# OpenData Processor: An Automation tool for the process of extracting and publishing open data to CKAN

Allyson Vilela and André Almeida
Federal Institute of Education, Science and Technology of Rio Grande do Norte
Natal, Rio Grande do Norte, Brazil
{allyson.barros,andre.almeida}@ifrn.edu.br

Frederico Lopes
Metropole Digital Institute
Federal University of Rio Grande do Norte
Natal, Rio Grande do Norte, Brazil
fred@imd.ufrn.br

## ABSTRACT

Public access to government information is an important aspect of modern society that allows an active participation of the population in monitoring government actions. Decree No. 8.777, signed on May 11, 2016, establishes the Open Data Policy of the Brazilian Federal Government. From this, the entities of the federal public administration, autarchic and foundational are obliged to make data available in open format. However, many of these institutions are failing to meet the commitments set out in the Decree. One possible explanation for this low number is the need for the technical team to have a good knowledge of their information systems and current legislation, allied to the difficulty of extracting the data, since in most institutions the whole process of data extraction, processing and publication of open data is done manually. In this sense, this work presents the OpenData Processor, an automation tool for the process of extracting, publishing and updating open data that brings agility in the publication and periodical updating, saving time and facilitating the management of open data portals.

## KEYWORDS

open data. extraction tool. automated process. CKAN.

## 1 INTRODUÇÃO

O acesso à informação é um importante aspecto da sociedade moderna, uma vez que permite uma participação ativa da população na fiscalização das ações governamentais e contribui com a melhoria da gestão pública e no combate à corrupção [7]. Para o Tribunal de Contas da União, a abertura dos dados na Administração Pública contribui para uma melhor transparência na gestão pública a partir da contribuição da sociedade através da criação de soluções inovadoras para fiscalização desses dados [7]. Nesse sentido, o Decreto nº 8777 [1], assinado pela Presidência da República em 11 de Maio de 2016, institui a Política de Dados Abertos do Poder Executivo Federal.

De acordo com o *Open Knowledge International* [5], uma organização global sem fins lucrativos focada em difundir a importância do acesso e uso dos dados abertos como na sociedade civil, "dados são abertos quando qualquer pessoa pode livremente acessá-los, utilizá-los, modificá-los e compartilhá-los para qualquer finalidade, estando sujeito a, no máximo, a exigências que visem preservar sua proveniência e sua abertura"[7]. Seguindo essa linha, dados abertos governamentais são aqueles gerados pelo governo e colocados à disposição da sociedade com o objetivo não disponibilizar apenas a sua leitura como também o seu acompanhamento, reutilização em sites e aplicações externas e o cruzamento com outros dados de órgãos/esferas diferentes [4].

Diante disso, os órgãos e entidades da administração pública federal direta, autárquica e fundacional são obrigados a publicarem os seus dados de forma aberta, segundo as diretrizes definidas na Infraestrutura Nacional de Dados Abertos (INDA) do Ministério do Planejamento, Orçamento e Gestão (MPOG) [3].

Este artigo tem por objetivo apresentar as principais características do OpenData Processor, uma aplicação web que automatiza todo o processo de extração, tratamento e catalogação de dados abertos seguindo os padrões definidos pela Infraestrutura Nacional de Dados Abertos (INDA) do Ministério do Planejamento, Desenvolvimento e Gestão (MPOG) e que pode ser utilizada por instituições públicas dos governos federal, estaduais e municipais. Essa versão do OpenData Processor utiliza o CKAN como plataforma de dados abertos. Tal plataforma é mais utilizada no mundo para esse fim. De código livre e desenvolvida em Python, a CKAN é utilizada no Portal Brasileiro de Dados Abertos e por governos e organizações públicas de diversos países [6].

## 2 PROBLEMÁTICA E MOTIVAÇÃO

O Decreto nº 8777, assinado pela Presidência da República em 11 de Maio de 2016, institui a Política de Dados Abertos do Poder Executivo federal. A partir dele, os órgãos e entidades da administração pública federal direta, autárquica e fundacional são obrigados a disponibilizarem os dados, que não estejam sob sigilo ou sob restrição de acesso, contidos em seus banco de dados sob a forma de dados abertos possibilitando assim um aprimoramento da cultura de transparência pública para a sociedade e facilitar a troca de informações entre órgãos e entidades da administração pública federal e as diferentes esferas da federação [1].

