DADOS RS

Manual para criação de dados e metadados para publicação automatizada no dados.rs.gov.br

Divisão de Dados e Indicadores – DEE – SPGG

Subchefia de Ética, Controle Público e Transparência – Casa Civil

Versão 2.0.1

5 de abril de 2022

Sumário

1. Objetivo

Esse manual tem o objetivo de orientar equipes de TI na geração de dados e metadados para publicação automatizada na plataforma CKAN onde estão catalogados os dados das instituições que integram o Governo do Estado do Rio Grande do Sul.

2. Plataforma CKAN

A plataforma CKAN é uma das mais utilizadas para portal de dados em software livre do mundo. Fornece uma solução completa e pronta que torna os dados acessíveis e utilizáveis. Provê ferramentas para simplificar a publicação, o compartilhamento e a utilização dos dados. Para facilitar a publicação de grandes quantidades de dados, a plataforma CKAN fornece uma robusta API para manipulação de metadados. O CKAN está direcionado a publicadores de dados (governos nacionais e regionais, companhias e organizações) que querem tornar seus dados abertos e disponíveis.

Na plataforma CKAN é importante conhecer dois conceitos: *dataset* e *resource*.

O *dataset* é o objeto primário, representa um conjunto de dados e suas informações. *datasets* podem conter zero ou mais *resources*.

O *resource* contém informações sobre um arquivo de dados representado dentro do *dataset*. O CKAN permite que os dados sejam armazenados em sua própria estrutura (DataStore) ou em um servidor da web externo, indicado sempre através de URLs.

A API do CKAN utiliza requisições HTTP para realizar a comunicação. O conteúdo enviado e recebido nessa comunicação é no formato JSON.

3. Formato JSON

O JSON (*JavaScript Object Notation*) é um formato de intercâmbio de dados de padrão aberto e independente de linguagem. É um formato bastante simples e fácil de ser manipulado.

No formato JSON é importante conhecer dois conceitos: *Object* e *Array*.

Object: uma coleção não ordenada de pares atributo-valor onde os atributos (ou nomes ou chaves) são *strings*. É recomendado, mas não obrigatório, que cada atributo seja único dentro de um objeto. *Objects* são delimitados por chaves ({}) e usam vírgulas para separar cada par, enquanto que no par o atributo e o valor ficam separados por dois pontos (:).

Array: uma lista ordenada de zero ou mais valores, cada um podendo ser de qualquer tipo. *Arrays* são delimitados por colchetes ([]), dentro dos quais ficam os valores, também conhecidos como elementos, separados por vírgulas.

Exemplo:

```
{"Alunos":[

     { "nome": "Edson Sales Arantes", "notas": [ 8, 9, 5 ]  },

     { "nome": "Luiz Livelli ", "notas": [ 8, 10, 7 ] },

     { "nome": "Caique Monteiro", "notas": [ 10, 10, 9 ] }

]}
```

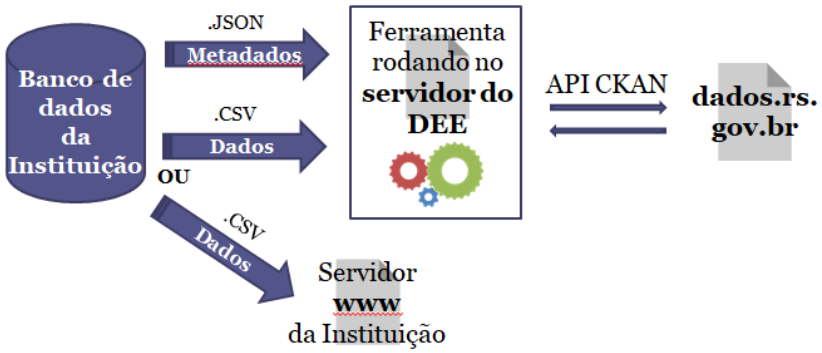
4. Publicação automatizada no portal dados.rs.gov.br

Grandes quantidades de dados demandam muito tempo para serem cadastradas manualmente no CKAN. Para isso a plataforma CKAN disponibiliza uma API para facilitar a automatização da publicação dos *datasets* e *resources*.

