Submission date: 06/10/2023
Resubmission date: 14/05/2024
Camera ready submission date: 21/10/2024

1st round notification: 12/01/2024
2nd round notification: 02/10/2024
Available online: 26/12/2024
Publication date: 26/12/2024

Section: Regular article

Implementation of Open Data Portals in Federal Higher Education Institutions
Implementação de Portais de Dados Abertos em Instituições Federais de Ensino Superior

Luiza Lima Meira de Menezes[1], **Maria da Conceição Moraes Batista**[1], **Roberta Macêdo Marques Gouveia**[1], **Taciana Pontual Falcão**[2], **Gabriel Alves de Albuquerque Júnior**[1]

[1]Departamento de Estatística e Informática – Universidade Federal Rural de Pernambuco (UFRPE)
Recife - Pernambuco - Brasil

[2]Departamento de Computação - Universidade Federal Rural de Pernambuco (UFRPE)
Recife - Pernambuco - Brasil

{menezes.luiza, maria.cmbatista, roberta.gouveia, taciana.pontual, gabriel.alves}@ufrpe.br

Abstract. *The Open Government Data initiative marked the beginning of a change process in public administration, with transparency as a way of social participation. One way to achieve this goal is an open data portal. However, government bodies face a number of challenges in implementing their open data initiatives. The main goal of this research is to chart the best technologies to develop and manage an open data portal for public education institutions and establish a procedure for extracting, manipulating and publishing such data. As proof of concept, a case study was carried out with the datasets of the Universidade Federal Rural de Pernambuco (UFRPE).*

Keywords. *Open Data; Open Government Data; Access to Information Law; CKAN.*

Resumo. *As iniciativas de Governo Aberto motivaram um processo de mudança na administração pública, visando a transparência como forma de participação social. Uma das formas de atingir esse objetivo consiste em um portal de dados abertos. No entanto, as instituições enfrentam uma série de desafios para por em prática suas estratégias de abertura de dados. A proposta desta pesquisa consiste em mapear as melhores tecnologias para desenvolver e gerenciar um portal de dados abertos em instituições de ensino público e estabelecer um procedimento para extração, manipulação e publicação dos dados. Como prova de conceito, foi realizado um estudo de caso com as bases de dados da Universidade Federal Rural de Pernambuco (UFRPE).*

Palavras-Chave. *Dados Abertos; Dados Abertos Governamentais; CKAN; Lei de Acesso à Informação.*

Cite as: Menezes, L.M.M., Batista, M.C.M., Gouveia, R.M.M. Falcão, T.P. & Alves, G. (2024). Implementação de Portais de Dados Abertos em Instituições Federais de Ensino Superior. iSys: Revista Brasileira de Sistemas de Informação (Brazilian Journal of Information Systems), Vol 17(1), 18:1-18:23. DOI: 10.5753/isys.2024.3732

1. Introdução

As políticas de transparência e abertura de dados, impulsionadas pela Parceria para Governo Aberto, ou *Open Government Data (OGD)*, incentivam os governos a disponibilizarem seus dados para a sociedade. O acesso a tais informações permite aos cidadãos entenderem o funcionamento e avaliarem a eficiência dos órgãos governamentais, contribuindo para o processo de tomada de decisão e possibilitando que situações de uso indevido de recursos sejam expostas e denunciadas [Ubaldi 2013].

No Brasil, o acesso aos dados governamentais, previsto na Constituição Federal de 1988, ganhou força em 2012, com a publicação da Lei de Acesso à Informação (LAI), a primeira de uma série de regulamentações federais para abertura de dados. A lei prevê que os órgãos e entidades públicos devem promover a disponibilização dos dados de interesse da população em portal na internet e responder às solicitações de informação.

Um dos meios de atender a esses requisitos é através de um portal de dados abertos, onde devem ser publicadas as bases de dados previstas para abertura no Plano de Dados Abertos (PDA) da instituição [Brasil 2019]. No entanto, muitos órgãos e instituições encontram dificuldades em atender a este requisito.

Segundo o Painel de Monitoramento de Dados Abertos[1], da Controladoria Geral da União (CGU), 57,8% deles estão sem PDA, 41,4% possuem PDA publicado e disponível, e 0,8% estão com PDA em construção. Dentre os órgãos com PDA publicado, 20% das bases previstas estão em atraso [Brasil 2022b].

Entre os desafios encontrados pelas instituições públicas ao abrir os seus dados, alguns autores listam o desconhecimento da legislação [Correa et al. 2014] e dificuldade de implementar os requisitos técnicos - como funcionalidades e requisitos de segurança - no portal [Lima et al. 2020]. Como resultado, a qualidade das informações divulgadas muitas vezes não possibilita sua utilização pela sociedade [Bachtiar et al. 2020]. As atividades necessárias para alcançar o sucesso no processo de abertura de dados e seu impacto ainda são pouco conhecidas. A falta de uma estratégia para abertura limita a qualidade das informações divulgadas e, consequentemente, seu potencial de reuso e geração de valor [Craveiro and Martano 2014].

Os benefícios dos dados abertos não estão estritamente na informação em si, mas em sua capacidade de contribuir para a sociedade. A prioridade dos órgãos não deve ser apenas tornar suas informações disponíveis, mas sim fornecer meios para que a população consiga encontrar, visualizar e analisar esses dados [Marijn Janssen and Zuiderwijk 2012]. Diante desta situação, a principal contribuição científica desta pesquisa consiste em realizar a implementação de um portal de dados abertos governamentais, estabelecendo quais etapas devem ser seguidas para garantir a publicação de dados de qualidade e elaborar um guia de boas práticas a ser seguido pelas instituições, com foco em Instituições Federais de Ensino Superior (IFES).

Ao longo da pesquisa, foi conduzido um estudo de caso com a Universidade Federal Rural de Pernambuco (UFRPE), resultando na implementação de seu portal de dados abertos e abordando as tecnologias disponíveis e as técnicas de extração e tratamento dos

[1]Disponível em: https://centralpaineis.cgu.gov.br/visualizar/dadosabertos. Acessado em: 26 fev. 2023.

dados para divulgação. O estudo de caso permitiu mapear fatores determinantes para sucesso para o processo.

Este artigo está organizado da seguinte forma: a seção 2 apresenta trabalhos relacionados; a seção 3 traz a fundamentação teórica, de forma a contextualizar o tema proposto; a seção 4 relata a metodologia de pesquisa e tecnologias adotadas no trabalho; a seção 5 relata as experiências adquiridas com o estudo de caso; a seção 6 propõe um guia de boas práticas para publicação de dados abertos e a seção 7 apresenta as conclusões e limitações da pesquisa e propõe trabalhos futuros.

2. Trabalhos Relacionados

Em artigo por Correa, Correa e Silva (2014), os autores analisam a aderência dos portais de transparência ativa de 20 municípios brasileiros aos princípios de OGD. Eles identificaram o baixo nível de maturidade dos governos locais em relação aos requisitos da LAI como a principal causa de insucesso de tais iniciativas.

Uma situação similar é identificada por Lima, Abdalla e Oliveira (2020), que avaliam a aderência das universidades federais brasileiras às determinações da LAI para transparência ativa e passiva. No caso dos portais de dados abertos, apenas 24 das 63 universidades analisadas atendiam aos requisitos. Os autores concluem que, mesmo com a LAI, ainda existe nessas instituições uma cultura de não divulgação das informações.

O artigo de Bachtiar, Suhardi e Muhamad (2020) estuda os principais desafios dos governos no processo de abertura de dados. Em relação ao portal em si, os principais problemas encontrados pelos autores foram a disponibilidade e interoperabilidade. Já Oliveira e Fonseca (2021) analisam os fatores por trás do sucesso das políticas de abertura de dados através de um estudo de caso com o Banco Central do Brasil (BC). Entre as descobertas, estão a importância da cultura de transparência e do apoio da alta administração.

O trabalho de Avila (2015) ressalta a importância do tratamento dos dados pessoais na publicação de dados abertos governamentais e da elaboração de documentação sobre a informação publicada, enquanto o artigo de Silva e Júnior (2018) destaca a necessidade de estabelecer processo de publicação de dados abertos em instituições públicas brasileiras. Já Dorobăt, e Posea (2021) avaliam a qualidade de dados abertos publicados pelos governos de diversos países, citando a adoção de boas práticas para a publicação como um critério de sucesso das iniciativas de transparência.