No entanto, apesar dessa obrigação, grande parte das instituições do poder executivo federal estão descumprindo os compromissos definidos no Decreto [2]. Uma possível explicação para esse número é a dificuldade enfrentada durante o processo de abertura dos dados devido a falta de integração entre os sistemas de informação utilizados nas instituições e as plataformas de dados abertos.

Além disso, em muitas instituições esse processo é feito de forma manual e exige que a equipe responsável pela abertura dos dados possuam bons conhecimentos de seus sistemas de informação e das legislações e recomendações vigentes.

In: XVII Workshop de Ferramentas e Aplicações (WFA 2018), Salvador, Brasil. Anais do XXIV Simpósio Brasileiro de Sistemas Multimídia e Web: Workshops e Pôsteres. Porto Alegre: Sociedade Brasileira de Computação, 2018.

ISBN 978-85-7669-435-9.

**Figura 1: Captura de tela do painel administrativo do OpenData Processor.**

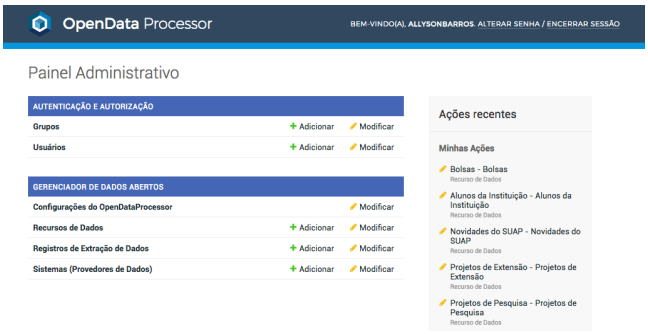



**Figura 2: Captura de tela da pré-visualização dos dados à serem extraídos.**

Amostra dos Dados Extraídos

| status | valor_liquido_contabil | campus | valor_inicial | estado_conservacao | codigo | id | descricao |
|---|---|---|---|---|---|---|---|
| ativo | 0,01 | CNAT | 0,01 | Não aplicar | 2 | 18070 | POLTRONA GIRATORIA MOD. C-1, C/ ASSENTO E ENCOSTODE MADEIRA. |
| baixado | 0,01 | CNAT | 0,01 | Não aplicar | 8 | 23827 | FICHARIO DE MADEIRA, C/11 GAVETAS, MED.0,84X0,53X0,28M. |
| baixado | 0,01 | CNAT | 0,01 | Não aplicar | 21 | 18071 | MESA, MOD.MM-2, C/01 GAVETA, MED.0,55X0,45X0,70M,C/RODIZIOS. |
| baixado | 0,01 | CNAT | 0,01 | Não aplicar | 22 | 16494 | FICHARIO DE ACO, MARCA "CONFIANCA", C/02 GAVETAS, P/FICHAS DE 5X8, MOD. 0258, DE COR CINZA. |
| baixado | 0,01 | CNAT | 0,09 | Não aplicar | 40 | 121 | ESTANTE DE ACO, C/06 PRATELEIRAS, MED. 0, 92X0, 30X1, 98M, DE COR CINZA. |
| baixado | 0,01 | CNAT | 0,09 | Não aplicar | 45 | 122 | ESTANTE DE ACO, C/06 PRATELEIRAS, MED. 0, 92X0, 30X1, 98M, DE COR CINZA. |
| baixado | 0,01 | CNAT | 0,09 | Não aplicar | 46 | 123 | ESTANTE DE ACO, C/06 PRATELEIRAS, MED. 0, 92X0, 30X1, 98M, DE COR CINZA. |

Nesse sentido, o desenvolvimento de uma ferramenta de automatização do processo de extração, publicação e atualização dos dados traz uma maior agilidade na publicação e atualização periódica dos dados abertos, economia de tempo e facilidade na gestão dos portais de dados abertos das instituições públicas.