A comunicação com a API do CKAN pode ser programada em diferentes linguagens (PHP, Python, Java, etc), no entanto, o desenvolvimento de qualquer solução exige pessoal treinado, tempo de aprendizado, ambiente de testes, etc.

A Divisão de Indicadores e Dados do Departamento de Economia e Estatística/SPGG em parceria com a Subchefia de Ética, Controle Público e Transparência da Casa Civil desenvolveu uma ferramenta para facilitar que equipes de TI executem a automatização da publicação de dados e metadados na plataforma CKAN.

Partindo do pressuposto que o formato JSON é nativo da plataforma CKAN, surgiu a proposta de criação de uma ferramenta genérica que execute a publicação de metadados na plataforma CKAN a partir de arquivos .JSON.



A instituição participante realiza a geração de seus dados, preferencialmente em formato aberto. Os dados podem ficar armazenados em um servidor da *web* ou dentro do próprio servidor CKAN (*DataStore*). Para que esses dados se tornem disponíveis no catálogo de dados abertos do Rio Grande do Sul (dados.rs.gov.br) é necessário que sejam gerados arquivos JSON onde estão descritos os metadados que identificam e qualificam os dados.

Para gerar o arquivo JSON devem ser seguidos os parâmetros descritos na documentação do CKAN. Alguns parâmetros são obrigatórios, sem eles o *dataset* não pode ser

criado e outros são opcionais. Para agregar qualidade aos dados, é desejável que a maior quantidade de informações possível seja fornecida no momento da criação do *dataset*. Para isso, junto a alguns parâmetros além das marcações “obrigatório” e “opcional”, foi colocada a marcação “importante” para ajudar no momento da escolha dos metadados.

4.1 Parâmetros possíveis para um *dataset*

https://docs.ckan.org/en/2.9/api/#ckan.logic.action.create.package_create

name (string) – o nome do *dataset*, deve ser único, deve ter entre 2 e 100 caracteres (obrigatório). Deve conter somente caracteres alfanuméricos minúsculos. Pode conter os símbolos - e _.

owner_org (string) – id da organização proprietária do *dataset* (obrigatório). Listagem de organizações disponível em https://dados.rs.gov.br/api/3/action/organization_list

title (string) – título do *dataset* (opcional) (importante)

author (string) – nome do autor do *dataset* (opcional)

author_email (string) – email do autor do *dataset* (opcional)

maintainer (string) – nome do mantenedor do *dataset* (opcional) (importante)

maintainer_email (string) – email do mantenedor do *dataset* (opcional) (importante)

license_id (license id string) – id da licença do *dataset* (opcional) Listagem de licenças acessível em https://dados.rs.gov.br/api/3/action/license_list

notes (string) – uma descrição do *dataset* (opcional) (importante). É possível formatar o texto com a sintaxe *Markdown* (veja seção 4.3)

url (string) – URL da fonte do *dataset* (opcional)

version (string, menor que 100 caracteres) – (opcional)

state (string) – o estado atual do *dataset* (opcional). Valores possíveis: 'active', 'deleted' (valor default: 'active')

resources (lista de resources) – esse é o *dataset resource* (opcional) (importante). Veja seção 4.2.

tags (lista de tags) – listagem de *tags* (opcional) (importante)

extras (lista de atributos extras) – informações extras (opcional), possuem formato 'key' : ' string', 'value' : 'string'

groups (list de grupos) – os grupos aos quais o metadado pertence (opcional) (importante). Possui formato 'name' : 'string'. Listagem de grupos acessível em https://dados.rs.gov.br/api/3/ action/group_list

4.2 Parâmetros possíveis para um *resource*

https://docs.ckan.org/en/2.9/api/#ckan.logic.action.create.resource_create

name (string) – nome do *resource* (obrigatório)

url (string) – url da *resource* (obrigatório)

format (string) – (opcional) (importante)

description (string) – (opcional) (importante)

upload (string) – indica o caminho relativo do arquivo para *upload* no *datastore* (opcional)

hash (string) – (opcional)

resource_type (string) – (opcional)

mimetype (string) – (opcional)

cache_url (string) – (opcional)

size (int) – (opcional)

created (iso date string) – (opcional)

last_modified (iso date string) – (opcional)

cache_last_updated (iso date string) – (opcional)

O *resource* é a estrutura que contém todas as informações sobre o arquivo de dados. Existem duas formas de disponibilizar os arquivos de dados no servidor CKAN:

1. Através do preenchimento do campo **url**, indicando o endereço onde o dado está hospedado. Por questões de compatibilidade com o sistema do dados.rs.gov.br é importante que o servidor de hospedagem seja **HTTPS**.