O artigo de Gao, Janssen e Zhang analisa como a publicação de dados abertos governamentais evoluiu ao longo da última década e conclui que a geração de valor a partir das informações divulgadas ainda é um desafio, assim como a sustentabilidade das estratégias de abertura e divulgação dos dados adotadas pelos governos.

Os trabalhos discutidos neste capítulo podem ser divididos em dois grupos. O primeiro relata desafios e dificuldades no processo de abertura de dados de instituições governamentais, enquanto o segundo traz casos de sucesso e orientações a serem seguidas. O principal diferencial desse trabalho consiste em propor uma forma de solucionar estes desafios, tendo como base as orientações de outros autores e a pesquisa realizada.

3. Fundamentação Teórica

Esta seção apresenta os conceitos necessários para a fundamentação teórica da pesquisa, incluindo a definição de Dados Abertos e as principais legislações que regem o tema no Brasil na esfera Federal, e justifica a necessidade de utilização de licenças abertas.

3.1. Dados Abertos Governamentais

O conhecimento pode ser considerado aberto quando qualquer pessoa está livre para acessá-lo, utilizá-lo, modificá-lo e compartilhá-lo, sujeito apenas às medidas que preservam a proveniência e abertura. Tal definição foi feita pela Open Knowledge Foundation (OKFN), organização que atua na promoção do conhecimento aberto e defende o direito de acessar, reutilizar e redistribuir informações [Molloy 2011].

Segundo Eaves (2009), os cidadãos precisam conseguir encontrar, utilizar e compartilhar um dado para que ele possa ser considerado aberto. Para o autor, a informação deve estar disponível na internet, em ferramentas de busca como o Google. Para ser utilizada, ela precisa ser publicada em formato não proprietário e legível por máquina. Por fim, para o dado ser compartilhado devem ser fornecidos mecanismos legais para isso, ou seja, a informação deve ser domínio público ou sob uma licença aberta.

A publicação de dados abertos governamentais é regida por oito princípios. Segundo tais diretrizes, as informações publicadas devem ser ser completas, primárias, atuais, acessíveis, processáveis por máquina, não discriminatórias, não proprietárias e publicadas sob licenças livres. Deste conceito surgiu a Parceria para Governo Aberto, iniciativa internacional fundada em 20 de setembro de 2011 por oito países, incluindo o Brasil, com o compromisso de incentivar globalmente práticas governamentais de transparência fiscal, acesso à informação, participação cidadã e divulgação de declarações patrimoniais por autoridades [Shintaku and Sales 2019].

Os dados governamentais são um recurso estratégico, com potencial de beneficiar diversos setores da sociedade. Esses setores incluem o próprio governo, melhorando a eficiência na alocação de recursos e reduzindo fraudes; e os cidadãos, que passam a ter meios de monitorar e responsabilizar o governo, com a possibilidade de identificar potenciais casos de corrupção e uso indevido ou ineficiente dos recursos públicos. As informações governamentais também são relevantes do ponto de vista econômico. Ao fornecer informações públicas sem custo, o governo promove a criação de produtos e serviços. Isso estimula a economia e gera recursos na forma de impostos [Ubaldi 2013].

3.2. Legislação Aplicada à Política de Dados Abertos

A publicação de Dados Abertos de IFES é regida por uma série de leis e decretos, entre os quais devem ser mencionados a LAI, a Lei Geral de Proteção de Dados (LGPD) e a Política de Dados Abertos do Poder Executivo Federal.

O acesso à informação é um direito fundamental, garantido no Brasil pela Constituição Federal de 1988 [Brasil 1988] e regulamentado pela LAI. Influenciada pelos princípios de dados abertos governamentais, a Lei regulamenta o direito de acesso dos cidadãos às informações públicas e se aplica para os poderes executivo, legislativo e judiciário da União, Estados, Distrito Federal e Municípios [Brasil 2011].

A LAI institui o acesso à informação pública como a regra, e o sigilo como exceção. É definido que a transparência dos dados deve ocorrer de forma ativa, que se dá quando a entidade disponibiliza por conta própria as informações, e passiva, como resposta aos pedidos de informação que podem ser feitos por qualquer cidadão [Brasil 2011].

Na esfera Federal, o acesso à informação é regido pela Política de Dados Abertos do Poder Executivo Federal, por meio da Infraestrutura Nacional de Dados Abertos (INDA). A CGU é responsável por monitorar a aplicação da política de Dados Abertos na esfera federal. O PDA é o principal instrumento esta política e organiza as ações de implementação e promoção da abertura de dados dos órgãos [Brasil 2016].

As iniciativas de acesso à informação devem estar de acordo com os princípios da LGPD [Brasil 2018]. Assim como o Regulamento Geral sobre Proteção de Dados (GDPR), a regulamentação da União Europeia para dados pessoais, a legislação brasileira tem como pilar a proteção, transparência no tratamento dos dados e direito de escolha a cerca de sua divulgação e publicidade [Lorenzon 2021].

Segundo a LGPD, é considerada informação pessoal aquela relacionada à pessoa natural identificada ou identificável. Já dados sobre origem racial ou étnica, convicção religiosa, opinião política, saúde ou à vida sexual, genético ou biométrico são classificados como sensíveis e possuem proteção especial [Brasil 2018].

3.3. Licenças Abertas

Para serem considerados dados abertos, não basta que os conjuntos de dados, ou *datasets*, estejam em formato aberto, eles também precisam estar legalmente abertos [Eaves 2009]. Em definição da (OKFN), *"A obra deve obrigatoriamente estar em domínio público ou ser fornecida sob uma licença aberta"*. Uma licença é considerada aberta quando permite irrevogavelmente o uso, redistribuição, modificação, separação e compilação para qualquer finalidade, sem necessidade de citar a fonte [Molloy 2011].

Muitos autores questionam se existe a necessidade de licenciar dados abertos governamentais. Contudo, a licença é uma ferramenta que torna explícito tudo que pode ser feito com determinado conjunto de dados. Portanto, o autor recomenda a utilização de licenças sem nenhuma restrição de uso em seu compartilhamento [Davies 2012].

Assim, o uso de uma licença aberta não só é necessário, mas essencial para o sucesso das iniciativas de dados abertos. O oferecimento de garantias legais que permitam o compartilhamento e reuso das bases de dados sem restrições encoraja o desenvolvimento de aplicações e ferramentas a partir dessas informações [Khayyat and Bannister 2014].

Por este motivo, o CGINDA determina que os dados pertencentes à União devem ser divulgados em uma das seguintes licenças abertas: Public Domain Dedication and Licen (PDDL), Open Data Commons Open Database Licen (ODbL), Creative Commons Zero (CC0) ou Creative Comons Atribuição 4 (CC4) [CGINDA 2017a].

4. Métodos e Ferramentas

Esta seção apresenta a metodologia de pesquisa e as tecnologias adotadas, ilustrada de forma sequencial na Figura 1. De forma a atender ao objetivo de estudar as tecnologias disponíveis para implantação de portais de dados abertos e elaborar um guia de boas

práticas, a primeira etapa desta pesquisa consistiu em pesquisa bibliográfica. São analisados trabalhos relevantes para a temática de dados abertos governamentais, de forma a conceitualizar o tema a ser estudado, além de identificar as práticas realizadas no Brasil e no mundo, o que possibilita a descoberta dos problemas e desafios comuns às instituições no processo de abertura de seus dados.

Figura 1. Método de pesquisa

Em seguida, é feita uma pesquisa documental com a legislação vigente acerca de dados abertos. Em relação à legislação, isto inclui a LAI, a Política de Dados Abertos do Governo Federal, uma vez que as IFES se enquadram na esfera federal, e a LGPD, assim como os manuais elaborados pela CGU.

Esta etapa também inclui o PDA da UFRPE, documento elaborado pelo Comitê de Transparência e Dados Abertos (CTDA) da instituição que contém as bases de dados previstas para abertura e seus respectivos responsáveis. Seguindo as recomendações da INDA, o PDA determina que todas as bases de dados precisam estar catalogadas também no Portal Brasileiro de Dados Abertos[2], preferencialmente de forma automatizada [PDA-UFRPE 2022].

A pesquisa bibliográfica e documental serve como base para a etapa seguinte, que consiste em identificar os requisitos técnicos que devem ser adotados pelos portais de dados abertos, incluindo as funcionalidades disponíveis para os usuários. Uma vez que tais requisitos sejam devidamente mapeados, é possível escolher a aplicação de gerenciamento de dados abertos que seria utilizada na implementação do portal, assim como as alterações que seriam necessárias para atendê-los, e colocar o portal no ar.