## 3 FUNCIONALIDADES

O objetivo do OpenData Processor é automatizar todo o processo de extração, publicação e atualização de dados abertos das instituições públicas seguindo os padrões definidos pela Infraestrutura Nacional de Dados Abertos (INDA) do Ministério do Planejamento, Desenvolvimento e Gestão (MPDG).

A ferramenta possui diversas funcionalidades que visam facilitar o gerenciamento dos dados abertos das instituições e tornar a publicação dos dados abertos mais fácil por parte da equipe de TI. A seguir serão detalhadas algumas dessas funcionalidades:

- **Painel Administrativo**: Através da interface administrativa é possível realizar o cadastro, edição ou exclusão de organizações, grupos, conjuntos de dados e executar as tarefas de extração e sincronização de dados. A Figura 1 apresenta a interface administrativa do OpenData Processor.
- **Sincronização de dados com o CKAN**: O OpenData Processor permite a integração com instâncias já existentes do CKAN. Com isso, após cadastrar ou atualizar as configurações relativas à integração com o CKAN, é realizada de forma automática o cadastro/atualização das Organizações, Grupos e Conjuntos de Dados existentes na instância CKAN da Instituição.
- **Suporte à múltiplos Provedores de Dados**: O OpenData Processor permite a extração de dados a partir de múltiplos provedores de dados diferentes: sistemas de informação, banco de dados e planilhas eletrônicas.
- **Extração de Dados via API REST**: Permite a extração de dados a partir de uma API REST já existente que utilize um dos seguintes mecanismos de autenticação: Nenhuma, HTTP Basic, Token Simples, OAUTH ou JSON Web Token (JWT).
- **Extração de Dados via Banco de Dados**: Permite a extração de dados a partir dos SGDBs mais utilizados no mercado como SQLite3, MySQL / MariaDB, PostgreSQL e SQL Server e também a partir de bancos de dados NoSQL como Redis, Memcached, MongoDB, Riak e CouchDB.
- **Pré-visualização dos dados**: Permite a pré-visualização dos dados à serem extraídos via API ou Banco de Dados. Dessa forma, é possível validar se as configurações relativas à extração foram realizadas corretamente. A Figura 2 ilustra a configuração do agendamento de extração automática.
- **Dicionário de Dados**: Permite a criação de uma página de especificação, denominada dicionário de dados, que contém o nome, tipo, formato e descrição de cada campo do conjunto de dados.
- **Agendamento de Extração Automática de Dados**: Ao cadastrar ou editar um Conjunto de dados é possível indicar se a extração automática será executada diariamente, semanalmente, mensalmente, semestralmente ou anualmente e em qual horário ela será executada. Ao salvar as alterações, o agendamento da extração é adicionado à fila de execução em segundo-plano. A Figura 3 ilustra a configuração do agendamento de extração automática.
- **Auditoria**: A ferramenta permite que os administradores possam visualizar as ações de cadastro, edição e exclusão realizadas pelos usuários do sistema a partir da interface administrativa.

**Figura 3: Captura de tela da configuração do agendamento de extração automática no OpenData Processor.**

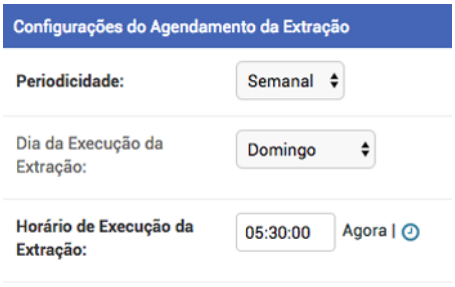
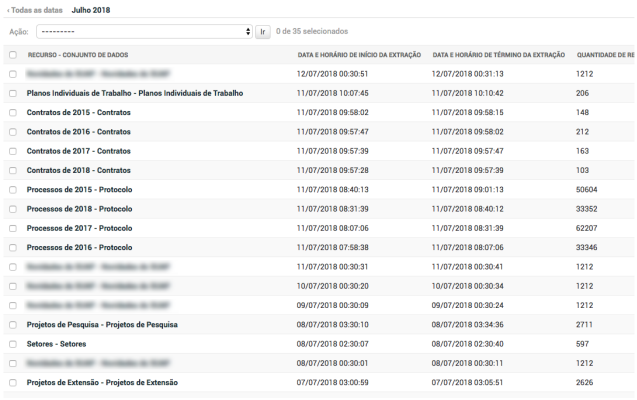


**Figura 4: Captura de tela de acompanhamento das extrações realizadas pelo OpenData Processor.**




- **Acompanhamento de Extrações de Dados Realizadas**: Ao executar uma extração de dados, independentemente do resultado, o OpenData Processor registra as informações relativas à execução tais como se a execução foi bem sucedida ou não, a quantidade de registros que foram extraídos e o horário de início e término da extração. A Figura 4 ilustra a funcionalidade de acompanhamento das extrações realizadas.

