindio.gov.br/memoteca/srav/semu/projeto_unesco5/ceramica/bilha_75_4_85/index.html">Bilha</a</v084> Linha 3401 - tag "a" erro na tag de fechamento "</a". Solução: Foi corrigida a *tag* de fechamento .
- Problema:<v084><ahref="http://base.museudoindio.gov.br/memoteca/srav/semu/projeto_unesco5/ceramica/tigela_2471/index.html">Tigela</></v084> Linha 9637 - tag "a" erro no caractere de fechamento "</>". Solução: Erro se repete, foi usado "Replace all" para trocar o caractere </> pela *tag* .
- Problema: Descrição com "|><|" caractere "<" dá erro no XML. Solução: Foi removido, o conjunto de caracteres na descrição ficou somente "||".

Esses problemas encontrados e corrigidos se devem a dificuldade do sistema antigo em exportar seus dados, já que alguns dos caracteres inseridos nas informações de caracterização dos objetos do museu foram confundidos com *tags* ou interferiram na leitura do documento, além disso, os links das imagens foram exportados com falhas de abertura ou fechamento de *tags,* impedindo que os dados desse atributo fossem obtidos corretamente.

A partir da validação dos arquivos XML da base de Tombo e da Ficha Catalográfica, sem opções viáveis e funcionais de transformação já existentes de arquivos XML para formato de tabela que se adequassem à estrutura das bases do museu, foi desenvolvido um código de programação na linguagem Python[3], para ler especificamente a estrutura do arquivo XML exportado pelo PHL Elysio e estruturar os dados no formato tabular, onde as *tags* de atributos são transformadas em colunas e os valores dentro dessas *tags* armazenados em células.

O código desenvolvido[4] pode ser interpretado em três partes principais:

A leitura e levantamento dos atributos do arquivo XML (Figura 4), onde é utilizado a biblioteca de funções *xml.etree.ElementTree (linha 4),* que permitiu a leitura e processamento dos dados do arquivo; e o levantamento de todas as *tags*/campos presentes no arquivo para transformá-las em um dicionário (linhas 20-24), cuja chave é o nome do campo, e os valores são os dados entre as *tags*.

```
1  # -*- coding: utf-8 -*-
2
3  #%% Importa a base em xml e faz a leitura.
4  import xml.etree.ElementTree as ET
5  import pandas as pd
6  tree = ET.parse('03 - base_tombo_museu.xml') #Nome do documento xml, deve estar na mesma pasta do script.
7  root = tree.getroot()
8
9  lista_tags = []
10 dict_base = {}
11
12 #%%Ler os elementos da base
13
14 #Pega todas as tags/colunas do xml e coloca em uma lista.
15 for element in root:
16     for tag in element:
17         lista_tags.append(tag.tag)
18 lista_tags = set(lista_tags) #Tira a repetição das tags/colunas
19
20 #Cria listas para cada tag em um dicionário.
21 for tag in lista_tags:
22     dict_base[tag] = []
23
24 print(dict_base.keys())
```

Figura 4 – Código de Conversão: Parte de levantamento de atributos e leitura do XML.

3 Python – https://www.python.org/
4 Link do Código: https://github.com/LuisMDlab/projeto_base_museus/edit/master/xml_to_Csv.py

O processo de conversão (Figura 5), em que para cada registro no arquivo XML, é verificado cada campo levantado na etapa anterior e adicionado seu valor ao dicionário criado em sua respectiva chave (campo/atributo). Além disso, para aqueles campos com mais de um valor, como o “v088” referente a descritores secundários e o “v089”, referente a informações contidas no objeto, foram tratados para serem armazenados em uma única célula, separadas por “;”.

```
#%% Processo de Conversão
lista_multiplos=[]

#Lê os valores das tags/colunas do txt e armazena nas respectivas lista no dicionário.
for i in range(len(root)):
    for tag in dict_base.keys():

        if root[i].find(tag) != None:
            if len(root[i].findall(tag)) > 1:
                lista_teste.append(tag)
                for campo in root[i].findall(tag):
                    lista_multiplos.append(campo.text)
                dict_base[tag].append(";".join(lista_multiplos))
                lista_multiplos=[]

            elif len(root[i].findall(tag)) == 1:
                if root[i].find(tag).find('a') == None:
                    for campo in root[i].findall(tag):
                        dict_base[tag].append(campo.text)
                elif root[i].find(tag).find('a') != None:
                    dict_base[tag].append(str(root[i].find(tag).find('a').attrib['href'])+';'+str(root[i].find(tag).find('a').text))

        elif root[i].find(tag) == None:
            dict_base[tag].append('Nulo')
```

Figura 5 – Código de Conversão: Processo de Conversão

Por fim, na etapa de retorno de uma planilha de dados (Figura 6), o dicionário criado e populado nas etapas anteriores a partir das *tags* e valores do arquivo XML é convertido em um formato tabular separado por vírgula (CSV), que pode ser aberto a partir de qualquer editor de planilha (Libre Office Calc ou Microsoft Excel), e assim importado pela plataforma Tainacan.

```
#%%Exportação
#Transforma o dicionário em um DataFrame para ser exportado para csv
base_csv = pd.DataFrame(dict_base)
base_csv.to_csv('base_resultante.csv', encoding='utf-8')
```

Figura 6 – Código de Conversão: Exportação da Planilha de Dados

4. Validação das Planilhas de Dados

O processo de validação das planilhas resultantes da conversão do formato XML para tabular se deu via e-mail, através do contato com o museu, em que foram enviadas as planilhas da base de Tombo e da base da Ficha Catalográfica. Na primeira tentativa de validação, foi retornado que a planilha gerada continha inconsistência nos dados, com informações em campos errados, e dados faltando.

A partir da reposta obtida pelo museu, que indicou quais os problemas na planilha de forma que facilitasse a compreensão de onde os erros se encontravas, foi identificado que alguns procedimentos no código de conversão estavam incompletos, causando as inconsistências, após a correção do código, foram enviadas novamente as planilhas resultantes da conversão via e-mail para o Museu do Índio.

O segundo processo de validação retornou com resultado positivo, indicando que os dados presentes nas planilhas realmente representavam os dados utilizados pelo museu, estabelecendo assim a conclusão deste produto que foi a conversão das bases do museu para o formato de importação na plataforma Tainacan.

5. Considerações sobre a conversão das bases

Com o produto das planilhas de importação convertidas do XML proveniente do sistema PHL Elysio utilizado pelo Museu do Índio, é sanada a necessidade de um formato comum para a conexão da base de dados do acervo museológico à plataforma Tainacan, provendo assim os produtos necessários para integração dos dados.

Esse processo se deu a partir de um esforço conjunto do L3P com a equipe do museu, em que, através do tratamento dos arquivos em XML e da construção de um código de programação personalizado para transformá-lo em formato tabular, além do auxílio constante da equipe do museu na validação dos resultados, permitiu o aperfeiçoamento e o alcance dos resultados necessários.

É uma etapa importante na integração dos dados, que foi alcançada graças à relevante parceria entre laboratório e museu, que através do empenho de ambos, o trabalho executado foi constituído com a qualidade necessária para proceder com a migração dos sistemas de informação utilizados, processo que demanda a atenção aos detalhes e dedicação de todas as partes envolvidas, como foi o processo que gerou este produto.