A proposta do estudo de caso feito com a UFRPE era de selecionar três bases de dados do PDA da instituição e fazer a publicação dos dados previstos na matriz de priorização. Essa publicação foi guiada pelos requisitos negociais mapeados na etapa de pesquisa. As bases escolhidas foram: Ensino Graduação, Contratos e Orçamento. A seleção foi feita com o objetivo de estudar três diferentes estratégias de publicação.

[2]Disponível em: https://dados.gov.br. Acessado em: 23 fev. 2023.

Os dados da base de Ensino de Graduação precisam ser extraídos do sistema acadêmico utilizado pela instituição. Os de Contratos vêm de sistemas do Governo Federal que já possuem suporte para exportação de relatório em formato aberto. O mesmo vale para as informações de Financeiro e Orçamento, sendo que neste caso o responsável já realizava um trabalho de tratamento e análise.

O estudo resulta na publicação das bases previstas, em processo que envolve os responsáveis por cada uma na seleção dos dados, elaboração do dicionário de dados e validação. As lições aprendidas ao longo do processo foram condensadas em um guia de boas práticas para publicação de dados abertos por IFES.

4.1. Atendimento às Normas e Regras

Segundo diretrizes da LAI e do Decreto nº 8777, as regras negociais a serem observadas nos portais de dados abertos são: garantir reuso irrestrito das bases; designar o responsável por sua atualização e manutenção; informar a periodicidade de atualização; fornecer descrição detalhada dos *datasets* em um dicionário de dados e utilizar linguagem de fácil compreensão [Brasil 2019]. Além de atender a esses critérios, as instituições devem se atentar ao tratamento de dados pessoais e dados pessoais sensíveis.

Em adição a essas regras, os PDAs das instituições podem trazer outras definições, detalhando alguns aspectos da implementação da LAI, que também devem ser seguidas. No caso da UFRPE, tais premissas incluem informar etiquetas ou palavras-chave, nome e e-mail do setor responsável pelos dados e periodicidade.

A LAI define em seu texto algumas informações que devem constar nos portais de transparência ativa. Um dos itens consiste em *"Dados gerais para o acompanhamento de programas, ações, projetos e obras de órgãos e entidades"* [Brasil 2011]. No caso das IFES, estas informações seriam relativas às atividades de ensino, que frequentemente possuem dados pessoais dos estudantes e docentes da instituição e precisam ser tratados seguindo as normas estabelecidas pela legislação.

Por determinação da LGPD, a divulgação de dados pessoais deve ser sempre acompanhada de justificativa, com a hipótese legal que dispensa o consentimento do titular. Conforme exposto na seção anterior, são considerados dados pessoais qualquer informação pertinente à pessoa natural identificada ou identificável, como nome, data de nascimento e endereço. No caso de portais de dados abertos, todo conjunto que inclua dados pessoais deve conter a hipótese legal que permite sua publicação.

Já os dados pessoais sensíveis - onde se enquadram informações biométricas e genéticas - devem ser removidos, caso permitam a identificação do titular, ou anonimizados, exceto nos casos indispensáveis. Nestas situações, a justificativa legal deve ser divulgada juntamente com os dados, de forma pública e acessível para o titular.

Outra recomendação consiste na utilização do Vocabulário Controlado do Governo Eletrônico (VCGE), um vocabulário para indexar informações - como bases de dados, documentos e sites - do governo de uma forma simples e compreensível. Seu objetivo é fornecer um guia de como catalogar essas informações e assim tornar esses dados mais fáceis de serem encontrados pelo público em ferramentas de indexação [Brasil 2022c].

Quanto à utilização de licenças abertas, foi escolhida a PDDL. A licença atende a todos os critérios para utilização em dados abertos e se destaca em relação às demais licenças recomendas pela CGINDA por ter em seu texto previsão para utilização em bases de dados, ao contrário da ODbL, por exemplo. Outro ponto positivo da PDDL está na simplicidade de utilização, que requer apenas uma menção à licença no portal, no conjunto de dados ou dicionário de dados, e um link para a página com sua descrição [OKFN 2022].

Tabela 1. Requisitos Negociais para Portais de Dados Abertos

Requisito	Legislação	Como atender
Informar detalhes do dataset	LAI, D8777 e PDA	Metadados
Atualizações periódicas	LAI, D8777 e PDA	Boas práticas
Instruções para contato	LAI e PDA	Campos do portal
Publicidade como regra	D8777	Boas práticas
Permissão irrestrita de reuso	D8777	Informar licença
Completude dos dados	D8777	Boas práticas
Informar responsável	D8777 e PDA	Campos do portal
Tratar dados pessoais	LGPD	Boas práticas
Tratar dados sensíveis	LGPD	Boas práticas
Informar etiquetas	PDA	Boas práticas
Informar assunto VCGE	PDA	Boas práticas

Para que o portal de dados abertos das IFES atenda às recomendações expostas nessa seção e listadas na Tabela 1, é necessário a adoção de uma série de boas práticas - ou metodologias de trabalho - por parte dos responsáveis.

4.2. Plataformas de Gerenciamento de Dados Abertos

Uma plataforma de gerenciamento de dados abertos, ou Open Data Platform (ODP), consiste em uma solução de software integrada composta por um portal, um gerenciador de metadados, biblioteca de Interface de Programação de Aplicação (API), serviços de busca e descoberta, ferramentas de visualização e relatórios. A ferramenta deve permitir publicação, compartilhamento e visualização de dados, servindo como ponto central onde o público pode descobrir, buscar, analisar e baixar informações [Osagie et al. 2015].

A aplicação fornece uma interface para que os responsáveis depositem conjuntos de dados diretamente ou através de uma API. Estes são armazenados e validados, podendo receber metadados, identificadores para controle de versão e localizadores, denominados tags ou etiquetas, que facilitam sua recuperação pelos usuários. Os editores podem adicionar e analisar os conjuntos de dados inseridos, enquanto os usuários podem pesquisar, visualizar e compartilhar as informações disponibilizadas [Costa et al. 2017].

4.3. Escolha do CKAN

A primeira etapa do processo de implementação do portal de dados abertos consiste na escolha da aplicação para gerenciamento de dados abertos. Com base no estudo da literatura e da legislação brasileira acerca dos requisitos a serem atendidos pelos portais, julgou-se que o CKAN era a melhor alternativa disponível no modelo *Open Source*.

O estudo da literatura mostrou que o CKAN é referência mundial para portais de dados abertos governamentais. Entre os governos que utilizam a aplicação, estão: Brasil (dados.gov.br), Estados Unidos (data.gov), Reino Unido (data.gov.uk), Canadá (canada.ca), Austrália (data.gov.au) e Indonésia (data.go.id) [Bachtiar et al. 2020].

Em artigo elaborado por Campelo e Neto (2020), que comparou quatro softwares gratuitos para criação de portais de dados abertos (CKAN, Dataverse, Invenio e Dspace), os autores demonstram que o CKAN é a plataforma mais eficiente para uso em portais de dados governamentais, o que justifica sua adesão em massa. Uma conclusão parecida foi obtida em pesquisa que comparou o CKAN com DKAN, Socrata e Semantic MediaWik. A aplicação foi considerada a com melhor usabilidade, além de atender nativamente ou com uso de *plugins* os principais critérios de *OGD* [Osagie et al. 2015].

Outro fator determinante para a adotação do CKAN foi o estudo dos requisitos técnicos, que podem ser resumidos em alguns pontos centrais. O primeiro deles consiste em fornecer acesso automatizado por sistemas externos em formatos abertos, estruturados e legíveis por máquina, e é atendido pelo CKAN através de seu *plugin DataStore*, que fornece uma API para leitura, busca e filtro dos conjuntos de dados presentes no portal.

Quanto aos formatos, as bases de dados devem ser publicadas exclusivamente em formatos não proprietários, suportados nativamente pelo CKAN. A próxima exigência diz respeito à permissão de gravação de relatórios em formatos abertos, já disponível na plataforma. Outro requisito envolve a disponibilização de ferramenta de pesquisa de conteúdo. A aplicação conta com a plataforma de busca integrada Solr. O campo de pesquisa é um dos componentes do cabeçalho do portal, sendo exibido em todas as páginas, e pode ser configurado para aparecer em destaque na página inicial [DOCS.CKAN 2022].

A legislação determina que o portal deve oferecer garantia de autenticidade e integridade das informações nele publicadas. No CKAN, apenas usuários autorizados podem criar, editar e excluir os conjuntos de dados publicados. A aplicação possui campos para que os dados do criador e mantedor dos *datasets* sejam informados.