## 4 ARQUITETURA E DEPLOY

O OpenData Processor é uma solução web escalável que foi desenvolvida utilizando as seguintes tecnologias: Python, Django, PostgreSQL, Redis, Python-RQ, Pandas, Docker e Docker Compose.

A ferramenta se adapta aos diferentes tipos de infraestrutura de TI disponíveis nas instituições públicas, ou seja, a partir de três opções arquiteturais disponíveis é possível implantá-la na infraestrutura de TI da instituição, em um provedor de nuvem como AWS e Azure, ou utilizá-la como um serviço (SaaS).

Nas Seções 4.1, 4.2 e 4.3 serão apresentadas cada uma dessas opções arquiteturais e na Seção 4.4 serão apresentadas informações relativas ao *deploy* do OpenData Processor.

### 4.1 Standalone

Esta opção arquitetural permite que o OpenData Processor seja executado em um ambiente completamente isolado do ambiente de produção dos sistemas corporativos utilizados na instituição. Dessa forma, todo o processo de extração dos dados é realizado em um único servidor virtual, evitando que esse processo interfira no desempenho dos sistemas da instituição. A figura 5 ilustra a separação dos componentes do OpenData Processor na versão "Standalone".

Os componentes dessa arquitetura são: **Queue**, responsável pelo gerenciamento da fila de execução de tarefas em *background*; **Scheduler**, responsável pelo agendamento de tarefas na fila de execução em *background* e **Worker**, responsável pela execução das tarefas em *background*.

A comunicação entre o OpenData Processor, o CKAN e os sistemas de informação utilizados nas instituições é feita através de

**Figura 5: Visão geral dos componentes da arquitetura "Standalone".**

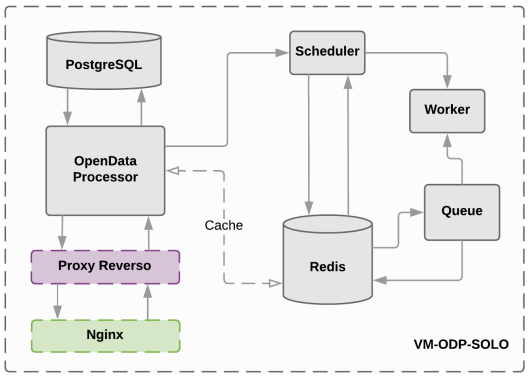
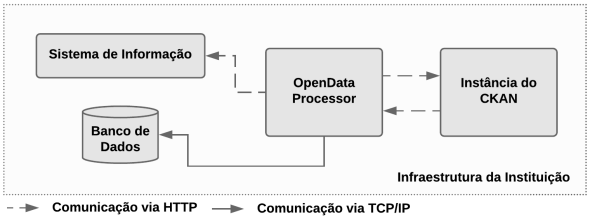



**Figura 6: Ilustração da forma de comunicação entre o OpenData Processor, os provedores de dados e o CKAN na versão Standalone.**




serviços web REST, já a comunicação com os bancos de dados são feita através de conexão TCP/IP. A Figura 6 ilustra a forma de comunicação entre o OpenData Processor, os provedores de dados (sistemas de informação ou banco de dados) e o CKAN na versão da arquitetura Standalone.

### 4.2 Distribuída

Esta opção de arquitetura é ideal para as instituições que possuem uma infraestrutura de TI mais robusta ou que utilizam provedores de serviços de hospedagem na nuvem como AWS, Azure, DigitalOcean e Scaleway uma vez que a arquitetura permite a separação de cada um de seus componentes em servidores virtuais separados ou através de tecnologias baseadas em contêineres Linux (*Linux Containers* - LXC) como Docker, por exemplo.