2. Através do preenchimento do campo **upload**, indicando o caminho relativo (localização) do arquivo de dados. Neste caso o campo **url** continua existindo no *resource,* mas deve ser deixado em branco. Assim, se o arquivo de dados está na mesma pasta do metadado basta indicar o nome do arquivo. Caso esteja em uma pasta diferente deve-se usar o seguinte formato: **nome-da-pasta/nome-arquivo.csv**. Note que, como o sistema roda em Linux, deve-se usar a barra de diretório simples “/”.

Importante:

- O atributo **name** do *resource* não deve ser alterado após a criação no CKAN, pois ele é importante para identificação do *resource* durante o processo de atualização. A alteração do atributo **name** implicará na criação de um novo *resource* dentro do *dataset*.

- A simples exclusão de um *resource* no arquivo do dataset (.json) não implica na exclusão desse *resource* no CKAN. Caso seja necessário excluir um *resource* de um *dataset* deve-se fazer *login* no portal dados.rs.gov.br e executar a exclusão manualmente.

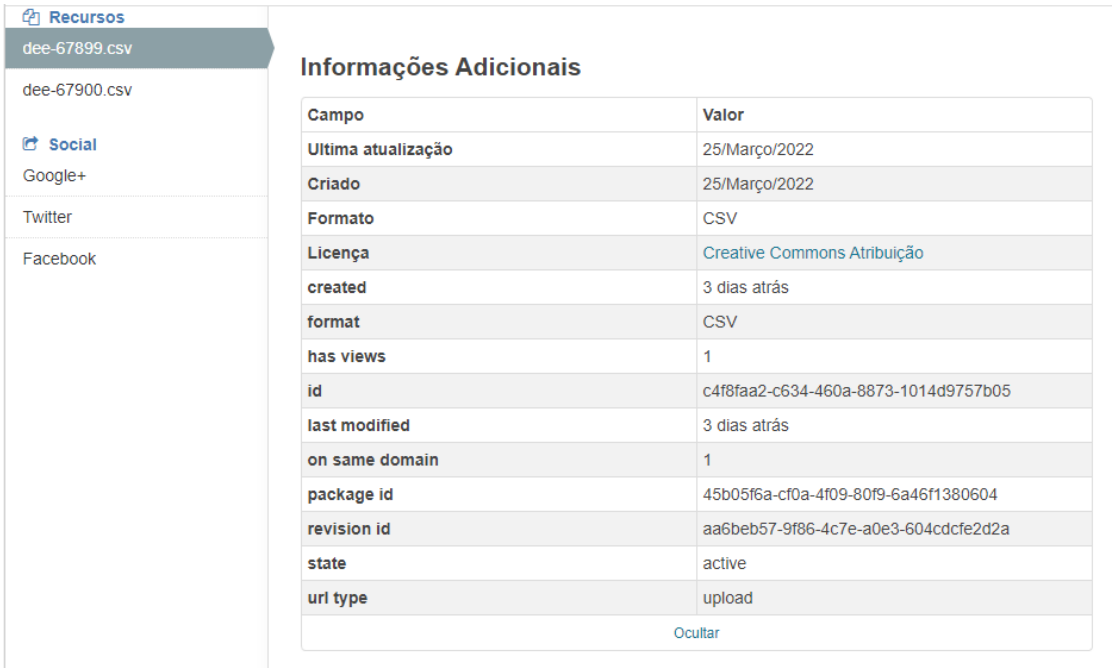
Exemplo

Dataset no formato JSON gerado para o DEE:


```
{
  "name": "dee-846",
  "owner_org": "spgg",
  "title": "Energia Elétrica - Consumo - Setor Público",
  "maintainer": "SPGG - DEE - Divisão de Dados e Indicadores",
  "maintainer_email": "biblioteca@planejamento.rs.gov.br",
  "notes": "Consumo anual de energia elétrica do setor público",
  "state": "active",
  "resources": [ {
      "format": "CSV",
      "name": "dee-67899.csv",
      "url": " https://dados.dee.planejamento.rs.gov.br/download/dee-67899.csv"
    } , {
      "format": "CSV",
      "name": "dee-67900.csv",
      "url": "",
      "upload": "arquivos/dee-67900.csv"
    }
],
  "groups": [ {"name": "infraestrutura" }  ],
  "tags": [   {  "name": "Consumo“  },    {   "name": "Setor Público"  } ],
  "extras": [  {
      "key": "Fonte dos Dados(1)",
      "value": "Distribuidoras de Energia Elétrica do Rio Grande do Sul."
    }  ]
}
```

O *dataset* do exemplo anterior carregado na interface do CKAN:

Informações do *resource* incluído no *dataset* exibidas pela interface do CKAN:



4.3 Sintaxe *Markdown* para formatação de textos no dataset

No campo **notes** do *dataset*, por questões de segurança do CKAN, a formatação HTML é ignorada. É possível utilizar a sintaxe *Markdown* para adicionar ênfase em algum conteúdo, fazer quebras de linhas, formatar cabeçalhos de seção, adicionar links ou imagens:

Negrito: adicione dois asteriscos **texto** ou dois traços-baixos __texto__ no início e no fim do conteúdo.