Apesar de não ter acessibilidade como foco, a pesquisa mostra que o CKAN atende parcialmente os requisitos de acessibilidade. Para contornar esta limitação, foi utilizado o *VLibras Widget*[3], um recurso mantido pelo Governo Federal que possibilita tradução automática de sites para a Língua Brasileira de Sinais (LIBRAS).

Outro ponto extremamente relevante para a utilização do CKAN foi o fato de a aplicação possibilitar a sincronização automática com o Portal Brasileiro de Dados Abertos[4]. Lançado em dezembro de 2011, o portal do Governo Federal é uma ferramenta para

[3]Disponível em: https://vlibras.gov.br/doc/widget/index.html. Acesso em: 22 fev. 2023.
[4]Disponível em: https://dados.gov.br. Acessado em: 20 abr. 2023.

catalogar as bases de dados criadas por órgãos e entidades da administração pública e serve como um ponto central para a busca de dados públicos no Brasil [Brasil 2019].

4.4. Instalação e Configuração do CKAN

O CKAN é desenvolvido majoritariamente na linguagem de programação Python. A aplicação utiliza o sistema de template Jinja e a biblioteca de JavaScript JQuery. Feita para funcionar em navegador WEB, ela utiliza o Servidor HTTP Apache, como servidor, o Nginx como proxy e o PostgreSQL como banco de dados. Também fazem parte dos recursos utilizados as plataformas de pesquisa Solr, que roda no servidor Jetty e armazena seus dados no Redis, um servidor de armazenamento chave-valor [Costa et al. 2017].

Após a conclusão da instalação, são feitos ajustes nas configurações padrão da aplicação, que incluem personalização de idioma, imagens e títulos. Com o objetivo de melhorar a usabilidade do portal, foi desenvolvido um novo layout, destacando as seções de Pesquisa e Estatísticas, e configurado o *VLibras Widget*.

Além disso, o CKAN possui diversos *plugins* que podem ser utilizados para personalizar a visualização dos *datasets*, sendo alguns já incluídos no pacote de instalação padrão e outros que devem ser inseridos à parte [DOCS.CKAN 2022].

Um dos *plugins* mais relevantes, que é pre-requisito para alguns dos serviços da aplicação, é o DataStore. A extensão armazena o conteúdo dos *datasets* em uma base de dados *ad hoc* - própria para este fim - e possibilita a pré-visualização dos recursos. O DataStore também fornece uma API para leitura, busca e filtragem dos conjuntos de dados. Nas Figuras 2 e 3 é possível ver o mesmo conjunto de dados, com o quantitativo de alunos dos cursos de licenciatura exibidos como tabela (Figura 2) e gráfico (Figura 3) com o DataStore Grid e o DataStore Graph, respectivamente.

Grade | Gráfico | Mapa 11 records « 1 – 11 » LICENCIATURA Go »

"CURSO"	ATIVO	ATIVO - FORMANDO	TRANCADO
LICENCIATURA EM CIÊNCIAS BIOLÓGICAS	617	29	1
LICENCIATURA EM COMPUTAÇÃO	332	0	1
LICENCIATURA EM EDUCAÇÃO FÍSICA	292	5	2
LICENCIATURA EM FÍSICA	241	4	1
LICENCIATURA EM HISTÓRIA	380	0	1
LICENCIATURA EM LETRAS - PORTUGUÊS E ESPANHOL	376	1	0
LICENCIATURA EM LETRAS - PORTUGUÊS E INGLÊS	275	4	2
LICENCIATURA EM MATEMÁTICA	470	2	8
LICENCIATURA EM PEDAGOGIA	234	0	0
LICENCIATURA EM QUÍMICA	426	14	2
LICENCIATURA EM QUÍMICA	220	3	0

Figura 2. Plugin para visualização de dados como tabela

O *plugin* funciona em conjunto com o DataPusher, que insere automaticamente os *datasets* na DataStore. Ambos já estão inclusos na instalação do CKAN por pacotes, sendo necessário apenas configurar uma nova base de dados no PostgreSQL.

Outra opção de *plugin* para visualização de dados é o Data Explorer, que não requer que o dataset esteja no DataStore. Isso é feito através de um recurso externo, o DataProxy, que converte e retorna o conteúdo para visualização. Em comparação com o DataStore Grid, ele suporta menos formatos de arquivos e codificações.

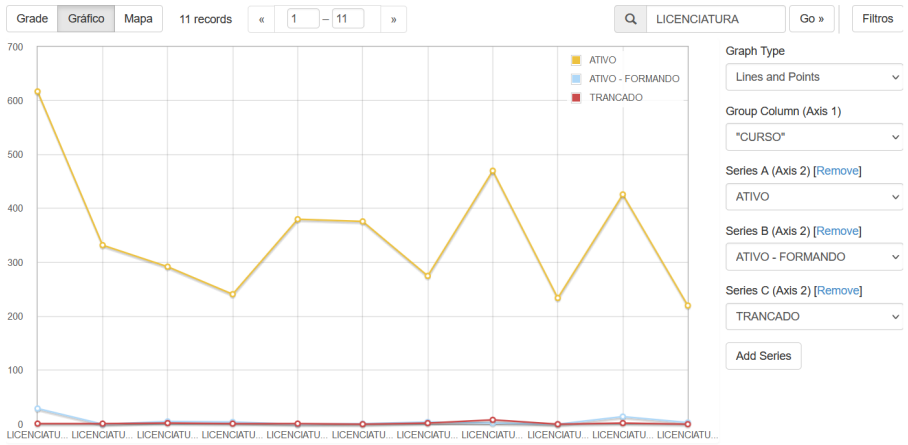

Figura 3. Plugin para visualização de dados como gráfico

Quando utilizados em conjunto, o Data Explorer serve como uma redundância para o DataStore Grid, oferecendo visualização de dados caso o serviço do DataStore/DataPusher esteja indisponível. Para isso, no arquivo de configuração principal da aplicação, existe uma seção em que são definidas as visualizações criadas por padrão.

O CKAN também possui extensões para visualização de arquivos de texto, imagens e gráficos. São eles o Text view, que mostra o conteúdo de arquivos de texto com a sintaxe realçada; Image view, que exibe a imagem e insere uma tag com referência para sua a URL; e DataStore Map, que gera um mapa interativo através de valores de latitude e longitude; o PDF Viewer, que renderiza os arquivos em PDF para pré-visualização.

O próximo passo no processo de implantação do portal se dá através da criação das Organizações e Grupos, além da personalização das páginas do CKAN de acordo com as necessidades da instituição. Na estrutura da aplicação, os conjuntos de dados pertencem a uma organização. Para catalogação no Portal Brasileiro de Dados Abertos, deve ser criada apenas uma organização por órgão ou entidade.

Os conjuntos de dados - que podem ser públicos ou privados - são compostos pelos seus metadados e recursos. Eles são agrupados por Organização e por Grupos temáticos. Cada conjunto pertence sempre a uma Organização e a um ou mais Grupos. Os metadados se aplicam para todos os recursos daquele conjunto [Costa et al. 2017]. Os grupos reúnem conjuntos de dados e devem refletir o que foi definido no PDA da instituição.

Um dos fatores determinantes para a escolha do CKAN como aplicação para o portal de dados abertos foi a possibilidade de fazer a integração de forma automática com o portal do Governo Federal. O processo de sincronização - feita sempre que um recurso é inserido ou atualizado - é realizado pela CGU. Qualquer órgão cujo portal de dados abertos utilize o CKAN pode solicitar a sincronização dos dados [Brasil 2022a].

5. Estudo de Caso

Esta seção apresenta um estudo de caso, com o objetivo de mapear as etapas do processo de abertura de dados pelos responsáveis. Para isso, foram escolhidas três bases de dados

do PDA da UFRPE: Ensino de Graduação, Financeiro/Orçamento e Contratos. O critério para seleção dos participantes do estudo de caso foi o próprio PDA, que contém a lista das bases e o responsável por sua abertura. Os encontros e reuniões foram organizados pelo CTDA, no papel de responsável pela gestão dos dados abertos da instituição.

O processo de publicação dessas bases teve início através de um diálogo com os responsáveis pelos dados, a fim de avaliar a melhor forma de extrair as informações dos sistemas informatizados. Em seguida, esses dados foram adaptados para o formato aberto e foi produzido um modelo de dicionário de dados. Esses artefatos foram aprovados pelos responsáveis pelas bases e disponibilizados no portal.