Ao utilizar essa opção arquitetural é possível distribuir o consumo de recursos computacionais, uma vez que ela permite a paralelização das rotinas de extração de dados através da criação de novas instâncias do componente Worker. A figura 7 ilustra a separação dos componentes da arquitetura em uma infraestrutura baseada em LXC.

A comunicação entre o OpenData Processor, o CKAN e os provedores de dados se dá da mesma forma que na versão "Standalone", no entanto, caso o OpenData Processor seja implantado em um provedor de Nuvem, é necessário a liberação de conexões externas aos servidores de banco de dados utilizados. A Figura 8 ilustra a

**Figura 7: Visão geral dos componentes da arquitetura "Distribuída"enfatizando a separação dos componentes em uma Infraestrutura Baseada em LXC.**

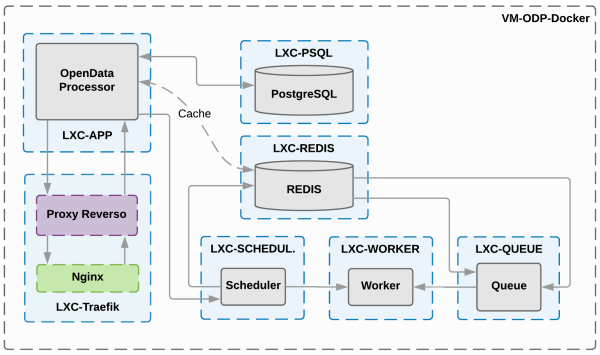
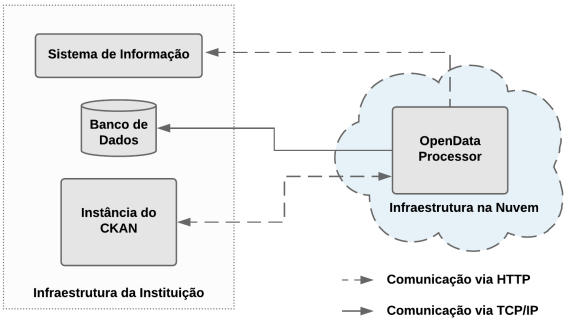



**Figura 8: Ilustração da forma de comunicação entre o OpenData Processor, os provedores de dados e o CKAN na versão "Distribuída"implantada em um provedor de Nuvem.**




forma de comunicação entre o OpenData Processor, os provedores de dados (sistemas de informação ou banco de dados) e o CKAN na versão da arquitetura Distribuída.

## 4.3 Software como Serviço (SaaS)

Esta opção arquitetural é ideal para as instituições que não possuem infraestrutura de TI e/ou equipe técnica para manter o OpenData Processor funcionando. Ou seja, por se tratar de um software como serviço, o fornecedor do software se responsabiliza por toda a estrutura necessária à disponibilização do sistema.

Nesta versão da arquitetura foi adicionado um novo componente, denominado "gerenciador de instâncias", o qual é responsável pela criação, exclusão, desativação e ativação das instâncias do OpenData Processor. Além disso, ele permite que as instituições solicitem a criação de uma instância de avaliação por um período de até 60 dias a partir do website[1] de divulgação da ferramenta.

Vale salientar que, as instâncias compartilham tanto a infraestrutura física de servidores quanto o banco de dados uma vez que