Itálico: adicione apenas um asterisco *texto* ou um traço-baixo _texto_ no início e no fim do conteúdo.

Links

[Texto do Link](url)

```
[Markdown Guide](https://www.markdownguide.org)
```

Adicionar imagens

![Descrição da imagem](url da imagem)

```
![Gaúchos](https://www.agenciapreview.com/wp-content/uploads/2017/09/Gauchos-ao-por-do-sol-0020.jpg)
```

Imagens com links

[![Descrição da imagem](url da imagem)](url destino)

```
[![Gaúchos](https://www.agenciapreview.com/wp-
content/uploads/2017/09/Gauchos-ao-por-do-sol-0020.jpg)](
https://www.agenciapreview.com/banco-de-imagens-do-rs/)
```

Quebra de linha:

Adicione em qualquer posição do texto a seguinte sequência de caracteres:

```
\r\n
```

Cabeçalhos

Adicione o(s) símbolo(s) # no texto conforme descrito a seguir (adicione uma quebra de linha após o final do texto do cabeçalho):

```
# Nível 1
## Nível 2
### Nível 3
#### Nível 4
##### Nível 5
###### Nível 6
```

5. Recomendações para criação do arquivo de dados

- Para armazenar os dados, utilize, de preferência, um formato de arquivo aberto, como o CSV, com os campos delimitados por aspas duplas e separados por vírgula, no formato “xxxxxx”,”xxxxxxxxxxx”,”xxxxxxxxxxxxxxx”

- Para salvar os dados utilize a codificação UTF-8

- O servidor onde os dados serão armazenados deve ser HTTPS

6 Recomendações para criação do arquivo de metadados

- Arquivo onde serão armazenados os metadados deve ser obrigatoriamente no formato JSON, codificado em UTF-8.

- Atributo “name” é o identificador do *dataset,* portanto deve ser único na organização. O atributo “name” não deve ser alterado no arquivo JSON após a publicação, pois essa alteração implicará na criação de um novo *dataset* em uma próxima execução do sistema. A mesma restrição serve para atributo “name” do *resource* (veja seção 4.2).

- Atributo “url” da lista de *resources* deve ser único na organização. Para o caso de armazenamento do dado no *DataStore* do CKAN o atributo deve ser deixado em branco

- É recomendável que nomes dos arquivos ou URLs armazenados nos *resources* não sejam alterados ao longo do tempo. As mudanças de endereços podem afetar sistemas que utilizam os dados armazenados como fonte

- Preferencialmente agrupar dados de uma série temporal de uma mesma “variável” em um arquivo único (*resource*). Evite que o arquivo especificado no *resource* tenha mais que 1000 registros (linhas), pois o *explorer*/*viewer* do CKAN limita a visualização em 1000 registros..

- A lista de *tags* não pode conter símbolos, nem caracteres especiais.