5.1. Caso I - Ensino de Graduação

A base de Ensino de Graduação foi escolhida para ser o foco deste primeiro estudo de caso. Segundo o PDA da UFRPE, as informações relacionadas ao Censo da Educação Superior e os demais dados dos cursos de graduação são geridos por setores diferentes. Por esse motivo, são criados dois grupos no CKAN: *"Censo da Educação Superior"* e *"Ensino de Graduação"*. Essa separação permite que cada responsável seja o único editor de seus dados e tenha acesso irrestrito para adicionar, remover e editar os dados.

No caso dos dados de Ensino de Graduação, esta pesquisa explora o processo de extração, tratamento e publicação das informações. Como o sistema acadêmico utilizado pela UFRPE não possibilita a extração de relatórios com a informação desejada em formato aberto, as informações foram extraídas diretamente da base de dados. O acesso à base de dados foi feito através do *software* DBeaver. As consultas, feitas em Structured Query Language (SQL), foram exportadas em formato CSV.

Por orientação da INDA, os dados publicados no portal devem estar em seu estado primário, ou seja, com o mínimo de agregações possíveis. Isso permite que os usuários possam realizar suas próprias análises, não ficando restritos ao que foi disponibilizado pela instituição. Agregações podem ser divulgadas, desde que de forma adicional.

Esta foi a lógica adotada para a extração dos dados básicos como de cursos, matrizes curriculares, componentes e docentes. As únicas combinações realizadas foram feitas com o objetivo de extrair descrições como nome do município ou curso, de forma a não exibir apenas o código. A nomenclatura das colunas foi ajustada através do uso de *aliases*. Em alguns casos, é necessário fazer a divisão pela dimensão temporal e, devido à natureza semestral dos cursos, optou-se por publicar as bases por semestre.

Durante a elaboração das listagens de alunos dos cursos de graduação e matriculados nas turmas de graduação, optou-se por utilizar quantitativos, como total de alunos por curso (Figura 4) ou total de matrículas por situação (Figura 5), como forma de respeitar à LGPD. Não devem ser exibidos nem dados pessoais nem identificadores que, em um eventual cruzamento de dados, permitam a identificação destas informações pessoais.

Enquanto alguns dados pessoais devem ser removidos ou anonimizados, existem situações em que as instituições têm obrigação legal de disponibilizar informações que contém esse tipo de dado. No caso das IFES, essas exceções podem incluir as listas de docentes e coordenadores de curso. Nestas situações, o dicionário de dados deve incluir uma menção à hipótese legal que permite a divulgação dos dados pessoais.

"CURSO"	ATIVO	ATIVO - FORMANDO	TRANCADO
BACHARELADO EM ADMINISTRAÇÃO	340	9	6
BACHARELADO EM ADMINISTRAÇÃO	221	3	0
BACHARELADO EM AGROECOLOGIA	82	0	1
BACHARELADO EM AGRONOMIA	424	4	4
BACHARELADO EM AGRONOMIA	243	2	1
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO	400	4	11
BACHARELADO EM CIÊNCIAS BIOLÓGICAS	325	1	3
BACHARELADO EM CIÊNCIAS BIOLÓGICAS	198	9	0
BACHARELADO EM CIÊNCIAS DO CONSUMO	249	3	1
BACHARELADO EM CIÊNCIAS ECONÔMICAS	304	2	2

Figura 4. Quantitativo de alunos de graduação no CKAN

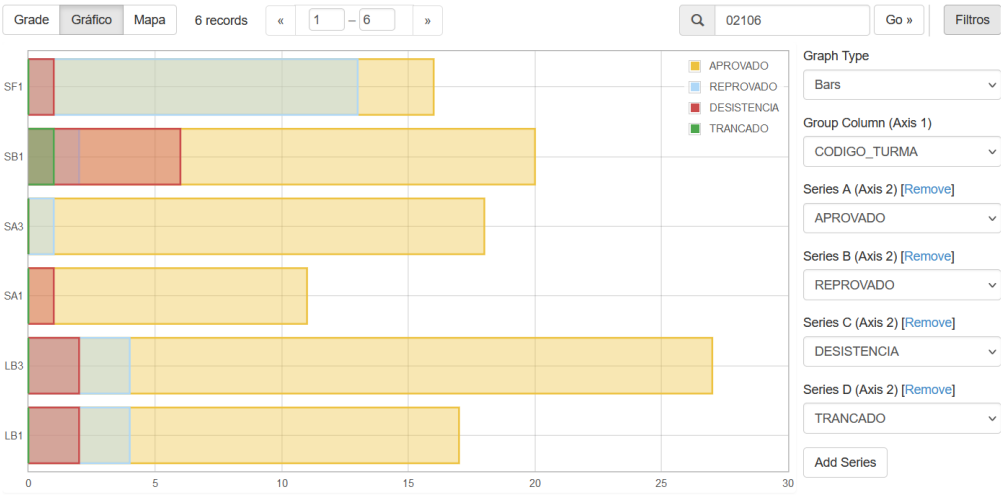

Figura 5. Quantitativos de matriculados nas turmas de graduação no CKAN

O processo de extração das informações do sistema acadêmico foi documentado, de forma a permitir extrações periódicas, conforme o prazo estabelecido no PDA da instituição. Os dados foram validados pelos responsáveis, que analisaram tanto à adequação à LGPD como a completude e relevância. Por fim, foi elaborado um dicionário de dados para cada base a ser publicado, com as descrições das colunas e tipo de dado.

O maior desafio encontrado na extração dos dados de Ensino de Graduação foi a quantidade de informações previstas para divulgação no PDA e o fato de o sistema acadêmico não possuir relatórios no modelo desejado. A coleta das informações, quando o sistema não possibilita geração de relatórios, exige um alto nível de conhecimento da estrutura do banco de dados, necessário tanto para a extração como para o tratamento. Assim, as instituições devem mensurar o esforço necessário e incluí-lo nos prazos do PDA, além de formalizar o envolvimento por parte do setor de TI da instituição, tanto na

extração e limpeza inicial dos dados quanto nas atualizações periódicas.

5.2. Caso II - Orçamento

Esta etapa do estudo de caso se propõe a investigar como se dá a publicação de dados vindos de sistemas do Governo Federal quando as IFES já fazem um tratamento dessas informações. O objetivo é reaproveitar tais dados para o portal de dados abertos, uma vez que o PDA não definiu quais dados seriam publicados, deixando a critério do setor responsável. Além de já divulgar os dados da execução orçamentária e financeira no Portal da Transparência, tal setor já disponibilizava estas informações em seu site institucional.

Nesta etapa do trabalho, o principal obstáculo encontrado foi a definição do PDA da UFRPE em separar a execução orçamentária e financeira em duas bases distintas. Juntamente com o servidor responsável pelos dados, foi estabelecido que essa divisão não faz sentido na prática. Por isso, foi criado apenas um grupo no CKAN denominado Orçamento e Financeiro, no qual foram publicadas as bases de Liquidações e Empenhos.

De forma a reduzir o retrabalho e incentivar que a atualização do portal de dados abertos fosse mantida de acordo com a periodicidade proposta no PDA - mensal -, os mesmos dados contidos no painel foram publicados no portal. Isso foi possível devido ao fato de o responsável já ter realizado todo o trabalho de limpeza e tratamento dos dados previamente, sendo necessário apenas converter o formato do arquivo para CSV.

A produção do dicionário de dados também se mostrou complexa. Apesar de os dados já estarem prontos para publicação, a nomenclatura das colunas não permitia o entendimento e análise por usuários leigos. Assim, foi preciso inserir essas definições no dicionário, juntamente com o tipo de dado e obrigatoriedade, para possibilitar seu uso.

Este estudo evidenciou a relevância das etapas de tratamento e limpeza no processo de abertura de dados. Como neste caso o responsável já realizava esse trabalho de forma rotineira, a inclusão no portal foi relativamente simples e ágil, não sendo necessário o envolvimento do setor de TI da instituição nas fases de extração e limpeza.

5.3. Caso III - Contratos

A última etapa deste estudo trata da base de Contratos. A proposta é mapear os passos envolvidos na publicação de dados oriundos dos sistemas do Governo Federal mas que, ao contrário do caso anterior, não passavam por nenhum tratamento dentro da instituição. O sistema utilizado pela UFRPE é o Contratos, módulo do *compras.gov.br*, a solução do Governo Federal para controle orçamentário e contratual.