[1]http://www.opendataprocessor.com

**Figura 9: Visão geral dos componentes da arquitetura "Software como Serviço".**

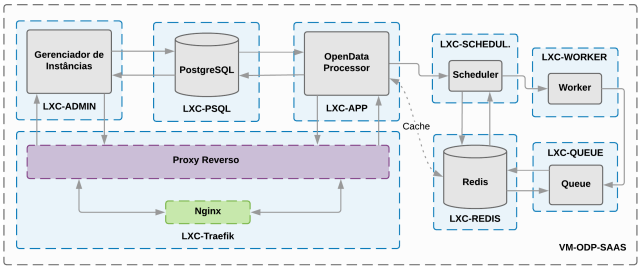
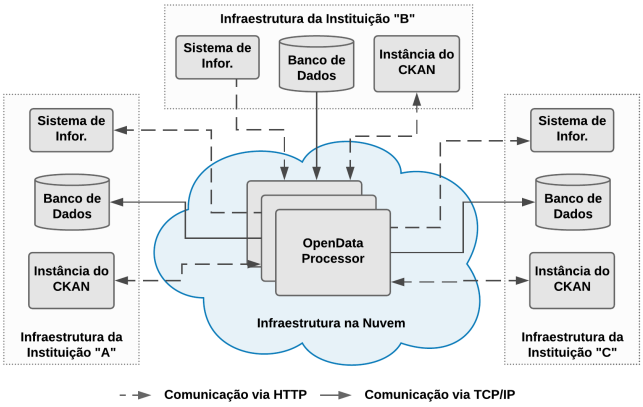



**Figura 10: Ilustração da forma de comunicação entre o OpenData Processor, os provedores de dados e o CKAN na versão SaaS.**




a separação dos dados é feita de forma lógica através de *schemas* do SGDB. Dessa forma, ao criarmos quatro novas instâncias serão criados quatro *schemas* no banco de dados, cada um referente à uma instância.

Da mesma forma que na versão "Distribuída", como a instância é executada fora da rede da instituição é necessário a liberação de conexões externas aos servidores de banco de dados utilizados como provedores de dados. A Figura 8 ilustra a forma de comunicação entre o OpenData Processor, os provedores de dados (sistemas de informação ou banco de dados) e o CKAN na versão da arquitetura Software como Serviço.

## 4.4 Deploy

Como visto anteriormente, as instituições podem optar por realizar a implantação do OpenData Processor em sua própria infraestrutura de TI, em um provedor de nuvem ou utilizá-lo como um serviço. Dessa forma, para implantar o OpenData Processor são necessárias as seguintes configurações mínimas, dependendo da opção arquitetural escolhida:

- **Standalone**: Um servidor virtual com 2 processadores, 2GB de Memória de RAM e 20GB de espaço em disco utilizando a distribuição Linux Debian ou Ubuntu Server. No

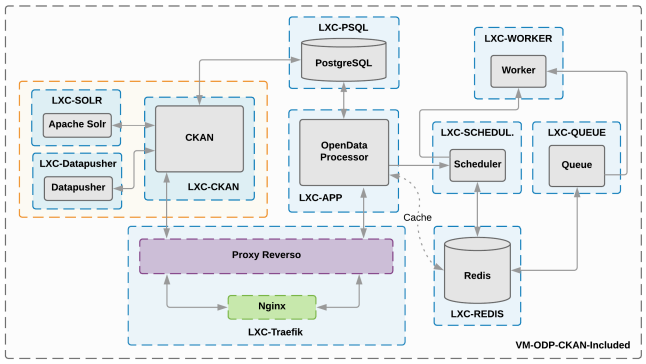



**Figura 11: Visão geral da arquitetura do OpenData Processor enfatizando a inclusão dos componentes da versão customizada do CKAN.**

entanto, caso a instituição deseja publicar conjuntos de dados contendo milhões de registros, será necessário a utilização de, no mínimo, 4GB de memória RAM.

- **Distribuída**: Pelo menos três servidores virtuais com 2GB de Memória de RAM e 20GB de espaço em disco utilizando a distribuição Linux Debian ou Ubuntu Server, sendo um para o banco de dados, outro para a execução das tarefas em *background* e outro para o servidor de aplicação. Ou, caso deseje executar os componentes em contêineres Linux, um servidor virtual com 4 processadores, 8GB de Memória de RAM e 40GB de espaço em disco utilizando a distribuição Linux Debian ou Ubuntu Server.
- **Software como Serviço**: Como a infraestrutura necessária para a implantação da instância é de responsabilidade do fornecedor do Software, a instituição não precisa se preocupar com as configurações mínimas.

Para facilitar a implantação do OpenData Processor nas instituições, para cada uma das versões arquiteturais são disponibilizados scripts que automatizam o processo de *deploy* através da utilização do Docker e Docker Compose. Dessa forma, todo o processo de instalação e configuração das dependências necessárias para o funcionamento do OpenData Processor é feito em poucos minutos sem a necessidade da intervenção humana.