Sugestão: No momento da criação dos primeiros *datasets* é interessante fazer a validação dos arquivos JSON para verificar se estão corretamente formatados. É possível fazer a validação em sites gratuitos como https://jsonformatter.curiousconcept.com e https://jsonlint.com.

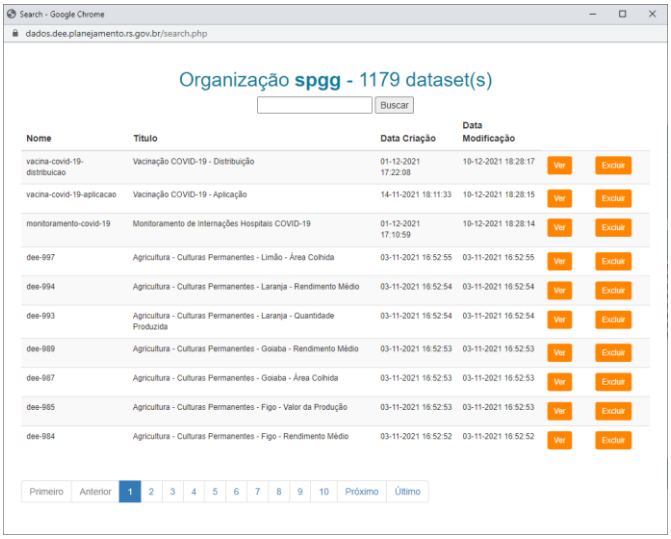
7 Sistema para publicação de dados e metadados no dados.rs.gov.br

O sistema, que está disponível em: https://dados.dee.planejamento.rs.gov.br/, foi construído para simplificar a publicação no portal dados.rs.gov.br. Será necessário um login para acessá-lo, solicite para josue-sperb@planejamento.rs.gov.br.

Na figura abaixo descrevemos as funcionalidades da tela inicial do sistema.

O Menu Principal dá acesso a cinco funcionalidades do sistema:

1. Gerenciar Medatados: Disponibiliza busca, navegação em todos os *datasets* da organização. Permite que o usuário execute duas ações: ver detalhes do *dataset* e excluir o *dataset* do servidor. A exclusão de *resources* não está prevista, deve ser feita manualmente no portal do CKAN.



2. Manual de utilização: Oferece acesso a este manual.
3. Configurações: Permite configurar duas informações importantes para o funcionamento do sistema. O nome da organização e a chave que permite a publicação no CKAN. O nome da organização pode ser obtido acessando a listagem disponibilizada no link https://dados.rs.gov.br/api/3/action/organization_list. Para obter a chave veja a seção 7.1. Importante salvar esses dados antes de iniciar o processo de publicação.

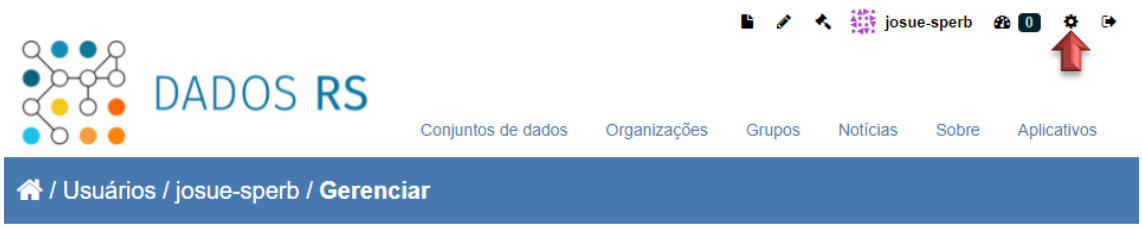
4. Logs de Execução: Permite exibir os logs de execução anteriores.
5. Alterar a senha: Permite que seja alterada a senha de acesso ao sistema.

6. Sair: Efetua *logoff* do sistema.

7.1 Como obter a chave de acesso para publicação automatizada

Caso a instituição ainda não tenha um *login* no dados.rs.gov.br deve fazer uma solicitação para a equipe da PROCERGS.

Faça o *login* no site dados.rs.gov.br e acesse a opção “Editar configurações” (imagem a seguir):



Na página de configurações do CKAN (imagem a seguir), no canto inferior direito existe o botão “Gerar Nova Chave para API”. A chave gerada deve ser salva no sistema disponível em https://dados.dee.planejamento.rs.gov.br/, escolhendo o item Configurações do Menu Principal. Essa chave deve ser gerada apenas uma vez.