Como o PDA não especificava em detalhes os dados que deveriam ser publicados, a primeira etapa do processo consistiu na escolha das informações que seriam divulgadas no portal de dados abertos. A decisão, tomada juntamente com o setor responsável, foi pela relação dos contratos vigentes da instituição, em formato similar ao divulgado na seção de Acesso à Informação do site da UFRPE. Como o volume de dados não é tão grande, eles foram divididos na dimensão anual. Conforme definido no PDA, a atualização será mensal. A cada atualização, os recursos são sobrescritos pela versão mais nova.

Quanto à extração dos dados, a princípio o sistema Contratos já permite que ela seja feita em arquivo CSV. No entanto, foi necessário um trabalho extenso de limpeza e correção de inconsistências. Ao avaliar as informações extraídas do sistema, o responsável identificou vários campos que eram desnecessários ou incorretos. Era o caso do número de parcelas e os dados da autoridade signatária, que sempre estão em branco.

Sendo assim, foi feita uma adaptação do arquivo extraído do sistema, visando melhorar a qualidade da informação e facilitar o entendimento por parte do público. Foram inseridas informações sobre a unidade de origem da licitação e situação (ativo ou encerrado). Esse modelo será utilizado para elaboração das próximas atualizações do arquivo.

Além do detalhamento das colunas, o dicionário de dados recebeu os possíveis valores de alguns dos campos utilizados, como categoria, natureza e termo do contrato. Devido à possibilidade de conter o nome e CPF dos fornecedores, o documento recebeu uma justificativa para que ele ficasse de acordo com a LGPD, uma vez que nesse caso essa divulgação se enquadra na hipótese de cumprimento de obrigação legal.

Essa etapa do estudo difere das anteriores principalmente no tratamento e limpeza dos dados. No caso dos dados de graduação, esse processo foi realizado durante a extração do banco de dados ou seguindo as especificações do Censo. Já os responsáveis pelo Orçamento já tinham as informações devidamente tratadas. Para Contratos, foi necessário estabelecer os critérios para essa limpeza, decidindo quais dados manter, corrigir ou eliminar, tendo sempre como critério a melhor experiência do usuário do portal.

5.4. Principais Descobertas

Esta subseção tem como objetivo reunir as principais descobertas do estudo de caso, que começou com a implantação do CKAN, seguida pelo processo de configuração e personalização. Neste momento, o PDA da instituição ainda estava em fase de elaboração, faltando os ajustes finais. Foi necessário iniciar o processo de abertura dos dados antes da publicação para poder cumprir os prazos propostos pelo documento.

Ao longo do estudo, foi possível observar a relevância da participação dos responsáveis pela publicação dos dados não só no momento de inserir as informações no portal, mas também na elaboração do PDA. Tanto em Graduação quanto em Orçamento foi preciso fazer ajustes no que o documento previa - a frequência de atualização e a denominação das bases de dados, respectivamente.

A pesquisa também mostrou os desafios das etapas de extração, tratamento e limpeza dos dados. O grau de complexidade pode variar bastante de acordo com a qualidade dos dados originais, do nível de conhecimento do responsável pela informação e da necessidade ou não de extrair as informações diretamente do banco de dados.

O estudo de caso abordou as etapas necessárias para a abertura dos dados por parte das IFES. Os problemas encontrados e suas respectivas soluções foram utilizados como base para elaboração de um guia de boas práticas para publicação de dados abertos.

6. Boas Práticas para Publicação de Dados Abertos

O processo Extraction-Transformation-Loading (ETL) pode ser definido como uma sequencia de etapas com o objetivo de realizar a extração, transformação e limpeza e carga

de dados. Os dados podem vir de uma ou mais fontes e recebem tratamento antes de serem inseridas em outra base de dados, como um *data warehouse* [Vassiliadis 2009]. Nesta seção, é proposta uma abordagem para ETL de dados de IFES com finalidade de realizar carga em portais de dados abertos. Seu principal diferencial consiste em incluir no processo etapas com orientações para garantir a adesão aos requisitos da legislação e às boas práticas de dados abertos governamentais.

6.1. Extração e Preparação do Conteúdo

As bases de dados a serem publicadas no portal são definidas no PDA de cada instituição. A legislação apresenta algumas normas a serem seguidas no processo de disponibilização dessas informações. Esta subseção reúne as recomendações da legislação e dos guias da CGU as consolida em um guia de boas práticas a serem seguidos, ilustrado na Figura 6.

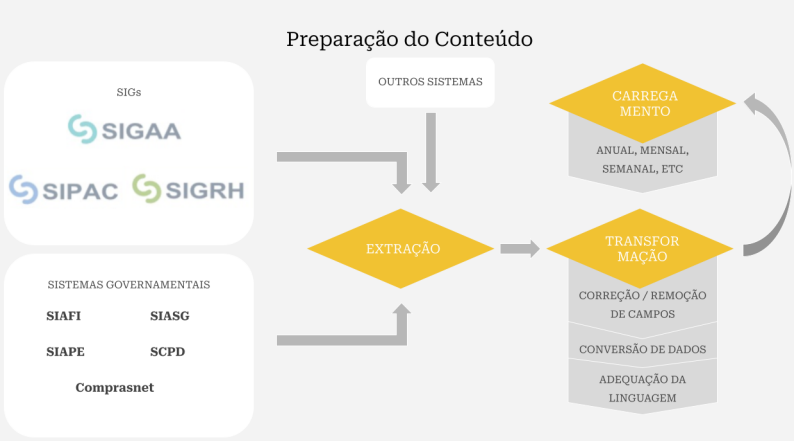

Figura 6. Processo de preparação dos dados

A primeira etapa consiste na extração dos dados dos sistemas informatizados, o que pode ser feito via relatório gerado pelo sistema ou acessando diretamente sua base de dados. Pode ser necessário remover campos que não sejam relevantes, converter dados ou ajustar informações que estejam incorretas no sistema.

Em relação ao conteúdo, uma das principais diretrizes consiste em publicar os dados em seu formato primário. Caso o órgão ou entidade deseje realizar algum tipo de agregação, as informações devem ser divulgadas das duas formas, primária e agregada [Brasil 2019], sempre com o devido tratamento e remoção de dados pessoais e sensíveis.

Sempre que possível, os dados devem estar em linguagem simples, que possa ser facilmente compreendida pelo usuário, evitando o uso de nomenclaturas pouco conhecidas ou termos técnicos mais específicos. Caso seja necessário fazer uso desses recursos, incluir um glossário ou vocabulário, como o VCGE [Brasil 2019].

Para conjuntos de dados muito grandes, é recomendado fazer sua divisão em subconjuntos menores para facilitar sua manipulação. Pode-se usar a dimensão temporal, por ano ou mês, geográfica, ou outra dimensão que o órgão julgue adequada.

6.2. Inclusão dos Metadados

A publicação de um conjunto de dados abrange não só os dados em si, mas também os seus metadados. Metadados são dados sobre os dados, ou seja, que permitem organizar, classificar, relacionar e inferir novas informações sobre eles. [Brasil 2022a]. Essas informações agregam valor ao *dataset* e permitem que ele seja encontrado e manipulado.

Por recomendação da legislação, os metadados devem conter no mínimo o nome e descrição do conjunto de dados, etiquetas, nome e e-mail do setor responsável pelos dados, periodicidade de atualização, escopo temporal e assuntos do VCGE [Brasil 2019]. Com exceção do último, todos esses campos existem por padrão no CKAN. Para o vocabulário, a aplicação permite que sejam inseridos campos personalizados, que podem ser usados para informar a categoria do VCGE mais adequada ao conjunto.

Os metadados também precisam informar a licença aberta, de forma a garantir a permissão irrestrita de uso dos dados. O CKAN conta com várias licenças pré-cadastradas, incluindo a PDDL, que foi escolhida para este trabalho. Ao selecionar a licença neste campo, a ferramenta adiciona o nome da licença nos metadados e um link para sua descrição e texto completo, atendendo assim os critérios para sua aplicação.

Outro componente essencial dos metadados é o dicionário de dados. O documento, que deve acompanhar todo conjunto de dados, possui as seguintes informações sobre os valores contidos em cada coluna do *dataset*: nome, descrição, tipo do dado e obrigatoriedade. Caso a base contenha algum dado pessoal, é preciso mencionar a hipótese legal que permite sua divulgação sem consentimento do titular. Uma justificativa para remoção de dados sensíveis também pode constar no documento.