Além disso, visando facilitar o processo de abertura dos dados em instituições públicas que ainda estão em fase de planejamento da abertura de seus dados, o OpenData Processor possibilita o *deploy* automático de uma instância do CKAN. Essa funcionalidade evita que a equipe de TI perca tempo com a instalação e configuração do CKAN. A Figura 11 ilustra a adição dos componentes do CKAN, destacados em amarelo, aos componentes da versão distribuída do OpenData Processor.

Os scripts de automatização do *deploy* do CKAN foram baseados em um projeto[2] de código aberto oferecido pela Plataforma FIWARE[3] e está disponível para utilização no endereço https://github.com/allysonbarros/docker-opendaprocessor-ckan. A Figura 12 ilustra o visual da instância customizada do CKAN.

[2]https://github.com/okfn/docker-fiware-ckan
[3]https://www.fiware.org/

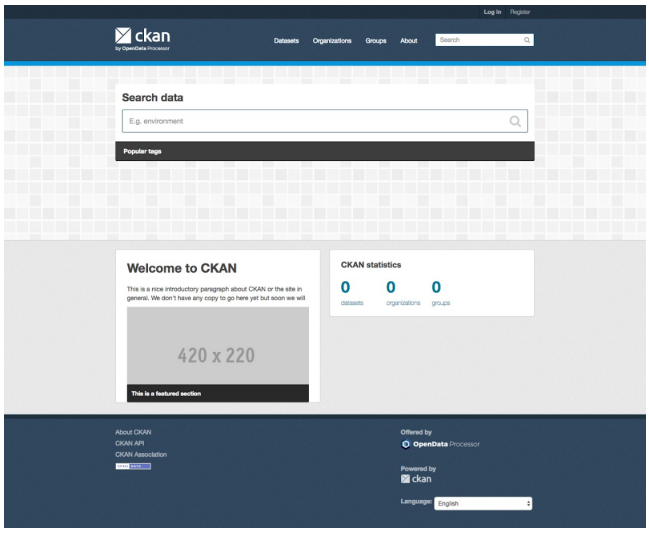



**Figura 12: Captura de tela da página inicial da instância customizada do CKAN oferecida pelo OpenData Processor.**

## 5 LICENÇA

O OpenData Processor é um software proprietário que foi registrado junto ao Instituto Nacional da Propriedade Industrial (INPI) e sua titularidade é compartilhada entre o IFRN e a UFRN.

Dessa forma, o processo de licenciamento para a utilização do OpenData Processor pelas instituições públicas se dá de duas formas:

- Através da assinatura de convênio de cooperação técnica para a realização da transferência de tecnologia entre a instituição e os titulares do software;
- Através da contratação de uma das empresas licenciadas pelos titulares para a realização da implantação, customização e suporte técnico do software.

## REFERÊNCIAS

[1] BRASIL. 2016. Decreto nº8777. (2016). http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2016/decreto/D8777.htm
[2] CONTROLADORIA-GERAL DA UNIÃO. 2018. Painel Monitoramento de Dados Abertos. (2018). http://paineis.cgu.gov.br/dadosabertos/index.htm
[3] MINISTÉRIO DO PLANEJAMENTO, DESENVOLVIMENTO E GESTÃO. 2017. Plano de Dados Abertos. (2017). http://wiki.dados.gov.br/Plano-de-Dados-Abertos.ashx
[4] NÚCLEO DE INFORMAÇÃO E COORDENAÇÃO DO PONTO BR. 2011. Manual dos dados abertos: Governos. (2011). http://www.w3c.br/pub/Materiais/PublicacoesW3C/Manual_Dados_Abertos_WEB.pdf
[5] OPEN KNOWLEDGE INTERNATIONAL. 2017. The Open Definition. (2017). http://opendefinition.org/
[6] OPEN KNOWLEDGE INTERNATIONAL. 2018. CKAN. (2018). http://ckan.org/about/
[7] TRIBUNAL DE CONTAS DA UNIÃO. 2016. Cinco Motivos para a Abertura de Dados na Administração Pública. (2016). http://portal.tcu.gov.br/lumis/portal/file/fileDownload.jsp?fileId=8A8182A24F0A728E014F0B366F2E2A40