7.2 Processo de publicação

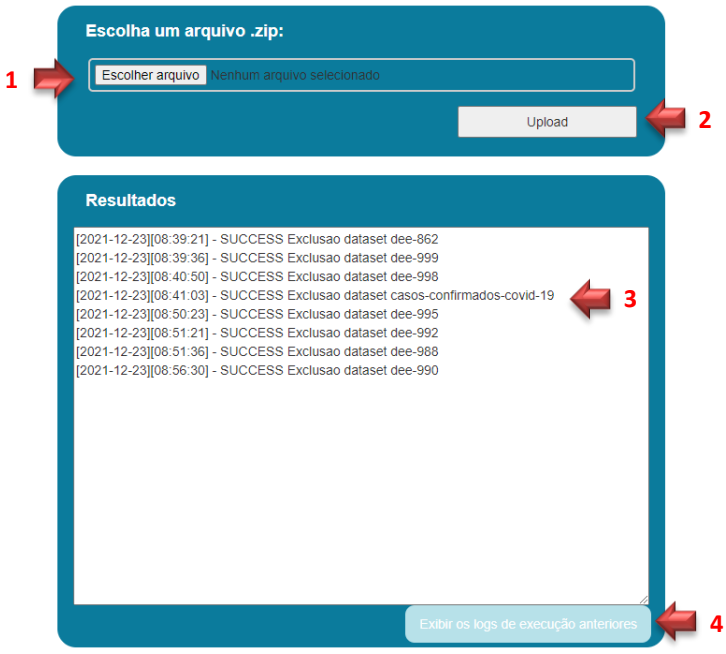
Com os arquivos de dados e metadados (.json) prontos, podemos iniciar o processo de publicação no portal do dados.rs.gov.br. Antes de iniciar o processo é necessário compactar os arquivos de metadados (.json) e dados (caso existam) em um arquivo no formato zip. Importante: os arquivos .json deve ficar alocados na raiz do arquivo .zip. Os arquivos de dados podem estar na raiz ou em uma pasta específica desde que indicado corretamente no campo **upload** do *resource* (veja seção 4.2). Primeiros passos:

- Faça o upload do arquivo .zip, para isso basta clicar no botão “Escolher arquivo” (1) e selecionar o arquivo .zip que contém os dados (opcional) e metadados.

- Clique no botão Upload (2). O sistema irá informar se o arquivo .zip foi carregado corretamente e descompactado.

- Aguarde alguns segundos para que o servidor faça a leitura e o processamento dos arquivos. O tempo máximo para iniciar o processamento é de 1 minuto.

- Visualize o resultado da execução na caixa Resultados (3). O sistema carregará o arquivo de log onde será mostrado se ouve sucesso ou erro no carregamento dos *datasets* e *resources*. Caso deseje visualizar os logs de execução anteriores clique em “Exibir todos os logs” (4).



Observações: A utilização do sistema com interface web é indicada para casos onde existe uma grande quantidade de dados/metadados e a atualização não é muito frequente (mensal ou anual). Para casos onde são necessárias atualizações a cada hora, diárias ou semanais recomendamos uma automatização total do processo de publicação, sem interação humana. O sistema foi projetado com essa característica, para utilizar esse recurso é necessária a criação de uma conta de FTP para transferência de arquivos. Solicite ao administrador do sistema pelo email josue-sperb@planejamento.rs.gov.br.

Referências

https://pt.wikipedia.org/wiki/JSON

http://ckan.org/

http://docs.ckan.org/en/2.9/

https://dados.gov.br/about

https://docs.ckan.org/en/538-package-install-docs/publishing-datasets.html

https://www.markdownguide.org/basic-syntax/

Dúvidas e Sugestões

Dúvidas e sugestões para melhoras no sistema e neste manual podem ser enviadas para josue-sperb@planejamento.rs.gov.br

Colaboradores: Josué Klafke Sperb (DEE/SPGG), José Henrique Schwanck Hinkel (DGTI/SES), Bruno Paim (DEE/SPGG), Fernando Ioannides Lopes da Cruz (DEE/SPGG).