Figura 7. Processo de elaboração dos metadados

A Figura 7 mostra os elementos envolvidos na elaboração dos metadados, separando-os em duas categorias, as pertencentes ao dicionário de dados e as que são

inseridas em campos próprios do CKAN. Tal distinção foi utilizada como forma de guiar os responsáveis pela publicação dos dados acerca de onde inserir cada um dos campos.

6.3. Processo de Publicação de Dados Abertos

Em relação ao formato, os dados publicados no portal devem estar em formato não proprietário, como: CSV, JSON, XML, ODS ou RDF. Optou-se por utilizar o CSV pois ele estava disponível em todas as fontes de dados utilizadas. Para dados geográficos, a e-PING recomenda SVG e GML. O formato PDF deve ser usado apenas no dicionário de dados. O uso de compactadores de arquivos é desaconselhado mas, caso seja indispensável, deve-se escolher um formato aberto, como 7Z, GZIP ou ZIP [Brasil 2022a].

Outra recomendação importante é a utilização de URLs amigáveis, que permitam ao usuário identificar o conteúdo do conjunto de dados. Todas as palavras devem estar em letras minúsculas, separadas por hífen e sem acentuação. Isso também proporciona uma melhor indexação do conteúdo por motores de busca [Brasil 2019]. O CKAN já gera as URLs dos conjuntos de dados nativamente nesse padrão. No entanto, a ferramenta também permite que o autor do *dataset* faça a pré-visualização da URL antes de publicar e realize qualquer alteração que julgar necessária.

Com base nas considerações apresentadas, foi estabelecido um processo a ser seguido para publicação das bases de dados, que consiste em três fases: preparação, metadados e publicação, ilustradas na Figura 8. A fase de preparação consiste na extração, limpeza e ajustes nos dados. A elaboração dos metadados consiste no cadastro de descrição, vocabulário e dicionário de dados. Já a publicação envolve a publicação dos dados no CKAN, com os ajustes finais no formato do arquivo e URL.

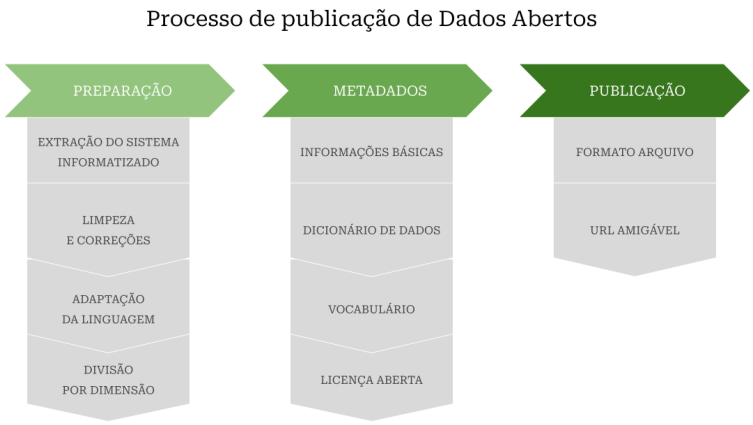

Figura 8. Processo de publicação de dados abertos governamentais

6.4. Papéis e responsabilidades

Esta subseção detalha os papéis e responsabilidades de cada um dos envolvidos no processo de publicação de dados abertos: os gestores de dados abertos do órgão, o setor de TI e os editores de dados. Tal atribuição é feita pelo PDA das instituições, conforme orientação da Política de Dados Abertos do Poder Executivo Federal [Brasil 2016].

Por determinação da CGINDA e da LAI, cada instituição possui um servidor na função de autoridade designada, que tem como atribuições instituir e monitorar a política de dados abertos daquele órgão [CGINDA 2017b]. Isto inclui a elaboração do PDA, que deve detalhar o papel de cada um dos envolvidos no processo. No caso da UFRPE, foi instituído o CTDA, trabalhando em conjunto com a autoridade designada.

No portal de dados abertos, cabe ao CTDA ou entidade similar a função de fazer o gerenciamento de permissões de usuários e administrar a criação de Grupos e Organizações dentro do CKAN. Os servidores apontados como responsáveis pela publicação dos dados têm a responsabilidade de realizar as rotinas de publicação de dados.

A TI da instituição, por sua vez, é responsável por fazer a extração dos dados em casos em que isso não seja possível pela interface do sistema. Também são atribuições do setor de TI a manutenção do portal e gerenciamento de segurança de informação. A Figura 9 mostra o fluxo de responsabilidade de cada um dos papéis envolvidos no processo.

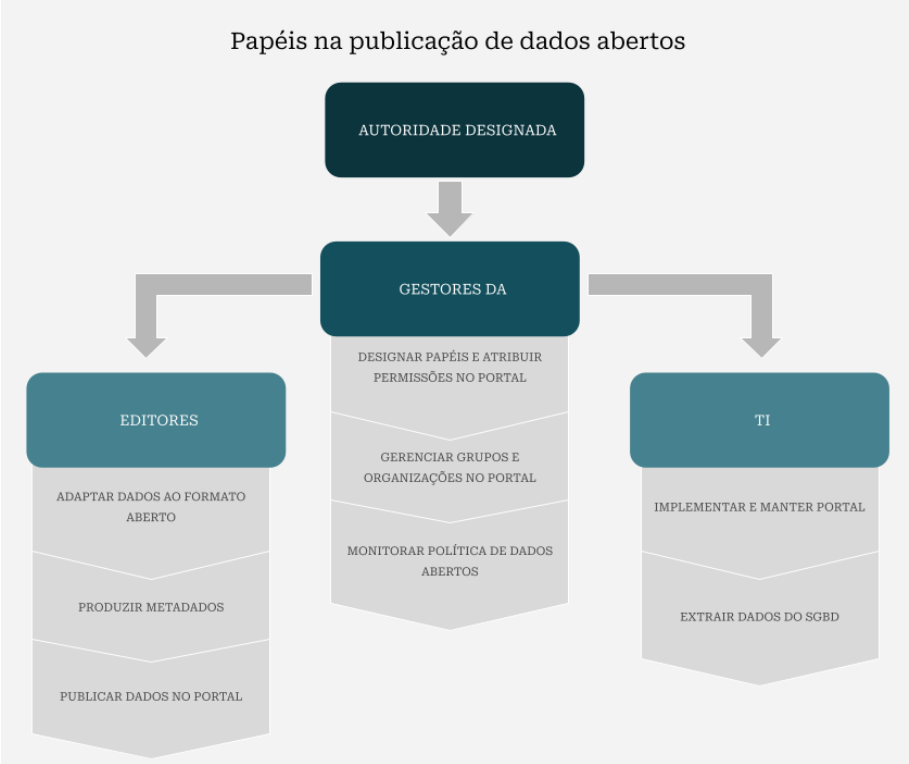

Figura 9. Papéis e responsabilidades na publicação de dados abertos

7. Conclusões

Esta pesquisa teve como objetivo principal estudar a implementação de um portal de dados abertos, identificar as melhores tecnologias disponíveis, e definir as melhores práticas para extração, tratamento e publicação das informações definidas pelo PDA das IFES. Através do estudo da literatura e da legislação, verificou-se que um dos principais desafios para o sucesso das iniciativas de transparência ativa consiste em compreender os requisitos necessários, tanto na parte técnica quanto no processo de escolha e tratamento dos dados.

Durante a pesquisa, o CKAN se mostrou a opção mais apropriada de gerenciador de conteúdo. No entanto, apesar de a aplicação ser extremamente versátil e completa, em sua versão padrão ela deixa de atender alguns dos requisitos necessários, como a disponibilização de API e questões de acessibilidade da e-MAG.

Como solução, foram implementados ajustes nas configurações padrão da aplicação, instalação de *plugins* - nativos ou não - e ajustes no código para melhorar a usabilidade. Apesar de não representar um impedimento para a utilização do CKAN, a necessidade de tais ajustes deve ser levada em conta pelas organizações durante a implantação.

Outro desafio considerável do processo de abertura de dados consiste nos próprios dados. A informação tem que ser vista como um meio, não um fim, e precisa permitir que análises e reusos sejam feitos a partir dela. Para que isso seja possível, é necessário um esforço por parte das organizações para limpar e adequar seus dados.

Conforme visto nos estudos de caso, os setores que utilizam sistemas do Governo Federal com possibilidade de geração de relatórios não podem simplesmente publicá-lo em sua formatação original. É preciso remover informações imprecisas ou irrelevantes, ajustar campos, anonimizar dados pessoais e convertê-los para formatos abertos.

O estudo de caso com as bases de dados de Orçamento e Contratos deixou claro a relevância da cultura organizacional no tratamento das informações e da política de abertura de dados como um todo. A base de Orçamento, por receber esse tratamento de forma rotineira, já se encontrava pronta para publicação. A base de Contratos, por sua vez, precisou passar pelas etapas de limpeza e transformação.

Já o estudo com as bases de Graduação mostrou que os casos em que os dados precisam ser extraídos de SGBDs são ainda mais complexos. Além de todo o trabalho de limpeza e formatação, os órgãos devem a divulgar os dados em sua forma primária, permitindo que o público faça suas próprias análises. Isso não impede que as instituições divulguem informações tratadas, com quantitativos ou combinações de bases.

Outra etapa fundamental é a elaboração dos metadados. Essas informações - como etiquetas, descrição e dicionário de dados - permitem que a sociedade encontre e consiga compreender as informações disponibilizadas. Os metadados também devem conter menção à licença aberta, que reafirma ao público o direito de utilizar os dados.

Para garantir que o processo de publicação seja feito corretamente e atinja os resultados esperados, foi definido um guia de boas práticas para o tratamento de dados, englobando desde o processo de extração da informação do sistema informatizado ao seu tratamento e publicação no portal, sempre atendendo aos requisitos da legislação.

Com isso, este trabalho entregou sua principal contribuição. Ao abordar os dois aspectos do processo de abertura de dados, técnico e negocial, sempre guiado pela legislação e boas práticas de dados abertos governamentais, o estudo mostra os meios pelos quais as IFES podem obter sucesso em sua iniciativa de abertura, publicando dados de qualidade e, consequentemente, incentivando o seu uso pelos cidadãos.

Em relação às limitações encontradas ao decorrer da pesquisa, uma delas foi a utilização dos *plugins* do CKAN. Apesar de a ferramenta incentivar o desenvolvimento

de extensões, gerando um número considerável de *plugins*, utilizar alguns deles se provou um desafio. Uma das razões foi falta de compatibilidade com a versão atual do CKAN, o que resultaria em um esforço considerável para adaptar e utilizar a extensão.

A falta de documentação também foi um fator determinante para que algumas soluções - como *plugins* adicionais de visualização e para fazer o download de todos os recursos presentes em um *dataset* - não fossem implementadas, assim como a ausência de modularidade em determinados *plugins* que ofereciam várias alterações na ferramenta, mas não permitiam que tais mudanças fossem selecionadas individualmente.

Com base na pesquisa, algumas propostas de trabalhos futuros que podem ser realizados de forma a contribuir para a temática são: analisar e propor formas de automatizar a extração dos dados; estudar estratégia para automatizar as etapas de limpeza e transformação dos arquivos a serem publicados no portal; estudar melhorias nas opções de acessibilidade do portal; promover estratégia para abertura de dados de setores pouco - ou não - informatizados; e promover a criação de reúsos para as bases de dados publicadas no portal, incentivando a participação da comunidade acadêmica.

Referências

[Avila 2015] Avila, T. J. T. (2015). Uma proposta de modelo de processo para publicação de dados abertos conectados governamentais. Master's thesis. 219 f. Dissertação (Mestrado em Modelagem Computacional de Conhecimento) - Instituto de Computação, Programa de Pós-Graduação em Modelagem Computacional de Conhecimento, Universidade Federal de Alagoas, Maceió.

[Bachtiar et al. 2020] Bachtiar, A., Suhardi, and Muhamad, W. (2020). Literature review of open government data. *2020 International Conference on Information Technology Systems and Innovation (ICITSI)*, pages 329–334.

[Brasil 1988] Brasil (1988). Constituição.

[Brasil 2011] Brasil (2011). Lei de acesso à informação.

[Brasil 2016] Brasil (2016). Decreto nº 8777.

[Brasil 2018] Brasil (2018). Lei geral de proteção de dados pessoais.

[Brasil 2019] Brasil (2019). Guia de transparência ativa (gta) para os órgãos e entidades do poder executivo federal. *Controladoria Geral da União (CGU)*.

[Brasil 2022a] Brasil (2022a). Cartilha de publicação de dados abertos.

[Brasil 2022b] Brasil (2022b). Painel de monitoramento de dados abertos.

[Brasil 2022c] Brasil (2022c). Vocabulário controlado do governo eletrônico.

[CGINDA 2017a] CGINDA (2017a). Resolução nº 2, de 24 de março de 2017.

[CGINDA 2017b] CGINDA (2017b). Resolução nº 3, de 13 de outubro de 2017.

[Correa et al. 2014] Correa, A. S., Correa, P. L. P., and da Silva, F. S. C. (2014). Transparency portals versus open government data: An assessment of openness in brazilian municipalities. In *Proceedings of the 15th Annual International Conference on Digital Government Research*, page 178–185. Association for Computing Machinery.

[Costa et al. 2017] Costa, L. R., Shintaku, M., Silveira, L. A., Macedo, D. J., and Fonseca, R. M. S. d. (2017). *Guia do Usuário CKAN*. Ibict, Brasília.

[Craveiro and Martano 2014] Craveiro, G. S. and Martano, A. M. (2014). Abertura e disponibilização de dados abertos governamentais: Estudos de caso. In *II Workshop de Transparência em Sistemas*, Londrina. Simpósio Brasileiro de Sistemas de Informação.

[Davies 2012] Davies, T. (2012). Ten building blocks of an open data initiative. Open Data Impacts – Research Note, Vol. 1.

[DOCS.CKAN 2022] DOCS.CKAN (2022). Plugins.

[Dorobăț and Posea 2021] Dorobăț, I. C. and Posea, V. (2021). Open data indicator: An accumulative methodology for measuring the quality of open government data. *13th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*.

[Eaves 2009] Eaves, D. (2009). The three laws of open data. *Eaves.ca*.

[Gao et al. 2023] Gao, Y., Janssen, M., and Zhang, C. (2023). Understanding the evolution of open government data research: towards open data sustainability and smartness. *International Review of Administrative Sciences*, 89:59–75.

[Khayyat and Bannister 2014] Khayyat, M. and Bannister, F. (2014). Open data licensing: More than meets the eye. *Information Polity*, 20.

[Lima et al. 2020] Lima, M. P. d., Abdalla, M. M., and Oliveira, L. G. L. (2020). A avaliação da transparência ativa e passiva das universidades públicas federais do brasil à luz da lei de acesso à informação. *Revista do Serviço Público - RSP, v. 71*, pages 232–263.

[Lorenzon 2021] Lorenzon, L. N. (2021). Análise comparada entre regulamentações de dados pessoais no brasil e na união europeia (lgpd e gdpr) e seus respectivos instrumentos de enforcement. *Revista do Centro de Excelência Jean Monnet da FGV Direito Rio, Rio de Janeiro, vol. 1*, pages 39–52.

[Marijn Janssen and Zuiderwijk 2012] Marijn Janssen, Y. C. and Zuiderwijk, A. (2012). Benefits, adoption barriers and myths of open data and open government. *Information Systems Management*, 29:258–268.

[Molloy 2011] Molloy, J. C. (2011). The open knowledge foundation: Open data means better science. *PLOS Biology*, 9:1–4.

[OKFN 2022] OKFN (2022). Open data commons public domain dedication and license (pddl).

[Oliveira and Fonseca 2021] Oliveira, W. Q. d. S. and Fonseca, I. F. d. (2021). Fatores de sucesso na abertura de dados: o caso do banco central do brasil. *Revista Do Serviço Público, 72*, pages 724–752.

[Osagie et al. 2015] Osagie, E., Mohammad, W., Stasiewicz, A., Hassan, I. A., Porwol, L., and Ojo, A. (2015). State-of-the-art report and evaluation of existing open data platforms. *ROUT-TO-PA*.

[PDA-UFRPE 2022] PDA-UFRPE (2022). Plano de dados abertos da ufrpe.

[Shintaku and Sales 2019] Shintaku, M. and Sales, L. (2019). *Ciência aberta para editores científicos*. ABEC, Botucatu, SP.

[Silva and Júnior 2018] Silva, R. d. O. and Júnior, G. S. d. A. (2018). A process proposal for implementation of open data in brazilian public institutions. *iSys - Brazilian Journal of Information Systems*, 11(1):30–54.

[Ubaldi 2013] Ubaldi, B. (2013). Open government data: Towards empirical analysis of open government data initiatives. *OECD Working Papers on Public Governance, No. 22*.

[Vassiliadis 2009] Vassiliadis, P. (2009). A survey of extract–transform–load technology. *International Journal of Data Warehousing and Mining*, pages 1–27.